Introduction to Statistics:

A Tour in 14 Weeks

Reinhard Furrer
and the Applied Statistics Group

Version February 21, 2023
Git b91ab56

Contents

Prologue

This document accompanies the lecture *STA120 Introduction to Statistics* that has been given each spring semester since 2013. The lecture is given in the framework of the minor in *Applied Probability and Statistics* (www.math.uzh.ch/aws) and comprises 14 weeks of two hours of lecture and one hour of exercises per week.

As the lecture's topics are structured on a week by week basis, the script contains thirteen chapters, each covering "one" topic. Some of chapters contain *consolidations* or *in-depth studies* of previous chapters. The last week is dedicated to a recap/review of the material.

I have thought long and hard about an optimal structure for this script. Let me quickly summarize my thoughts. It is very important that the document contains a structure that is tailored to the content I cover in class each week. This inherently leads to 14 "chapters." Instead of covering Linear Models over four weeks, I framed the material in four *seemingly* different chapters. This structure helps me to better frame the lectures: each week having a start, a set of learning goals and a predetermined end.

So to speak, the script covers not 14 but essentially only three topics:

1. Background
2. Statistical foundations in a (i) frequentist and (ii) Bayesian approach
3. Linear Modeling

We will not cover these topics chronologically. For a smoother setting, we will discuss some part of the background (multivariate Gaussian distribution) before linear modeling, when we need it. This also allows a recap of several univariate concepts. Similarly, we cover the Bayesian approach at the end. The book is structured according the path illustrated in Figure 1.

In case you use this document outside the lecture, here are several alternative paths through the chapters, with a minimal impact on concepts that have not been covered:

- Focusing on linear models: Chapters 1, 2, 3, 7, 4, 5, 9, 10, 11 and 12
- Good background in probability: You may omit Chapters 2, 3 and 7
- Bare minimum in the frequentist setting: Chapters 2, 3, 7, 4, 5, 9, 10, and 11

All the datasets that are not part of regular CRAN packages are available via the URL www.math.uzh.ch/furrer/download/sta120/. The script is equipped with appropriate links that facilitate the download.

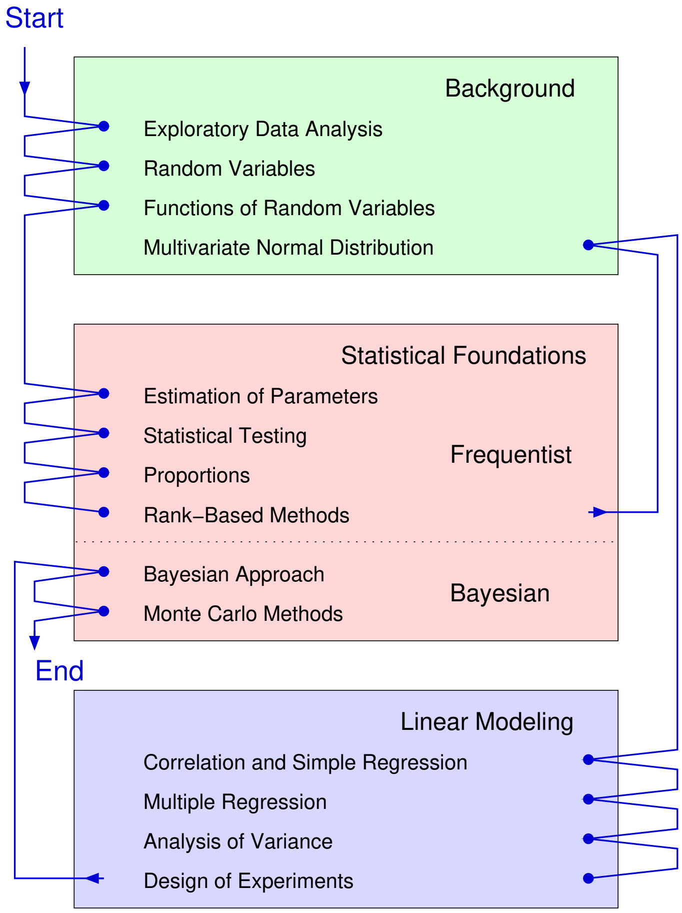


Figure 1: Structure of the script

The lecture *STA120 Introduction to Statistics* formally requires the prerequisites *MAT183 Stochastic for the Natural Sciences* and *MAT141 Linear Algebra for the Natural Sciences* or equivalent modules. For the content of these lectures we refer to the corresponding course web pages www.math.uzh.ch/fs20/mat183 and www.math.uzh.ch/hs20/mat141. It is possible to successfully pass the lecture without having had the aforementioned lectures, some self-studying is necessary though. This book and the accompanying exercises require differentiation, integration, vector notation and basic matrix operations, concept of solving a linear system of equations. Appendix B and C give the bare minimum of relevant concepts in calculus and in linear algebra. We review and summarize the relevant concepts of probability theory in Chapter 2 and in parts of Chapter 3.

I have augmented this script with short video sequences giving additional – often more technical – insight. These videos are indicated in the margins with a 'video' symbol as here. 6 min

Some more details about notation and writing style.

- I do not differentiate between 'Theorem', 'Proposition', 'Corollary', they are all termed as 'Property'.
- There are very few mathematical derivations in the main text. These are typical and important. Quite often we only state results. For the interested, the proofs of these are given in the 'theoretical derivation' problem at the end of the chapter.

- Some of the end-of-chapter problems and exercises are quite detailed others are open. Of course there are often several approaches to achieve the same; R-Code of solutions is *one* way to get the necessary output.
- Variable names are typically on the lower end of explicitness and would definitely be criticized by a computer scientists scrutinizing the code.
- I keep the headers of the R-Codes rather short and succinct. Most of them are linked to examples and figures with detailed explanations.
- R-Code contain short comments and should be understandable on its own. The degree of difficulties and details increases over the chapters. For example, in earlier chapters, we use explicit loops for simulations, whereas in the latter chapters we typically vectorize.
- The R-Code of each chapter allows to reconstruct the figures, up to possible minor differences in margin specifications. There are a few illustrations made in R, for which the code is presented only online and not in the book.
- At the end of each chapter there are standard references for the material. In the text I only cite particularly important references.
- For clarity, we omit the prefix 'http://' or 'https://' from the URLs - unless necessary. The links have been tested and worked at the time of the writing.

Many have contributed to this document. A big thanks to all of them, especially (alphabetically) Zofia Baranczuk, Federico Blasi, Julia Braun, Matteo Delucci, Eva Furrer, Florian Gerber, Michael Hediger, Lisa Hofer, Mattia Molinaro, Franziska Robmann, Leila Schuh and many more. Kelly Reeve spent many hours improving my English. Without their help, you would not be reading these lines. Please let me know of any necessary improvements and I highly appreciate all forms of contributions in form of errata, examples, or text blocks. Contributions can be deposited directly in the following Google Doc sheet.

Major errors that are detected *after* the lecture of the corresponding semester started are listed www.math.uzh.ch/furrer/download/sta120/errata.txt.

Reinhard Furrer
February 2023

Chapter 1

Exploratory Data Analysis and Visualization of Data

Learning goals for this chapter:

- ◇ Understand the concept and the need of an exploratory data analysis (EDA) within a statistical data analysis
- ◇ Know different data types and operations we can perform with them
- ◇ Calculate different descriptive statistics from a dataset
- ◇ Perform an EDA in R
- ◇ Sketch schematically, plot with R and interpret a histogram, barplot, boxplot
- ◇ Able to visualize multivariate data (e.g. scatterplot) and recognize special features therein

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter01.R.

It is surprisingly difficult to start a statistical study from scratch (somewhat similar as starting the first chapter of a book). Hence to get started, we assume a rather pragmatic setup: suppose we have "some" data. This chapter illustrates the first steps thereafter: exploring and visualizing the data. The exploration consists of understanding the types and the structure of the data, including its features and peculiarities. A valuable graphical visualization should quickly and unambiguously transmit its message. Depending on the data and the intended message, different types of graphics can be used — but all should be clear, concise and stripped of clutter. Of course, much of the visualization aspects are not only used at the beginning of the study but are also used after the statistical analysis. Figure 1.1 shows one representation of a statistical data analysis flowchart and we discuss in this chapter the right most box of the top line, *performing an exploratory data analysis.* Of course, many elements thereof will be again very helpful when we summarize the results of the statistical analysis (part of the bottom right most box in the workflow). Subsequent chapters come back to questions we should ask ourselves before we start collecting data, i.e., before we start an experiment, and how to conduct the statistical data analysis.

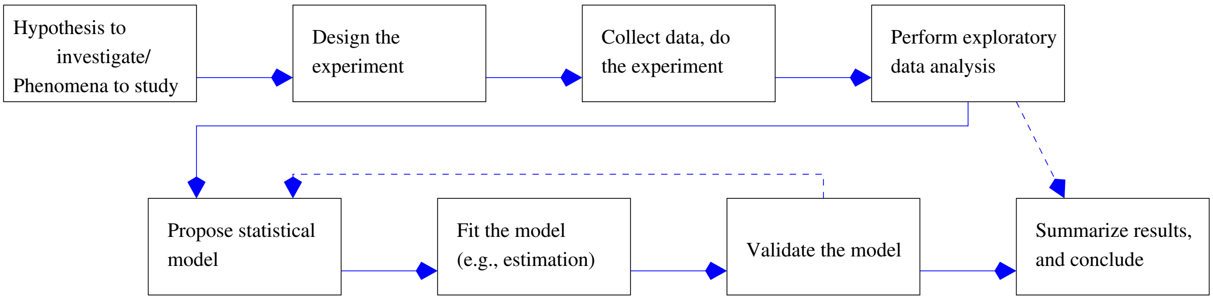

Hypothesis to investigate/ Phenomena to study
Design the experiment
Collect data, do the experiment
Perform exploratory data analysis
Propose statistical model
Fit the model (e.g., estimation)
Validate the model
Summarize results, and conclude

Figure 1.1: Data analysis workflow seen from a statistical perspective. There might be situations where an EDA shows that no statistical analysis is necessary (dashed arrow on the right). Similarly, model validation may indicate that the proposed model is not adequate (dashed back-pointing arrow).

The workflow is of course not always as linear as indicated, even trained statisticians may need to revisit statistical models used. Moreover, the workflow is just an extract of the scientific cycle, or the scientific method: conclusions lead to new or refined hypotheses that require again data and analyses.

1.1 Structure and Labels of Data

At the beginning of any statistical data analysis, an *exploratory data analysis* (EDA) should be performed (Tukey, 1977). An EDA summarizes the main characteristics of the data by representing observations or measured values graphically and describing them qualitatively and quantitatively. Each dataset tells us a 'story' that we should try to understand before we begin with the analysis. To do so, we should ask ourselves questions like

- What is the data collection or data generating process? (discussed in Section 1.1.1)
- What types of data do we have? (discussed in Section 1.1.2)
- How many data points/missing values do we have? (discussed in Section 1.2)
- What are the key summary statistics of the data? (discussed in Sections 1.2 and 1.3.1)
- What patterns/features/clusters exist in the data? (discussed in Section 1.3.2)

At the end of a study, results are often summarized graphically because it is generally easier to interpret and understand graphics than values in a table. As such, graphical representation of data is an essential part of statistical analysis, from start to finish.

1.1.1 Accessing the Data

Assuming that a data collection process is completed, the "analysis" of this data is one of the next steps. This analysis is typically done in an appropriate software environment. There are many of such but our prime choice is R (R Core Team, 2020), often used alternatives are SPSS, SAS, Minitab, Stat, Prism besides other general purpose programming languages and platforms. Appendix A gives links to R and to some R resources. We assume now a running version of R.

The first step of the analysis is loading data in the software environment. This task sounds trivial and for pre-processed and readily available datasets often is. Cleaning own and others' data is typically very painful and eats up much unanticipated time. We do load external data but in this book we will not cover the aspect of data cleaning — be aware of this step when planning your analysis.

When storing own data it is recommended to save it in a tabular, comma-separated values format, typically called a CSV file. Similarly, the metadata should be provided in a separate text file. This ensures that the files are readable on "all" computing environments with open source software and remain so in the foreseeable future. The next two examples illustrate how to access data in R.

Example 1.1. There are many datasets available in R, the command `data()` lists all that are currently available on your computing system. Additional datasets are provided by many packages. Suppose the package `spam` is installed, e.g., by executing `install.packages("spam")`, then the function call `data(package="spam")` lists the datasets included in the package `spam`. A specific dataset is then loaded by calling `data()` with argument the name of the dataset (quoted or unquoted work both). (The command `data( package=.packages( all.available=TRUE))` would list all datasets from all R packages that are installed on your computing system.) ♣

Example 1.2. Often, we will work with own data and hence we have to "read-in" (or load or import) the data, e.g., with R functions `read.table()`, `read.csv()`. In R-Code 1.1 we read-in observations of mercury content in lake Geneva sediments. The data is available at www.math.uzh.ch/furrer/download/sta120/lemanHg.csv and is stored in a CSV file, with observation number and mercury content (mg/kg) on individual lines (Furrer and Genton, 1999). After importing the data, it is of utmost importance to check if the variables have been properly read, that (possible) row and column names are correctly parsed. Possible R functions for this task are `str()`, `head()`, `tail()`, or visualizing the entire dataset with `View()`. The number of observations and variables should be checked (if known), for example with `dim()` for matrices (a two-by-two arrangement of numbers) and dataframes (a handy tabular format of R), or `length()` for vectors or from the output of `str()`. Note that in subsequent examples we will not always display the output of all these verification calls to keep the R display to a reasonable length. We limit the output to pertinent calls such that the given R-Code can be understood by itself without the need to run it in an R session (although this is recommended).

R-Code 1.1 includes also a commented example code that illustrates how the format of the imported dataset changes when arguments of importing functions (here `read.csv()`) are not properly set (commented second but last line). ♣

R-Code 1.1 Loading the ‘*lemanHg*’ dataset in R.

```
Hg.frame <- read.csv('data/lemanHg.csv')
str( Hg.frame)      # dataframe with 1 numeric column and 293 observations
## 'data.frame': 293 obs. of  1 variable:
##  $ Hg: num  0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
head( Hg.frame, 3)  # column name is 'Hg'
##     Hg
## 1 0.17
## 2 0.21
## 3 0.06
Hg <- Hg.frame$Hg   # equivalent to `Hg.frame[,1]` or `Hg.frame[,"Hg"]`
str( Hg)            # now we have a vector
##  num [1:293] 0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
sum( is.na( Hg))    # check if there are NAs, alt: `any( is.na( Hg))`
## [1] 0
# str( read.csv('data/lemanHg.csv', header=FALSE))
            # Wrong way to import data. Result is a `factor` not `numeric`!
```

The tabular format we are using is typically arranged such that each *observation* (data for a specific location or subject or time point) that may consist of several *variables* or individual measurements (different chemical elements or medical conditions) is on one line. We call the entirety of these data points a *dataset*.

When analyzing data it is always crucial to query the origins of the data. In the following, we state some typical questions that we should ask ourselves as well as resulting consequences or possible implications and pitfalls. “What was the data collected for?”: The data may be of different quality if collected by scientists using it as their primary source compared to data used to counter ‘fake news’ arguments. Was the data rather observational or from a carefully designed experiment. The latter often being more representative and less prone to biases due to the chosen observation span. “Has the data been collected from different sources?”: This might imply heterogeneous data because the scientists or labs have used different standards, protocols or tools. In case of specific questions about the data, it may not be possible to contact all original data owners. “Has the data been recorded electronically?”: If electronic data recording has lower chances of erroneous entries. No human is perfect (especially reading off and entering numbers) and such datasets are more prone to contain errors compared to automatically recorded ones.

In the context of an educational data analysis, the following questions are often helpful “Is the data a classical textbook example?”: In such setting, we typically have cleaned data that serve to illustrate one (or at most a couple) pedagogical concepts. The *lemanHg* dataset used in this chapter is such a case, no negative surprises to be expected. “Has the data has been analyzed elsewhere before?”: such data have been typically cleaned and we already have one reference for an analysis. “Are the data stemming from simulated data?”: in such cases, there is a known underlying generation processes.

Of in all the cases mentioned before we can safely apply the proverb “the exception proves the rule”.

1.1.2 Types of Data

Presumably we all have a fairly good idea of what data is. However, often this view is quite narrow and boils down to *data are numbers*. But “data is not just data”: data can be hard or soft, quantitative or qualitative.

Hard data is associated with quantifiable statements like “The height of this female is 172 cm.” Soft data is often associated with subjective statements or fuzzy quantities requiring interpretation, such as “This female is tall”. Probability statements can be considered hard (derived from hard data) or soft (due to a lack of quantitative values). In this book, we are especially concerned with hard data.

An important distinction is whether data is qualitative or quantitative in nature. *Qualitative data* consists of categories and are either on *nominal scale* (e.g., male/female) or on *ordinal scale* (e.g., weak<average<strong, nominal with an ordering). *Quantitative data* is numeric and mathematical operations can be performed with it.

Quantitative data is either discrete, taking on only specific values (e.g., integers or a subset thereof), or continuous, taking on any value on the real number line. Quantitative data is measured on an *interval scale* or *ratio scale*. Unlike the ordinal scale, the interval scale is uniformly spaced. The ratio scale is characterized by a meaningful absolute zero in addition to the characteristics of the interval scale. Depending on the measurement scale, certain mathematical operators and thus summary measures or statistical measures are appropriate. The measurement scales are classified according to Stevens (1946) and summarized in Table 1.1. We will discuss the statistical measures based on data next and their theoretical counterparts and properties in later chapters.

Non-numerical data, often gathered from open-ended responses or in audio-visual form, is considered qualitative. We will not discuss such type of data here.

Example 1.3. The classification of elements as either “C” or “H” results in a nominal variable. If we associate “C” with *cold* and “H” with *hot* we can use an ordinal scale (based on temperature).

In R, nominal scales are represented with factors. R-Code 1.2 illustrates the creation of nominal and interval scales as well as some simple operations. It would be possible to create ordinal scales as well, but we will not use it in this book.

When measuring temperature in Kelvin (absolute zero at -273.15°C), a statement such as “The temperature has increased by 20%” can be made. However, a comparison of twice as hot (in degrees Celsius) does not make sense as the origin is arbitrary. ♣

Table 1.1: Types of scales according to Stevens (1946) and possible mathematical operations. The statistical measures are for a description of location and spread.

		Measurement scale			
		nominal	ordinal	interval	real
Mathematical operators		$=, \neq$	$=, \neq$ $<, >$	$=, \neq$ $<, >$ $-, +$	$=, \neq$ $<, >$ $-, +$ $\times, /$
Statistical measures	location	mode	mode median	mode median arithmetic mean	mode median arithmetic mean geometric mean
	spread			range standard deviation	range studentized range standard deviation coefficient of variation

R-Code 1.2 Example of creating ordinal and interval scales in R.

```
ordinal <- factor( c("male","female"))
ordinal[1] == ordinal[2]
## [1] FALSE
# ordinal[1] > ordinal[2]  # warning '>' not meaningful for factors
interval <- c(2, 3)
interval[1] > interval[2]
## [1] FALSE
```

1.2 Descriptive Statistics

With *descriptive statistics* we summarize and describe quantitatively various aspects and features of a dataset.

Although rudimentary information, very basic summary of the data is its size: number of observations, number of variables, and of course their types (see output of R-Code 1.1). Another important aspect is the number of missing values which deserve careful attention. In R a missing value is represented as `NA` in contrast to `NaN` (not a number) or `Inf` (for "infinity"). One has to evaluate if the missing values are due to some random mechanism, emerge consistently or with some deterministic pattern, appear in all variables, for example.

For a basic analysis one often neglects the observations if missing values are present in any variable. There exist techniques to fill in missing values but these are quite complicated and not treated here.

As a side note, with a careful inspection of missing values in ozone readings, the Antarctic "ozone hole" would have been discovered more than one decade earlier (see, e.g., en.wikipedia.org/wiki/Ozone_depletion#Research_history).

Informally a *statistic* is a single measure of some attribute of the data, in the context of this chapter a statistic gives a good first impression of the distribution of the data. Typical statistics for the *location* (i.e., the position of the data) include the sample mean, truncated/trimmed mean, sample median, sample quantiles and quartiles. The *trimmed mean* omits a small fraction of the smallest and the same small fraction of the largest values. A trimming of 50% is equivalent to the sample median. Sample quantiles or more specifically sample percentiles link observations or values with the position in the ordered data. For example, the sample median is the 50th-percentile, half the data is smaller than the median, the other half is larger. The 25th- and 75th-percentile are also called the lower and upper quartiles, i.e., the quartiles divide the data in four equally sized groups. Depending on the number of observations at hand, arbitrary quantiles are not precisely defined. In such cases, a linearly interpolated value is used, for which the precise interpolation weights depend on the software at hand. It is important to know this potential ambiguity less important to know the exact values of the weights.

Typical statistics for the *spread* (i.e., the dispersion of the data) include the *sample variance*, *sample standard deviation* (square root of the sample variance), *range* (largest minus smallest value), *interquartile range* (third quartile minus the first quartile), *studentized range* (range divided by the standard deviation, representing the range of the sample measured in units of sample standard deviations) and the *coefficient of variation* (sample standard deviation divided by the sample mean). Note that the studentized range and the coefficient of variance are dimension-less and should only be used with ratio scaled data.

We now introduce mathematical notation for several of these statistics. For a univariate dataset the observations are written as $x_1, \dots, x_n$ (or with some other latin letter), with n being the *sample size*. The ordered data (smallest to largest) is denoted with $x_{(1)} \leq \dots \leq x_{(n)}$. Hence, we use the following classical notation:

sample mean: $$\bar{x} = \sum_{i=1}^{n} x_i, \tag{1.1}$$

sample median: $$\operatorname{med}(x_i) = \begin{cases} x_{(n/2+1/2)}, & \text{if } n \text{ odd}, \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{if } n \text{ even}, \end{cases} \tag{1.2}$$

sample variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1.3}$$

sample standard deviation: $$s = \sqrt{s^2}. \tag{1.4}$$

If the context is clear, we may omit *sample*. Note that some books also use the quantifier *empirical* instead of sample. The symbols for the sample mean, sample variance and sample standard deviation are quite universal. However, this is not the case for the median. We use the notation $\operatorname{med}(x_1, \dots, x_n)$ or $\operatorname{med}(x_i)$ if no ambiguities exist.

Example 1.4 (continued from Example 1.2)**.** In R-Code 1.3 several summary statistics for 293 observations of mercury in lake Geneva sediments are calculated (see R-Code 1.1 to load the data). Of course all standard descriptive statistics are available as predefined functions in R. ♣

R-Code 1.3 A quantitative EDA of the '*lemanHg*' dataset.

```
c( mean=mean( Hg), tr.mean=mean( Hg, trim=0.1), median=median( Hg))
##    mean tr.mean  median
## 0.46177 0.43238 0.40000
c( var=var( Hg), sd=sd( Hg), iqr=IQR( Hg))   # capital letters for IQR!
##      var       sd      iqr
## 0.090146 0.300243 0.380000
summary( Hg)              # min, max, quartiles and mean
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.010   0.250   0.400   0.462   0.630   1.770
range( Hg)                # min, max, but not the difference
## [1] 0.01 1.77
tail( sort( Hg))          # sorts and then list the 6 largest values
## [1] 1.28 1.29 1.30 1.31 1.47 1.77
```

For discrete data, the *sample mode* is the most frequent value of a sample frequency distribution; in order to calculate the sample mode, only the operations $\{=, \neq\}$ are necessary. Continuous data are first divided into categories (by discretization or binning) and then the mode can be determined.

In subsequent chapters, we will discuss (statistical) properties of these different statistics. Then, it will be important to emphasize when we are referring to the sample mean or to the theoretical mean of our statistical model (also known as the expectation).

Another important aspect of EDA is the identification of outliers, which are defined (verbatim from Olea, 1991): "In a sample, any of the few observations that are separated so far in value from the remaining measurements that the questions arise whether they belong to a different population, or that the sampling technique is faulty. The determination that an observation is an outlier may be highly subjective, as there is no strict criteria for deciding what is and what is not an outlier". Graphical respresentations of the data often help in the identification of outliers, as we will see in the next section.

1.3 Visualizing Data

In the remaining part of the chapter, we discuss how to "graphically" present and summarize data. With graphically we mean a visualization using a graph, plot, chart or even a diagram. The exact definition and meaning of the latter four terms is not universal but a rough characterization is as follows. A *diagram* is the most generic term and is essentially a symbolic representation of information (Figure 1.1 being an example). A *chart* is a graphical representation of data (a pie chart being an example). A *plot* is a graphical representation of data in a coordinate system (we will discuss histograms, boxplots, scatterplots, and more below). Finally, a *graph* is "continuous" line representing data in a coordinate system (for example, visualizing a function).

Remark 1.1. There are several fundamentally different approaches to creating plots in R: base graphics (package *graphics*, which is automatically loaded upon starting R), trellis graphics (packages *lattice* and *latticeExtra*), and the *grammar of graphics* approach (package *ggplot2*). In this book we focus on base graphics. This approach is in sync with the R source code style and we have a clear direct handling of all elements. *ggplot* functionality may produce seemingly fancy graphics at the price of certain black box elements and more complex code structure. ♣

1.3.1 Graphically Represent Univariate Data

Univariate data is composed of one variable, or a single scalar component, measured at several instances or for several individuals or subjects. Graphical depiction of univariate data is usually accomplished with bar plots, histograms, box plots, or Q-Q plots among others. We now look at these examples in more details.

Ordinal data, nominal data and count data are often represented with bar plots (also called bar charts). The height of the bars is proportional to the frequency of the corresponding value. The German language discriminates between bar plots with vertical and horizontal orientation ('Säulendiagramm' and 'Balkendiagramm').

Example 1.5. R-Code 1.4 and Figure 1.2 illustrate bar plots with data giving aggregated CO_2 emissions from different sources (transportation, electricity production, deforestation, . . .) in the year 2005, as presented by the SWISS Magazine 10/2011-01/2012, page 107 (SWISS Magazine, 2011) and also shown in Figure 1.11. Here and later we may use a numerical specification of the colors which are, starting from one to eight, black, red, green, blue, cyan, magenta, yellow and gray.

Note that the values of such emissions vary considerably according to different sources, mainly due to the political and interest factors associated with these numbers. ♣

Histograms illustrate the frequency distribution of observations graphically and are easy to construct and to interpret (for a specified partition of the x-axes, called bins, observed counts or proportions are recorded). Histograms allow one to quickly assess whether the data is symmetric or rather *left-skewed* (more smaller values on the left side of the bulk of the data) or *right-skewed* (more larger values on the right side of the bulk of the data), whether the data is *unimodal* (has rather one dominant peak) or several or whether exceptional values are present. Several valid rules of thumb exist for choosing the optimal number of bins. However, the number of bins is a subjective choice that affects the look of the histogram, as illustrated in the next example.

R-Code 1.4 Representing emissions sources with barplots. (See Figure 1.2.)

```r
dat <- c(2, 15, 16, 32, 25, 10)    # see Figure 1.11
emissionsource <- c('Air', 'Transp', 'Manufac', 'Electr', 'Deforest', 'Other')
barplot( dat, names=emissionsource, ylab="Percent", las=2)
barplot( cbind('2005'=dat), col=c(2,3,4,5,6,7), legend=emissionsource,
      args.legend=list(bty='n'), ylab='Percent', xlim=c(0.2,4))
```

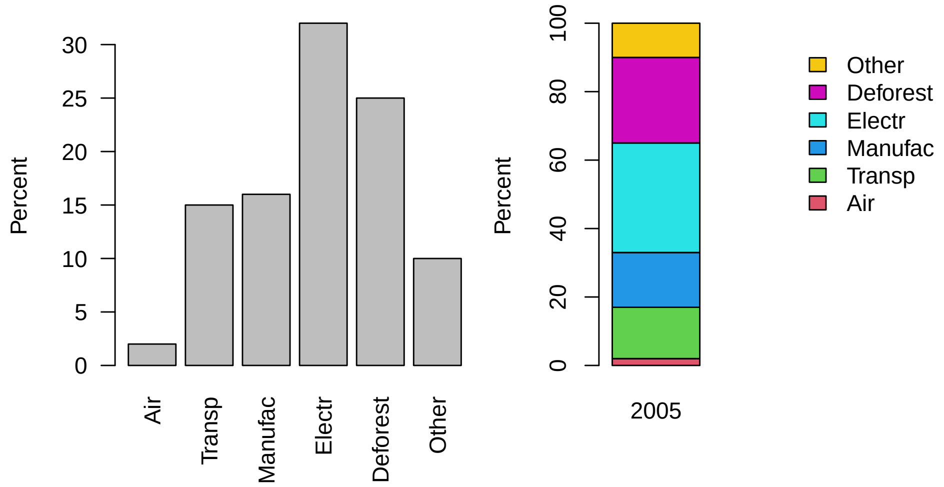


Figure 1.2: Bar plots: juxtaposed bars (left), stacked (right) of CO_2 emissions according to different sources taken from SWISS Magazine (2011). (See R-Code 1.4.)

Example 1.6 (continued from Examples 1.2 and 1.4). R-Code 1.5 and the associated Figure 1.3 illustrate the construction and resulting histograms of the mercury dataset. The different choices illustrate that the number of bins needs to be carefully chosen. Often, the default values work well, which is also the case here. In one of the histograms, a "smoothed density" has been superimposed. Such curves will be helpful when comparing the data with different statistical models, as we will see in later chapters. When adding smoothed densities, the histogram has to represent the fraction of data in each bin (argument `probability=TRUE`) as opposed to the frequency (the default).

The histograms in Figure 1.3 shows that the data is unimodal, right-skewed, no exceptional values (no outliers). Important statistics like mean and median can be added to the plot with vertical lines. Due to the slight right-skewedness, the mean is slightly larger than the median.

♣

When constructing histograms for discrete data (e.g., integer values), one has to be careful with the binning. Often it is better to manually specify the bins. To represent the result of several dice tosses, it would be advisable to use `hist( x, breaks=seq( from=0.5, to=6.5, by=1))`, or possibly use a bar plot as explained above. A *stem-and-leaf plot* is similar to a histogram, in which each individual observation is marked with a single dot (`stem()` in R). Although this plot gives the most honest representation of data, it is rarely used today, mainly beacuse it does not work for very large sample sizes. Figure 1.4 gives an example.

R-Code 1.5 Different histograms (good and bad ones) for the '*lemanHg*' dataset. (See Figure 1.3.)

```
histout <- hist( Hg)     # default histogram
hist( Hg, col=7, probability=TRUE, main="With 'smoothed density'")
lines( density( Hg))     # add smooth version of the histogram
abline( v=c(mean( Hg), median( Hg)), col=3:2, lty=2:3, lwd=2:3)
hist( Hg, col=7, breaks=90, main="Too many bins")
hist( Hg, col=7, breaks=2, main="Too few bins")
str( histout[1:3])       # Contains essentially all information of histogram
## List of 3
##  $ breaks : num [1:10] 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8
##  $ counts : int [1:9] 58 94 61 38 26 9 5 1 1
##  $ density: num [1:9] 0.99 1.604 1.041 0.648 0.444 ...
```

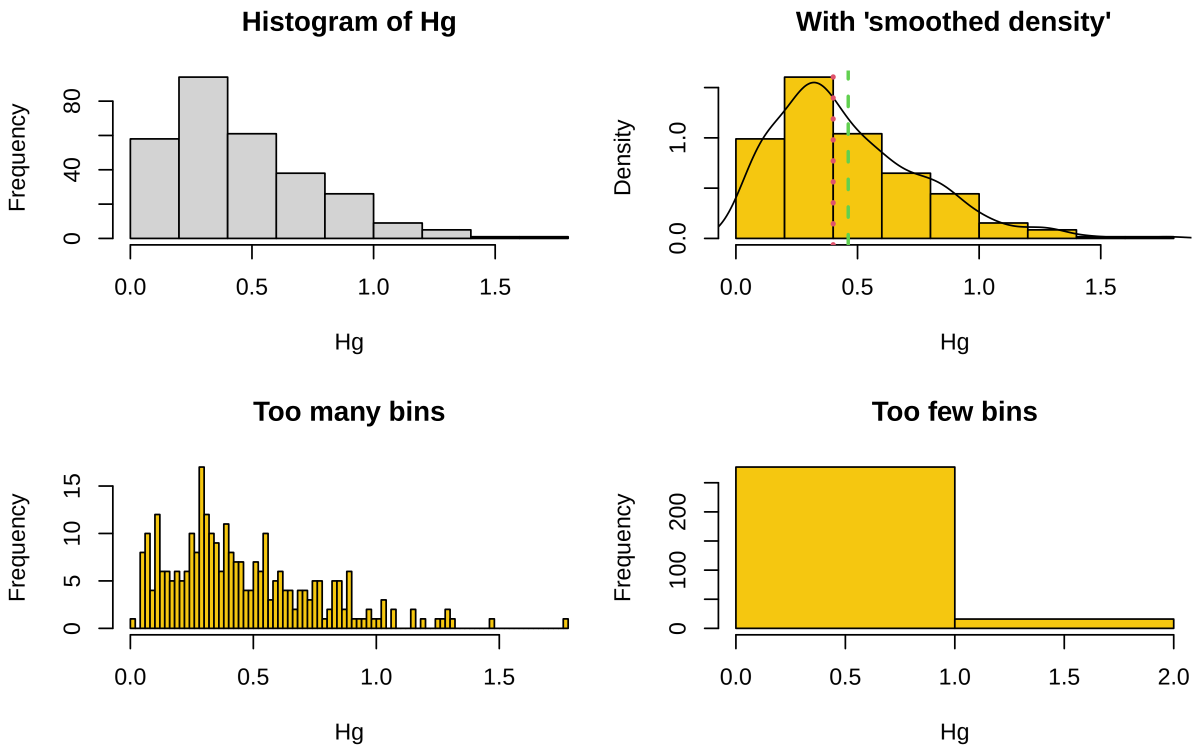
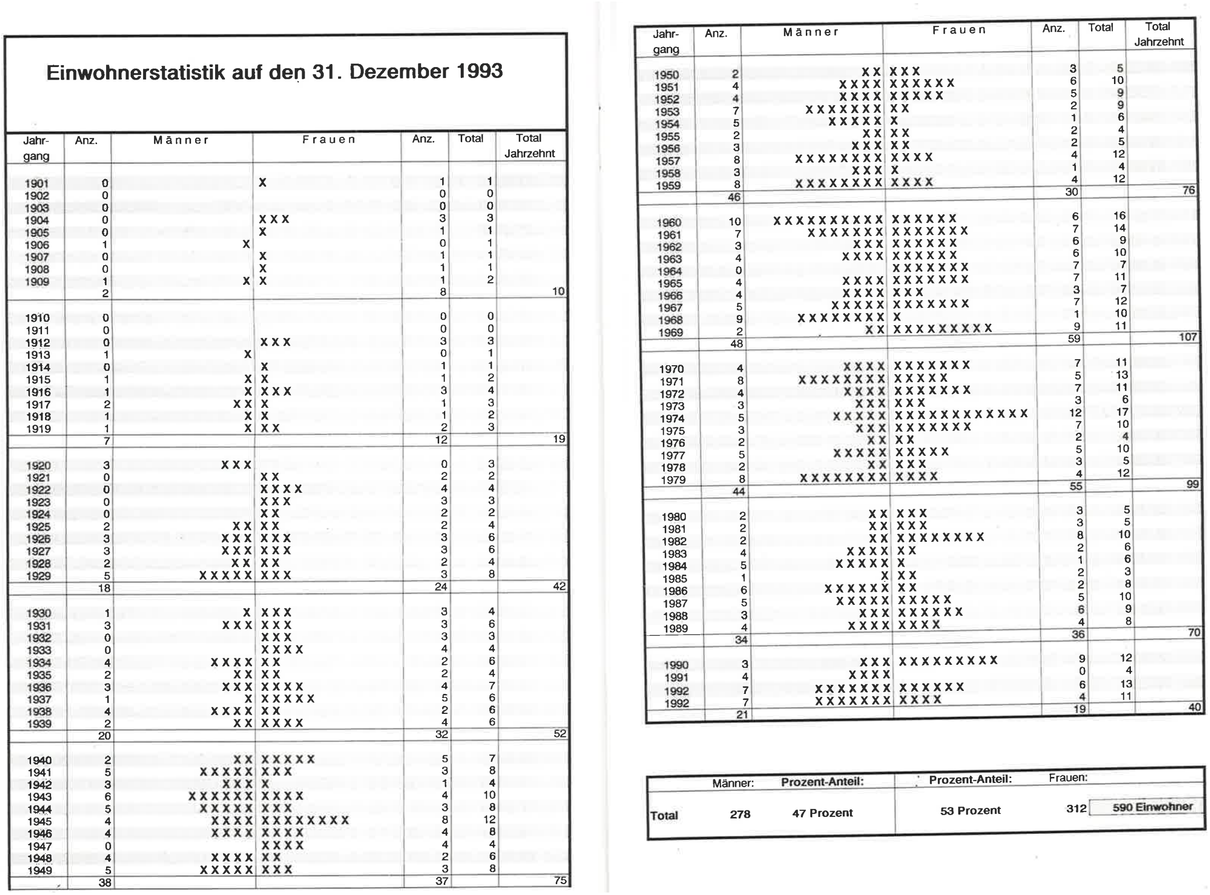


Figure 1.3: Histograms of mercury data with various bin sizes. In the top left panel, the smoothed density is in back, the mean and median are in green dashed and red dotted vertical lines, respectively. (See R-Code 1.5.)

A *boxplot* or a *box-and-wishker plot* is a graphical representation of five location statistics of the observations: median, the lower and upper quartiles and the minimum and maximum values. A box ranging extending to both quartiles contains thus half of the data values. Typically, all points smaller than the first quartile minus 1.5·IQR and larger than the third quartile plus



Einwohnerstatistik auf den 31. Dezember 1993

Jahrgang	Anz.	Männer	Frauen	Anz.	Total	Total Jahrzehnt
1901	0		X	1	1	
1902	0			0	0	
1903	0			0	0	
1904	0		XXX	3	3	
1905	0		X	1	1	
1906	1	X		0	1	
1907	0		X	1	1	
1908	0		X	1	1	
1909	1	X	X	1	2	
	2			8		10
1910	0			0	0	
1911	0			0	0	
1912	0		XXX	3	3	
1913	1	X		0	1	
1914	0		X	1	1	
1915	1	X	X	1	2	
1916	1	X	XXX	3	4	
1917	2	XX	X	1	3	
1918	1	X	X	1	2	
1919	1	X	XX	2	3	
	7			12		19
1920	3	XXX		0	3	
1921	0		XX	2	2	
1922	0		XXXX	4	4	
1923	0		XXX	3	3	
1924	0		XX	2	2	
1925	2	XX	XX	2	4	
1926	3	XXX	XXX	3	6	
1927	3	XXX	XXX	3	6	
1928	2	XX	XX	2	4	
1929	5	XXXXX	XXX	3	8	
	18			24		42
1930	1	X	XXX	3	4	
1931	3	XXX	XXX	3	6	
1932	0		XXX	3	3	
1933	0		XXXX	4	4	
1934	4	XXXX	XX	2	6	
1935	2	XX	XX	2	4	
1936	3	XXX	XXXX	4	7	
1937	1	X	XXXXX	5	6	
1938	4	XXXX	XX	2	6	
1939	2	XX	XXXX	4	6	
	20			32		52
1940	2	XX	XXXXX	5	7	
1941	5	XXXXX	XXX	3	8	
1942	3	XXX	X	1	4	
1943	6	XXXXXX	XXXX	4	10	
1944	5	XXXXX	XXX	3	8	
1945	4	XXXX	XXXXXXXX	8	12	
1946	4	XXXX	XXXX	4	8	
1947	0		XXXX	4	4	
1948	4	XXXX	XX	2	6	
1949	5	XXXXX	XXX	3	8	
	38			37		75
1950	2	XX	XXX	3	5	
1951	4	XXXX	XXXXXX	6	10	
1952	4	XXXX	XXXXX	5	9	
1953	7	XXXXXXX	XX	2	9	
1954	5	XXXXX	X	1	6	
1955	2	XX	XX	2	4	
1956	3	XXX	XX	2	5	
1957	8	XXXXXXXX	XXXX	4	12	
1958	3	XXX	X	1	4	
1959	8	XXXXXXXX	XXXX	4	12	
	46			30		76
1960	10	XXXXXXXXXX	XXXXXX	6	16	
1961	7	XXXXXXX	XXXXXXX	7	14	
1962	3	XXX	XXXXXX	6	9	
1963	4	XXXX	XXXXXX	6	10	
1964	0		XXXXXXX	7	7	
1965	4	XXXX	XXXXXXX	7	11	
1966	4	XXXX	XXX	3	7	
1967	5	XXXXX	XXXXXXX	7	12	
1968	9	XXXXXXXXX	X	1	10	
1969	2	XX	XXXXXXXXX	9	11	
	48			59		107
1970	4	XXXX	XXXXXXX	7	11	
1971	8	XXXXXXXX	XXXXX	5	13	
1972	4	XXXX	XXXXXXX	7	11	
1973	3	XXX	XXX	3	6	
1974	5	XXXXX	XXXXXXXXXXXX	12	17	
1975	3	XXX	XXXXXXX	7	10	
1976	2	XX	XX	2	4	
1977	5	XXXXX	XXXXX	5	10	
1978	2	XX	XXX	3	5	
1979	8	XXXXXXXX	XXXX	4	12	
	44			55		99
1980	2	XX	XXX	3	5	
1981	2	XX	XXX	3	5	
1982	2	XX	XXXXXXXX	8	10	
1983	4	XXXX	XX	2	6	
1984	5	XXXXX	X	1	6	
1985	1	X	XX	2	3	
1986	6	XXXXXX	XX	2	8	
1987	5	XXXXX	XXXXX	5	10	
1988	3	XXX	XXXXXX	6	9	
1989	4	XXXX	XXXX	4	8	
	34			36		70
1990	3	XXX	XXXXXXXXX	9	12	
1991	4	XXXX		0	4	
1992	7	XXXXXXX	XXXXXX	6	13	
1992	7	XXXXXXX	XXXX	4	11	
	21			19		40

	Männer:	Prozent-Anteil:	Prozent-Anteil:	Frauen:	
Total	278	47 Prozent	53 Prozent	312	590 Einwohner

Figure 1.4: Stem-and-leaf plot of residents of the municipality of Staldenried as of 31.12.1993 (Gemeinde Staldenried, 1994).

1.5·IQR are marked individually. The closest non-marked observations to these bounds are called the whiskers.

A *violin plot* combines the advantages of the box plot and a "smoothed" histogram by essentially merging both. Compared to a boxplot a violin plot depicts possible multi-modality and for large datasets de-emphasizes the marked observations outside he whiskers (often termed "outliers" but see our discussion in Chapter 7).

Example 1.7 (continued from Examples 1.2, 1.4 and 1.6)**.** R-Code 1.6 and Figure 1.5 illustrates the boxplot and violin plot for the mercury data. Due to the right-skewedness of the data, there are several data points beyond the upper whisker. Notice that the function `boxplot()` has several arguments for tailoring the appearance of the box plots. These are discussed in the function's help file. ♣

A quantile-quantile plot (*QQ-plot*) is used to visually compare the ordered sample, also called sample quantiles, with the quantiles of a theoretical distribution. For the moment we think of this theoretical distribution in form of a "smoothed density" similar to the superimposed line in the top right panel of Figure 1.3. We will talk more about "theoretical distributions" in the next two chapters. The theoretical quantiles can be thought of as the n values of a "perfect" realization. If the points of the QQ-plot are aligned along a straight line, then there is a good

R-Code 1.6 Boxplot and violin plot for the ‘*lemanHg*’ dataset. (See Figure 1.5.)

```
out <- boxplot( Hg, col="LightBlue", ylab="Hg", outlty=1, outpch='')
  # 'out' contains numeric values of the boxplot
quantile( Hg, c(0.25, 0.75))  # compare with summary( Hg) and out["stats"]
##  25%  75%
## 0.25 0.63
IQR( Hg)                      # interquartile range
## [1] 0.38
quantile(Hg, 0.75) + 1.5 * IQR( Hg)          # upper boundary of the whisker
## 75%
## 1.2
Hg[ quantile(Hg, 0.75) + 1.5 * IQR( Hg) < Hg] # points beyond the whisker
## [1] 1.25 1.30 1.47 1.31 1.28 1.29 1.77
require(vioplot)                          # R package providing violin plots
vioplot( Hg, col="Lightblue", ylab="Hg")
```

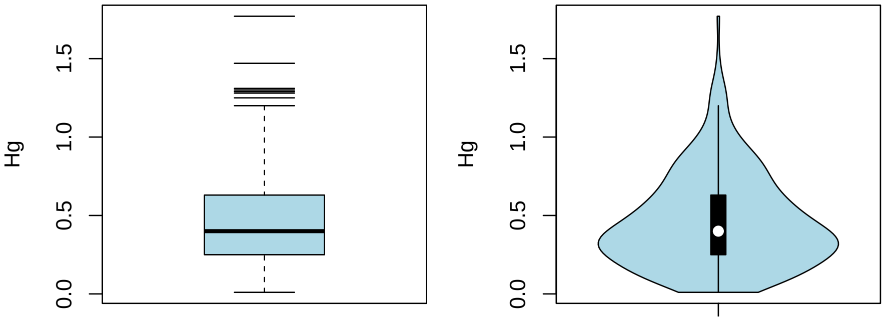


Figure 1.5: Box plots and violin plot for the ‘*lemanHg*’ dataset. (See R-Code 1.6.)

match between the sample and theoretical quantiles. A deviation of the points indicate that the sample has either too few or too many small or large points. Figure 1.6 illustrates a QQ-plot based on nine data points (for the ease of understanding) with quantiles from a symmetric (a) and a skewed (b) distribution respectively. The last panel shows the following cases with respect to the reference quantiles: (c1) sample has much smaller values than expected; (c2) sample has much larger values than expected; (c3) sample does not have as many small values as expected; (c4) sample does not have as many large values as expected.

Notice that the QQ-plot is invariant with respect to changing the location or the scale of the sample or of the theoretical quantiles. Hence, the omission of the scales in the figure.

Example 1.8 (continued from Examples 1.2, 1.4 to 1.7)**.** R-Code 1.7 and Figure 1.7 illustrate a QQ-plot for the mercury dataset by comparing it to a “bell-shaped” and a right-skewed distribution. The two distributions are called normal or Gaussian distribution and chi-squared distribution and will be discussed in subsequent chapters. To “guide-the-eye”, we have added the a line passing through the lower and upper quartile. The right panel shows a suprisingly good

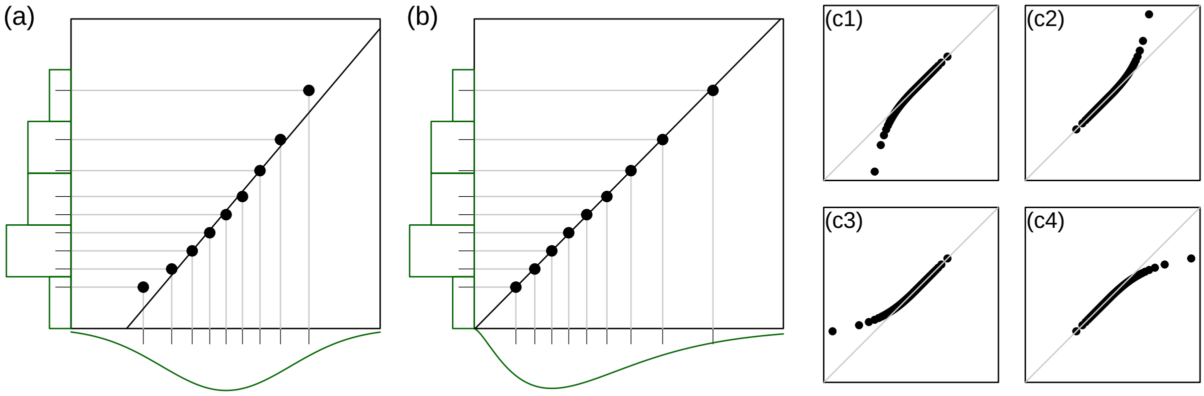


Figure 1.6: (left) QQ-plots with a symmetric (a) and a right-skewed (b) theoretical distribution. The histogram of the data and the theoretical density are shown in the margins in green. (c): schematic illustration for the different types of deviations. See text for an interpretation.

fit. ♣

R-Code 1.7 QQ-plot of the '*lemanHg*' dataset. (See Figure 1.7.)

```
qqnorm( Hg)              # QQplot with comparing with bell-shaped theoretical
qqline( Hg, col=2, main='')  # add read line
theoQuant <- qchisq( ppoints( 293), df=5) # minor mystery for the moment
# hist( theoQuant, prob=TRUE); lines(density(theoQuant)) # convince yourself
qqplot( theoQuant, Hg, xlab="Theoretical quantiles")
# For 'chisq' some a priori knowledge was used, for 'df=5' minimal
# trial and error was used.
qqline( Hg, distribution=function(p) qchisq( p, df=5), col=2)
```

1.3.2 Visualizing Multivariate Data

Multivariate data means that two or more variables are collected for each observation and are of interest. Additionally to the univariate EDA applied for each variable, visualization of multivariate data is often accomplished with scatterplots (*plot(x,y)* for two variables or *pairs()* for several) so that the relationship between pairs of variables is illustrated. For three variables, an interactive visualization based on the function *plot3d()* from the package *rgl* might be helpful.

In a scatterplot, "guide-the-eye" lines are often included. In such situation, some care is needed as there is an perception of asymmetry between y versus x and x versus y. We will discuss this further in Chapter 9.

In the case of several frequency distributions, bar plots, either stacked or grouped, may also be used in an intuitive way. See R-Code 1.8 and Figure 1.8 for two slightly different par-

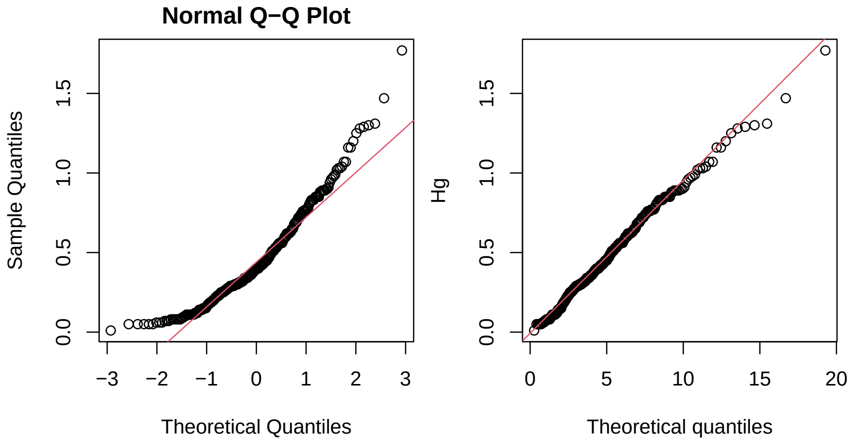


Figure 1.7: QQ-plots using the normal distribution (left) and a so-called chi-squared distribution with five degrees of freedom (right). The red line passes through the lower and upper quantiles of both the sample and theoretical distribution. (See R-Code 1.7.)

titions of CO_2 emission sources, based on (SWISS Magazine, 2011) and www.c2es.org/facts-figures/international-emissions/sector (approximate values).

R-Code 1.8 Bar plots for emissions by sectors for the year 2005 from two sources. (See Figure 1.8.)

```
dat2 <- c(2, 10, 12, 28, 26, 22) # source c2es.org
mat <- cbind( SWISS=dat, c2es.org=dat2)
barplot(mat, col=c(2,3,4,5,6,7), xlim=c(0.2,5), legend=emissionsource,
        args.legend=list(bty='n'), ylab='Percent', las=2)
barplot(mat, col=c(2,3,4,5,6,7), xlim=c(1,30), legend=emissionsource,
        args.legend=list(bty='n'), ylab='Percent', beside=TRUE, las=2)
```

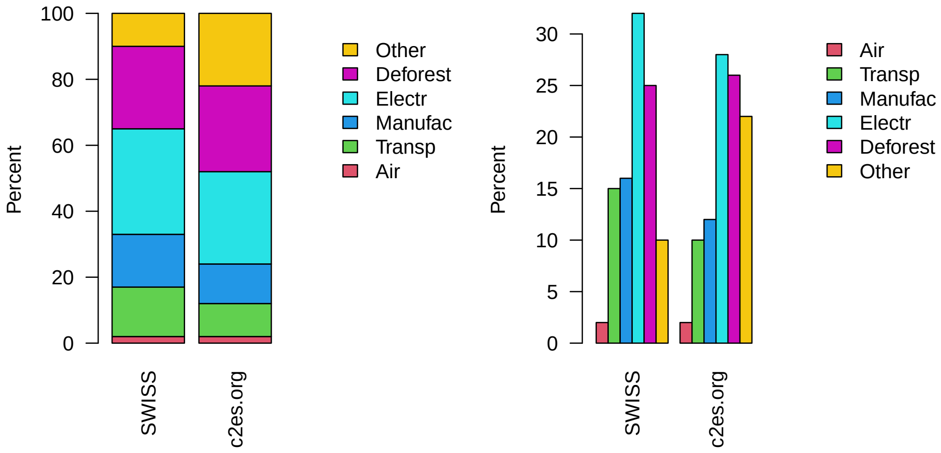


Figure 1.8: CO_2 emissions by sectors visualized with bar plots for the year 2005 from two sources (left: stacked, right: grouped). (See R-Code 1.8.)

Example 1.9. In this example we look at a more complex dataset containing several measurements from different penguin species collected over three years on three different islands. The package *palmerpenguins* provides the CSV file called `penguins.csv` in the package directory `extdata`. The dataset provides bill length, bill depth, flipper length, (all in millimeters) and body mass (in gram) of three penguin species *Adelie*, *Chinstrap*, *Gentoo*, next to their habitat island, sex and the year when the penguin has been sampled (Horst *et al.*, 2020).

R-Code 1.9 loads the data, performs a short EDA and provides the code for some representative figures. We deliberately use more arguments for several plotting functions to highlight different features in the data or of the R functions. The eight variables are on the nominal (e.g., *species*), interval (*year*) and real scale (e.g., *bill_length_mm*). Two of the length measurements have been rounded to integers (*flipper_length_mm* and *body_mass_g*). For two out of the 344 penguins we do not have any measures and sex is missing for nine additional penguins (*summary(penguins$sex)*). Due to their natural habitat, not all penguins have been observed on all three islands, as shown by the cross-tabulation with *table()*.

Figure 1.9 shows some representative graphics. The first two panels show that, overall, flipper length is bimodal with a more pronounced first mode. When we partition according to the species, we notice that the first mode corresponds species Adelie and Chinstrap, the latter slightly larger. Comparing the second mode with the violin plot of species Gentoo (middle panel) we notice that some details in the violin plot are "smoothed" out.

The third panel shows a so-called *mosaic plot*, were we represent a two dimensional frequency table, i.e., specifically a visualization of the aforementioned *table(...)* call. Here, in such a mosaic plot, the width of each vertical band corresponds to the proportion of penguins on the corresponding island. Each band is then divided according to the species proportion. For example, we quickly realize that Adeline is home on all three islands and that on Torgersen island the smallest sample has been taken. Empty cells are indicated with dashed lines. A mosaic plot is a qualitative assessment only and the areas of the rectangles are harder to compare than the corresponding numbers of the equivalent *table()* output.

The final plot is a scatterplot of the four length measures. Figure 1.9 shows that for some variables specific species cluster well (e.g., Gentoo with flipper length), whereas other variables are less separated (e.g., Adelie and Chinstrap with bill depth). To quickly increase information content, we use different colors or plotting symbols according to the ordinary variables. With the argument *row1attop=FALSE* we obtain a fully symmetric representation that allows quick comparisons between both sides of the diagonal. Here, with a black-red coloring we see that male penguins are typically larger. In fact, there is even a pretty good separation between both sexes. Some care is needed that the annotations are not overdone: unless there is evident clustering more than three different symbols or colors are often too much. R-Code 1.9 shows how to redefine the different panels of a scatterplot (see *help(pairs)* for a more professional panel definition). By properly specifying the *diag.panel* argument, it possible to add charts or plots to the diagonal panels of *pairs()*, e.g., the histograms of the variables. ♣

R-Code 1.9: Constructing histograms, boxplots, violin plots and scatterplot with *penguing* data. (See Figure 1.9.)

```
require(palmerpenguins)
penguins <- read.csv(       # `path.package()` provides local location of pkg
        paste0( path.package('palmerpenguins'),'/extdata/penguins.csv'))
str(penguins, strict.width='cut')

## 'data.frame': 344 obs. of  8 variables:
##  $ species          : chr  "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ island           : chr  "Torgersen" "Torgersen" "Torgersen" "Torgers"..
##  $ bill_length_mm   : num  39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42..
##  $ bill_depth_mm    : num  18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2..
##  $ flipper_length_mm: int  181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int  3750 3800 3250 NA 3450 3650 3625 4675 3475 42..
##  $ sex              : chr  "male" "female" "female" NA ...
##  $ year             : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 ..

summary(penguins[, c(3:6)])  # others variables can be summarized by `table()`

##  bill_length_mm bill_depth_mm  flipper_length_mm  body_mass_g
##  Min.   :32.1   Min.   :13.1   Min.   :172       Min.   :2700
##  1st Qu.:39.2   1st Qu.:15.6   1st Qu.:190       1st Qu.:3550
##  Median :44.5   Median :17.3   Median :197       Median :4050
##  Mean   :43.9   Mean   :17.2   Mean   :201       Mean   :4202
##  3rd Qu.:48.5   3rd Qu.:18.7   3rd Qu.:213       3rd Qu.:4750
##  Max.   :59.6   Max.   :21.5   Max.   :231       Max.   :6300
##  NA's   :2      NA's   :2      NA's   :2         NA's   :2

#penguins <- na.omit(penguins) # remove NA's
table(penguins[,c(2,1)])     # tabulating species on each island

##            species
## island      Adelie Chinstrap Gentoo
##   Biscoe        44         0    124
##   Dream         56        68      0
##   Torgersen     52         0      0

hist( penguins$flipper_length_mm, main='', xlab="Flipper length [mm]", col=7)
box()                        # rectangle around for similar view with others
with(penguins,  vioplot(flipper_length_mm[species=="Gentoo"],
      flipper_length_mm[species=="Chinstrap"], flipper_length_mm[
      species=="Adelie"],  names=c("Gentoo", "Chinstrap", "Adelie"),
      col=5:3, xlab="Flipper length [mm]", horizontal=TRUE, las=1))

mosaicplot( table( penguins[,c(2,1)]), col=3:5, main='', cex.axis=.75, las=1)

upper.panel <- function( x,y, ...)   # see `?pairs` for a better way
  points( x, y, col=as.numeric(as.factor(penguins$species))+2,...)
lower.panel <- function( x,y, ...)
```

```
  points( x, y, col=as.numeric(as.factor(penguins$sex)))
pairs( penguins[,c(3:6)], gap=0, row1attop=FALSE,    # scatterplot
     lower.panel=lower.panel, upper.panel=upper.panel)
 #  pch=as.numeric(as.factor(penguins$island)) # clutters & is not helpful
```

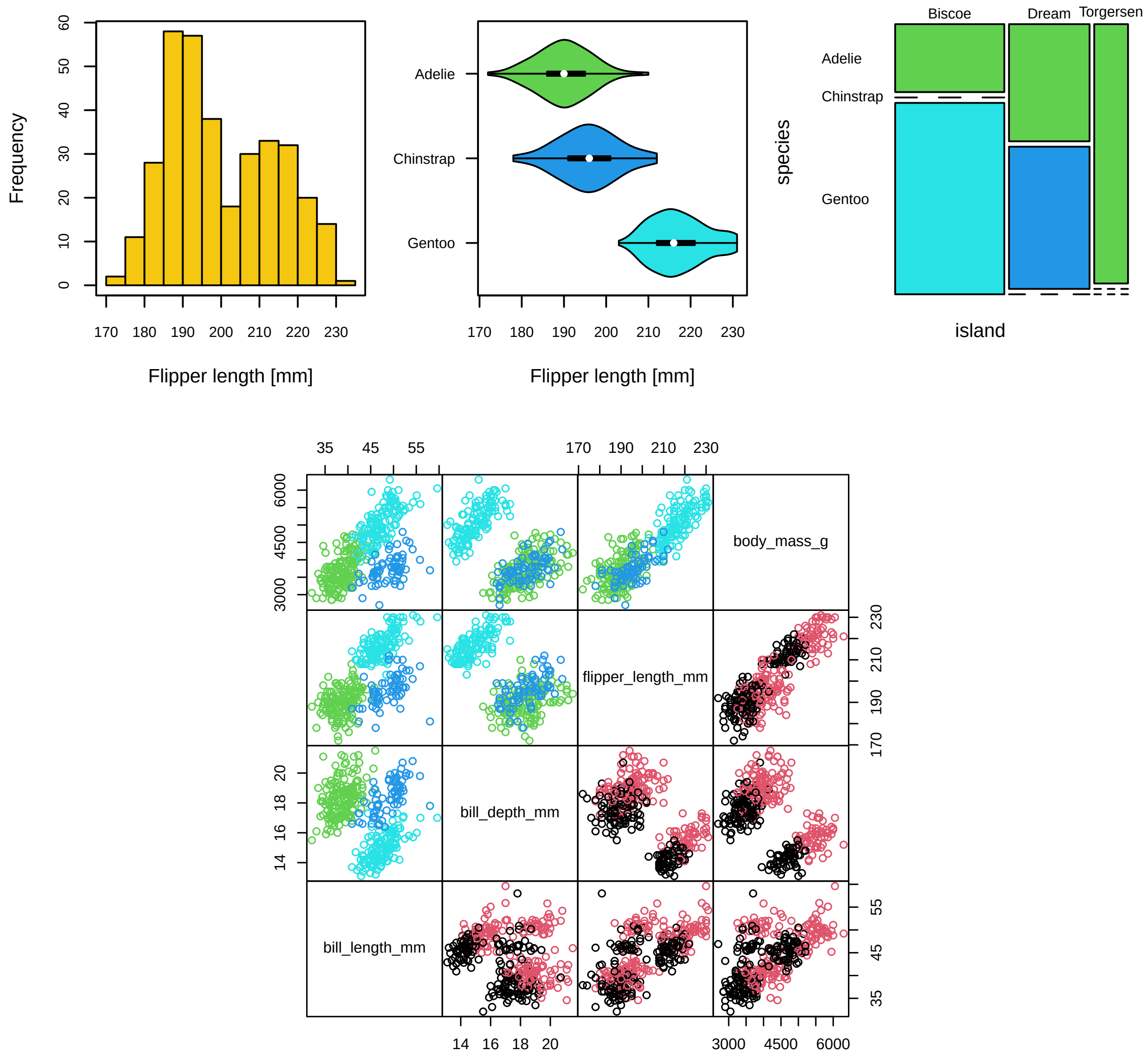


Figure 1.9: Visualization of `palmerpenguins` data. Top left to right: histogram for the variable flipper length, violin plots for flipper length stratified according to species and mosaic plot for species on each island (green: Adeline; blue Chinstrap; cyan: Gentoo). Bottom: scatterplot matrix with colors for species (above the diagonal, colors as above) and sex (below the diagonal, black female, read male). (See R-Code 1.9.)

For a dozen and more variables, scatterplots are no longer helpful as there are too many panels and nontrivial underlying structures are hardly visible. *Parallel coordinate plots* are a popular way of representing observations in high dimensions. Each variable is recorded along a vertical axis and the values of each observation are then connected with a line across the various variables. That means that points in the usual (Euclidean) representation correspond to lines in a parallel coordinate plot. In a classical version of the plot, all interval scaled variables are normalized to $[0, 1]$. Additionally, the inclusion of nominal and ordinal variables and their comparison with interval scaled variables is possible. The natural order of the variables in such a plot may not be optimal for an intuitive interpretation.

Example 1.10. The dataset *swiss* (provided by the package *datasets*) contains 6 variables (standardized fertility measure and socio-economic indicators) for 47 French-speaking provinces of Switzerland at about 1888.

R-Code 1.10 and Figure 1.10 give an example of a parallel coordinate plot. The lines are plotted according to the percentage of practicing catholics (the alternative being protestant). Groups can be quickly detected and strong associations are spotted directly. For example, provices with more catholics have a higher fertility rate or lower rates on examination (indicated by color grouping). Provinces with higher fertility have also higher values in agriculture (lines are in general parallel) whereas higher agriculture is linked to lower examination (lines typically cross). We will revisit such associations in Chapter 9. ♣

R-Code 1.10 Parallel coordinate plot for the swiss dataset. (See Figure 1.10.)

```
dim( swiss)   # in   package:datasets, available without the need to load.
## [1] 47  6
# str( swiss, strict.width='cut')    # or even:
# head( swiss);  summary( swiss)
require( MASS)    # package providing the function `parcoord()`
parcoord( swiss, col=2-(swiss$Catholic<40) + (swiss$Catholic>60))
```

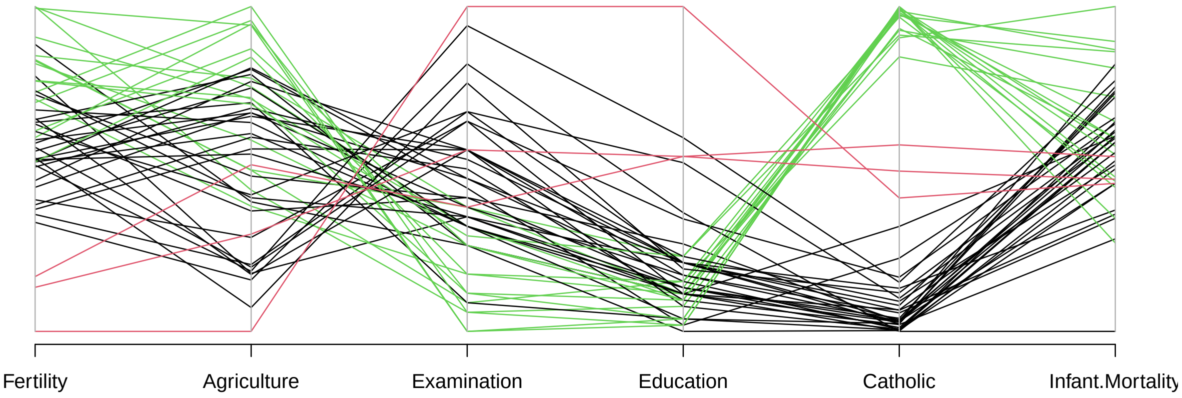


Figure 1.10: Parallel coordinate plot of the swiss dataset. The provinces (lines) are colored according to the percentage catholic (black less than 40%, red between 40% and 60%, green above 60%). (See R-Code 1.10.)

An alternative way to represent high-dimensional data is to project the data to two or three dimensions, which can be represented with classical visualization tools. The basic idea of *projection pursuit* is to find a projection, which highlights an interesting structure of the dataset (Friedman and Tukey, 1974). These projections are often varied and plotted continuously to find as many possible structures, see Problem 1.7 for a guided example.

1.4 Constructing Good Graphics

A good graphic should immediately convey the essence without distraction and influential elements. Hence, when creating a figure, one should start thinking what is the message the figure is conveying. Besides this fundamental rule additional basic guidelines of graphics and charts are:

- Construct honest graphs without hiding facts. Show all data, not omitting some observations or hiding information through aggregation. In R, it is possbile to construct a boxplot based on three values, but such a representation would suggest to the reader the availability of a larger amount of data. Thus hiding the fact that only three values are present.
- Construct graphs that are not suggestive. A classical deceptive example is to choose the scale such that small variations are emphasized (zooming in) or the variations are obscured by the large scale (zooming out). See bottom panel of Figure 1.11 where by starting the y-axis at 3 a stronger decrease is suggested.
- Use appropriate and unambiguous labels with units clearly indicated. For presentations and publications the labels have to be legible. See top left panel of Figure 1.11 where we do not know the units.
- To compare quantities, one should directly represent ratios, differences or relative differences. When different panels need to be compared, they should have the same plotting range.
- Carefully choose colors, hues, line width and type, symbols. These need to be explained in the caption or a legend. Certain colors are more dominant or are directly associated with emotions. Be aware of color blindness
- Never use three-dimensional renderings of lines, bars etc. The human eye can quickly compare lengths but not angles, areas, or volumes.
- Never change scales mid axis. If absolutely necessary a second scale can be added on the right or on the top.

Of course there are situations where it makes sense to deviate from individual bullets of the list above. In this book, we carefully design our graphics but in order to keep the R-code to reasonable length, we bend the above rules quite a bit.

From a good-scientific-practice point of view we recommend that figures should be constructed such that they are “reproducible”. To do so

- create figures based on a script. The figures in this book can be reconstructed based on sourcing the dedicated R scripts.
- do no post-process figures with graphics editors (e.g., with PhotoShop, ImageMagick)

- as some graphics routines are based on random numbers, initiate the random number generator before the construction (use `set.seed()`, see also Chapter 2).

Do not use pie charts unless absolutely necessary. Pie charts are often difficult to read. When slices are similar in size it is nearly impossible to distinguish which is larger. Bar plots allow an easier comparison, compare Figure 1.11 with either panel of Figure 1.2.

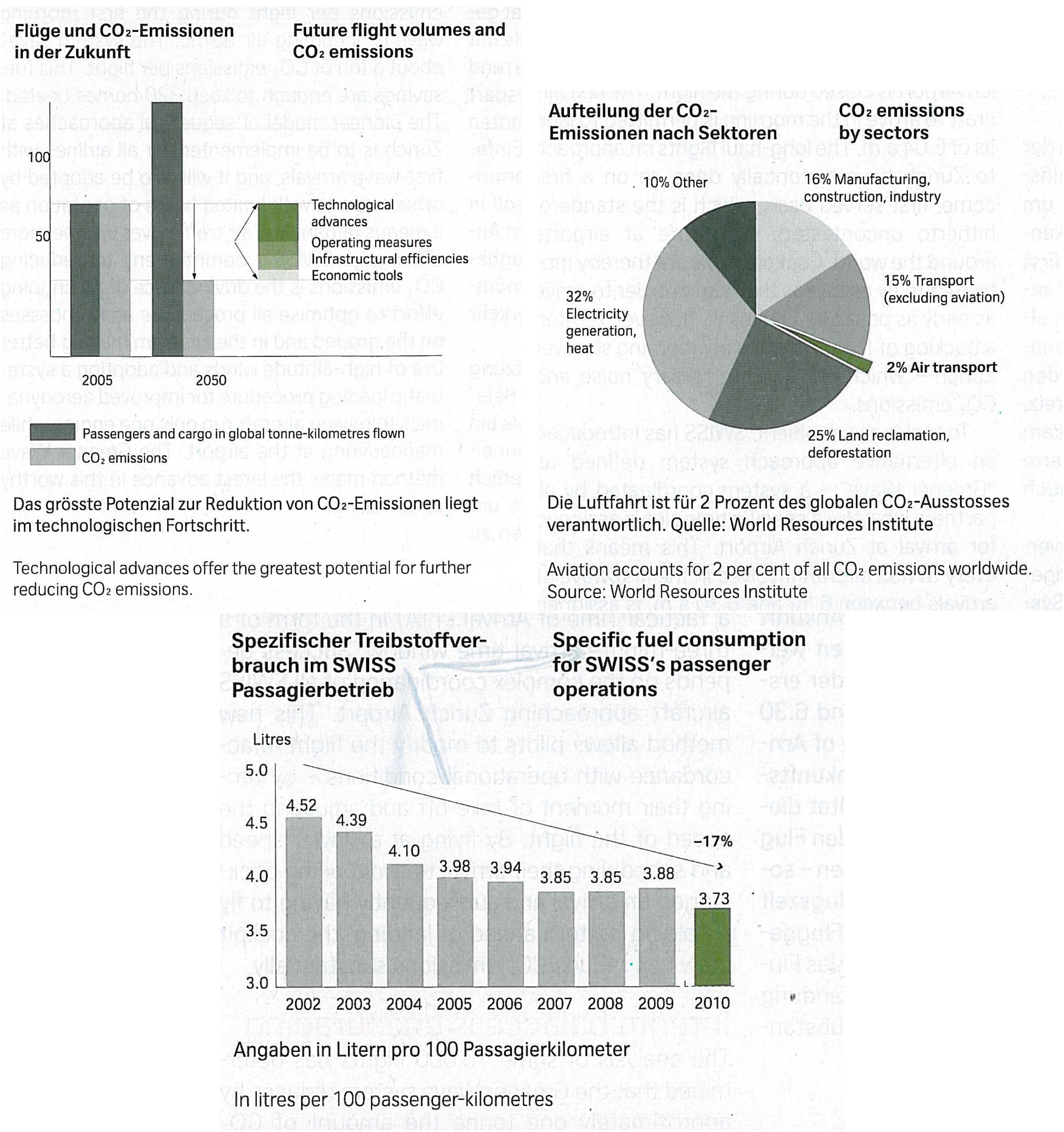


Figure 1.11: SWISS Magazine 10/2011-01/2012, 107.

Figures 1.12 and 1.11 are examples of badly designed charts and plots. The top panel Figure 1.12 contains an unnecessary 3-D rendering. The lower panel of the figure is still suboptimal because of the shading not all information is visible. Depending on the intended message, a plot instead of a graph would be more adequate.

Bad graphics can be found everywhere including in scientific journals. Figure 1.13 is a snapshot of the webpage http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/ that includes the issues of the figures and a discussion of possible improvements.

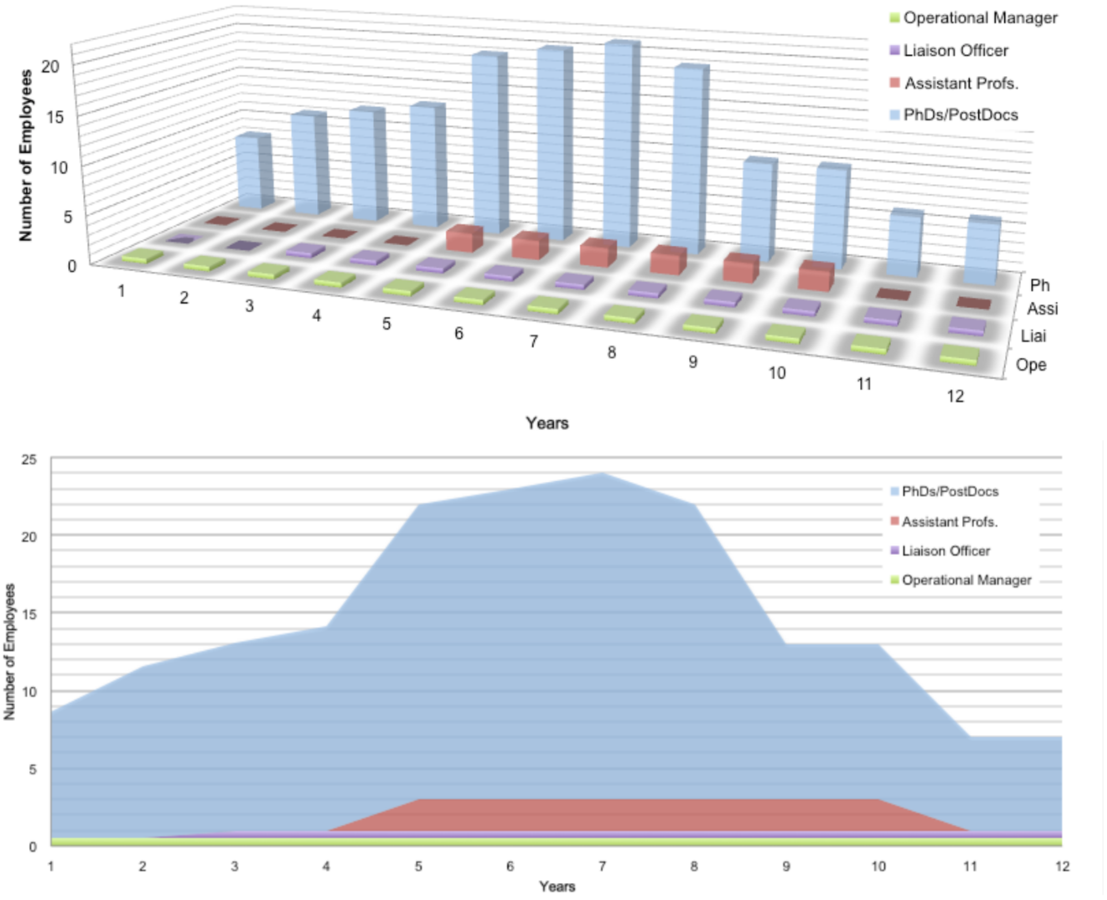


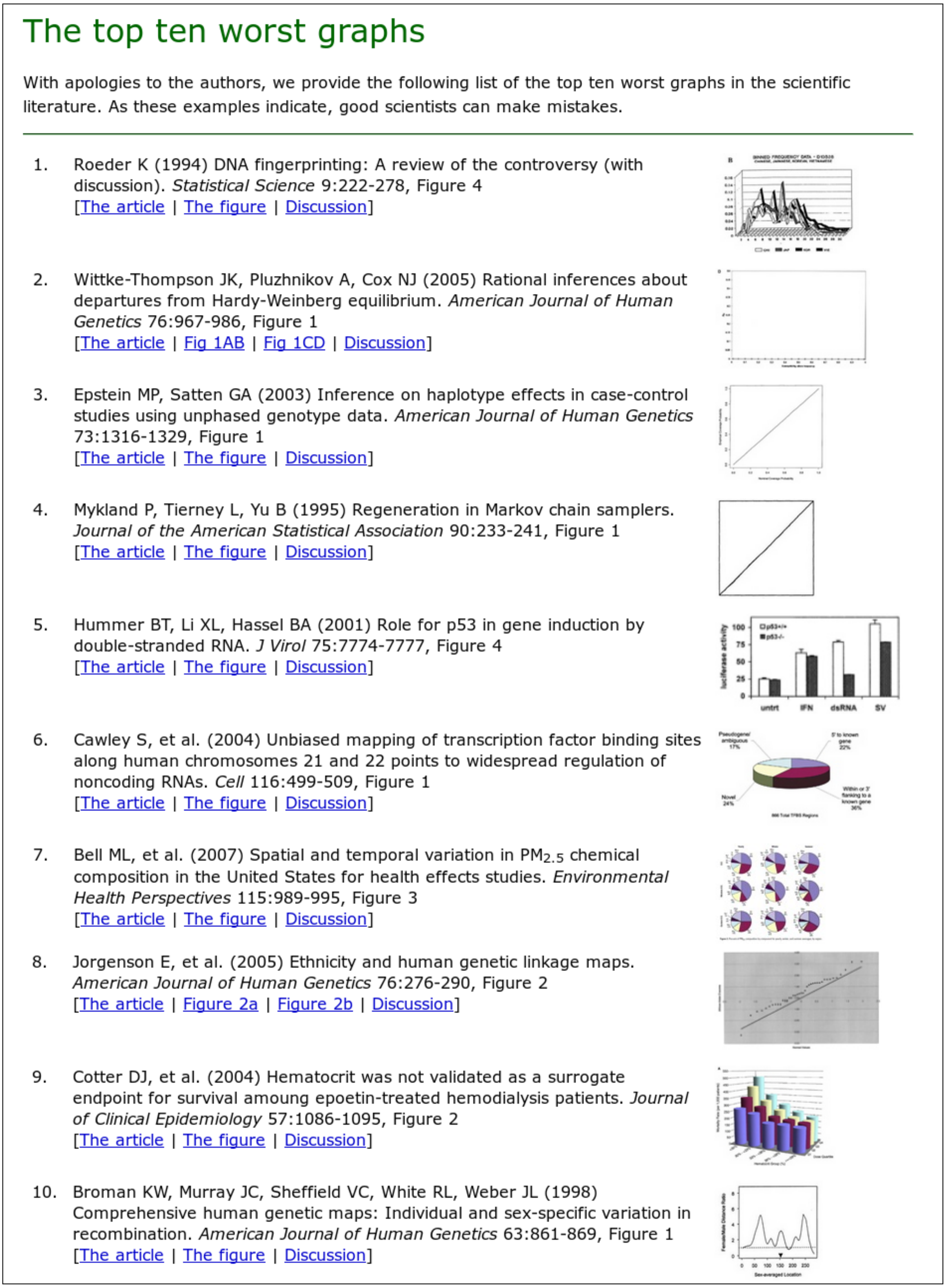
Figure 1.12: Bad example (above) and improved but still not ideal graphic (below). Figures from university documents.

1.5 Bibliographic Remarks

A "complete" or representative list of published material about and tutorials on displaying information is beyond the scope of this section. Here are a few links to works that I consider relevant.

Many consider John W. Tukey to be the founder and promoter of exploratory data analysis. Thus his EDA book (Tukey, 1977) is often seen as the (first) authoritative text on the subject. In a series of books, Tufte rigorously yet vividly explains all relevant elements of visualization and displaying information (Tufte, 1983, 1990, 1997b,a). Many university programs offer lectures on information visualization or similar topics. The lecture by Ross Ihaka is one example worth mentioning: www.stat.auckland.ac.nz/ ihaka/120/lectures.html.

Friendly and Denis (2001) give an extensive historical overview of the evolvement of cartography, graphics and visualization. The document at euclid.psych.yorku.ca/SCS/Gallery/milestone/



The top ten worst graphs

With apologies to the authors, we provide the following list of the top ten worst graphs in the scientific literature. As these examples indicate, good scientists can make mistakes.

1. Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4
   [The article | The figure | Discussion]

2. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1
   [The article | Fig 1AB | Fig 1CD | Discussion]

3. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73:1316-1329, Figure 1
   [The article | The figure | Discussion]

4. Mykland P, Tierney L, Yu B (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association* 90:233-241, Figure 1
   [The article | The figure | Discussion]

5. Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. *J Virol* 75:7774-7777, Figure 4
   [The article | The figure | Discussion]

6. Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509, Figure 1
   [The article | The figure | Discussion]

7. Bell ML, et al. (2007) Spatial and temporal variation in PM$_{2.5}$ chemical composition in the United States for health effects studies. *Environmental Health Perspectives* 115:989-995, Figure 3
   [The article | The figure | Discussion]

8. Jorgenson E, et al. (2005) Ethnicity and human genetic linkage maps. *American Journal of Human Genetics* 76:276-290, Figure 2
   [The article | Figure 2a | Figure 2b | Discussion]

9. Cotter DJ, et al. (2004) Hematocrit was not validated as a surrogate endpoint for survival amoung epoetin-treated hemodialysis patients. *Journal of Clinical Epidemiology* 57:1086-1095, Figure 2
   [The article | The figure | Discussion]

10. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics* 63:861-869, Figure 1
    [The article | The figure | Discussion]

Figure 1.13: Examples of bad plots in scientific journals. The figure is taken from www.biostat.wisc.edu/~kbroman/topten_worstgraphs/. The website discusses the problems with each graph and possible improvements ('[Discussion]' links).

milestone.pdf has active links for virtually endless browsing. See also the applet at www.datavis.ca/milestones/.

There are many interesting videos available illustrating good and not-so-good graphics. For example, www.youtube.com/watch?v=ajUcjR0ma4c. The websites qz.com/580859/the-most-misleading-charts-of-2015-fixed/ and www.statisticshowto.com/misleading-graphs/ illustrate and discuss misleading graphics.

The page en.wikipedia.org/wiki/List_of_statistical_software gives an extensive list of available statistical software programs and environments (from open-source to proprietary) and most of these are compared at en.wikipedia.org/wiki/Comparison_of_statistical_packages.

The raw `swiss` data is available at https://opr.princeton.edu/archive/Download.aspx?FileID=1113 and documentation at https://opr.princeton.edu/archive/Download.aspx?FileID=1116, see also https://opr.princeton.edu/archive/pefp/switz.aspx.

1.6 Exercises and Problems

Problem 1.1 (Introduction to R/RStudio) The aim of this exercise is to get some insight on the capabilities of the statistical software environment R and the integrated development environment `RStudio`.

a) R has many built-in datasets, one example is `volcano`. Based on the help of the dataset, what is the name of the Volcano? Describe the dataset in a few words.

b) Use the R help function to get information on how to use the `image()` function for plotting matrices. Display the volcano data.

c) Install the package `fields`. Display the volcano data with the function `image.plot()`. What is the maximum height of the volcano?

d) Use the the R help function to find out the purpose of the function `demo()` and have a look at the list of available demos. The demo of the function `persp()` utilizes the volcano data to illustrate basic three-dimensional plotting. Call the demo and have a look at the plots.

Problem 1.2 (EDA of bivariate data) On www.isleroyalewolf.org/data/data/home.html the file isleroyale_graph_data_28Dec2011.xlsx contains population data from wolves and moose. Download the data from the STA120 course page. Have a look at the data.

a) Construct a boxplot and a QQ-plot of the moose and wolf data. Give a short interpretation.

b) Jointly visualize the wolves and moose data, as well as their abundances over the years. Give a short interpretation of what *you* see in the figures. (Of course you may compare the result with what is given on the aforementioned web page).

Problem 1.3 (EDA of trivariate data) Perform an EDA of the dataset www.math.uzh.ch/furrer/download/sta120/lemanHgCdZn.csv, containing mercury, cadmium and zinc content in sediment samples taken in lake Geneva.

Problem 1.4 (EDA of multivariate data) In this problem we want to explore the classical *mtcars* dataset (directly available through the package *datasets*). Perform an EDA thereof and provide at least three meaningful plots (as part of the EDA) and a short description of what they display. In what measurement scale are the variables stored and what would be the natural or original measurement scale?

Problem 1.5 (Beyond Example 1.9) In this problem we extend R-Code 1.9 and create further graphics summarizing the *palmerpenguins* dataset.

a) What is `mosaicplot( table( penguins[,c(2,1,8)]))` displaying? Give an interpretation of the plot. Can you increase the clarity (of the information content) with colors?

Would it be possible to further display sex? Compared to the mosaic-plot of Figure 1.9 is there any further insight?

b) Add the marginal histograms to the diagonal panels of the scatterplot matrix in Figure 1.9.

c) In the panels of Figure 1.9 only one additional variable is used for annotation. With the upper- and lower-diagonal plots, it is possible to add two variables. Convince yourself that adding additionally year or island is rather hindering interpretability than helpful. How can the information about island and year be summarized/visualized?

Problem 1.6 (Parallel coordinate plot) Construct a parallel coordinate plot using the built-in dataset *state.x77*. In the left and right margins, annotate the states. Give a few interpretations that can be derived from the plot.

Problem 1.7 (Feature detection with *ggobi*) The open source visualization program ggobi may be used to explore high-dimensional data (Swayne *et al.*, 2003). It provides highly dynamic and interactive graphics such as tours, as well as familiar graphics such as scatterplots, bar charts and parallel coordinates plots. All plots are interactive and linked with brushing and identification. The package *rggobi* provides a link to R.

a) Install the required software ggobi and R package *rggobi* and explore the tool with the dataset *state.x77*.

b) The synthetic dataset *whatfeature*, available at www.math.uzh.ch/furrer/download/sta120/whatfeature.RData has a hidden feature. Try to find it using projection pursuit in *rggobi* and notice how difficult it is to find pronounced structures in small and rather low-dimensional datasets.

Problem 1.8 (BMJ Endgame) Discuss and justify the statements about ‘Skewed distributions’ given in doi.org/10.1136/bmj.c6276.

Chapter 2

Random Variables

Learning goals for this chapter:

- ◇ Describe in own words a cumulative distribution function (cdf), a probability density function (pdf), a probability mass function (pmf), and a quantile function
- ◇ Schematically sketch, plot in R and interpret a pdf/pmf, a cdf, a quantile function
- ◇ Verify that a given function is a pdf/pmf or a cdf. Find a multiplicative constant that makes a given function a pdf/pmf or cdf.
- ◇ Pass from a cdf to a quantile function, pdf or pmf and vice versa
- ◇ Given a pdf/pmf or cdf calculate probabilities
- ◇ Give the definition and intuition of an expected value (E), variance (Var), know the basic properties of E, Var used for calculation
- ◇ Describe a binomial, Poisson and Gaussian random variable and recognize their pmf/pdf

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter02.R.

Probability theory is the prime tool of all statistical modeling. Hence we need a minimal understanding of the theory of probability in order to well understand statistical models, the interpretation thereof, and so forth. Probability theory could (or rather should?) be covered in entire books. Hence, we boil the material down to the bare minimum used in subsequent chapters and thus to some more experienced reader, there seem many gaps in this chapter.

2.1 Basics of Probability Theory

Suppose we want to (mathematically or stochastically) describe an experiment whose outcome is perceived as “random”, the outcome is rather according to chance than to plan or determinism. To do so, we need probability theory, starting with the definition of a probability function, then random variable and properties thereof.

Assume we have a certain experiment. The set of all possible outcomes of this experiment is called the *sample space*, denoted by Ω. Each outcome of the experiment $\omega \in \Omega$ is called an *elementary event*. A subset of the sample space Ω is called an *event*, denoted by $A \subset \Omega$.

Example 2.1. Tossing two coins results in a sample space {HH, HT, TH, TT}. The event “at least one head” {HH, HT, TH} consists of three elementary events. ♣

A *probability measure* is a function $\mathrm{P} : \Omega \to [0, 1]$, that assigns to an event A of the sample space Ω a value in the interval $[0, 1]$, that is, $\mathrm{P}(A) \in [0, 1]$, for all $A \subset \Omega$. Such a function cannot be chosen arbitrarily and has to obey certain rules that are required for consistency. For our purpose, it is sufficient to link these requirements to Kolmogorov’s axioms. More precisely, a probability function must satisfy the following axioms:

1. $0 \le \mathrm{P}(A) \le 1$, for every event A,
2. $\mathrm{P}(\Omega) = 1$,
3. $\mathrm{P}\big(\bigcup_i A_i\big) = \sum_i \mathrm{P}(A_i)$, for $A_i \cap A_j = \emptyset$, $i \ne j$.

In the last bullet, we only specify the index without indicating start and end, which means sum over all possible indices i, say $\sum_{i=1}^{n}$, where n may be finite or countably infinite. (Similarly for the union).

Summarizing informally, a probability is a function that assigns a value to each event of the sample space constraint to:

1. the probability of an event is never smaller than zero or greater than one,
2. the probability of the whole sample space is one,
3. the probability of several events is equal to the sum of the individual probabilities, if the events are mutually exclusive.

Probabilities are often visualized with Euler diagrams (Figure 2.1), which clearly and intuitively illustrate consequences of Kolmogorovs axioms, such as:

$$\begin{aligned}
&\text{monotonicity} && \text{if } C \subset B \implies \mathrm{P}(C) \le \mathrm{P}(B), && (2.1)\\
&\text{empty set} && \mathrm{P}(\varnothing) = 0 && (2.2)\\
&\text{union of two events} && \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B), && (2.3)\\
&\text{complement of an event} && \mathrm{P}(B^c) = \mathrm{P}(\Omega \backslash B) = 1 - \mathrm{P}(B), && (2.4)\\
&\text{conditional probability} && \mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, && (2.5)\\
&\text{law of total probability} && \mathrm{P}(A) = \mathrm{P}(A \mid B)\,\mathrm{P}(B) + \mathrm{P}(A \mid B^c)\,\mathrm{P}(B^c). && (2.6)
\end{aligned}$$

For the second but last statement, we require that $\mathrm{P}(B) > 0$ and conditioning is essentially equivalent to reducing the sample space from Ω to B. The last statement can be written for arbitrary number of events B_i with $B_i \cap B_j = \emptyset$, $i \neq j$ and $\bigcup_i B_i = \Omega$ yielding $\mathrm{P}(A) = \sum_i \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)$.

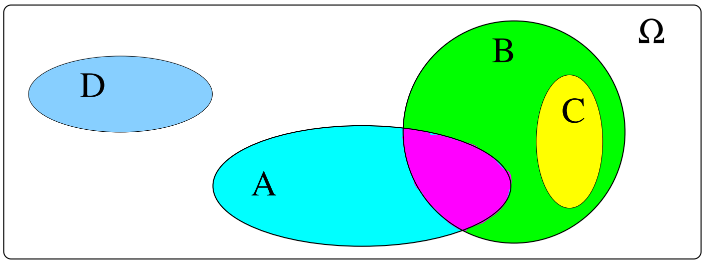


Figure 2.1: Euler diagram where events are illustrated with ellipses. The magenta area represents $A \cap B$ and the event C is in the event B, $C \subset B$.

We now need to formalize this figurative description of probabilities by introducing random variables. A random variable is a function that assigns values to the elementary events of a random experiment, that is, these values or ranges of values are assumed with certain probabilities. The outcomes of the experiment, i.e., the values are called realizations of the random variable.

The following definition introduces a random variable and gives a (unique) characterization of random variables. In subsequent sections, we will see additional characterizations. These, however, will depend on the type of values the random variable takes.

Definition 2.1. Let $\mathrm{P}(\cdot)$ be a probability measure. A random variable X is a measurable function from the sample space Ω to $\mathbb{R}$ and represents a possible numerical outcome of an experiment (Measurable in terms of the probability measure $\mathrm{P}(\cdot)$). The distribution function (cumulative distribution function, cdf) of a random variable X is

$$F(x) = F_X(x) = \mathrm{P}(X \leq x), \tag{2.7}$$

for all $x \in \mathbb{R}$. ♢

Random variables are denoted with uppercase letters (e.g. X, Y), while realizations (i.e., their outcomes, the observations of an experiment) are denoted by the corresponding lowercase letters (x, y). This means that the theoretical “concept“ are denoted by uppercase letters. Actual values or data, for example the columns in your dataset, would be denoted with lowercase letters.

The first two following statements are a direct consequence of the monotonicity of a probability measure and probability of empty set/sample space. The last requires a bit more work to show but will be intuitive shortly.

Property 2.1. *A distribution function $F_X(x)$ is*

1. *monotonically increasing, i.e., for $x < y$, $F_X(x) \leq F_X(y)$;*
2. *normalized, i.e. $\lim\limits_{x\searrow -\infty} F_X(x) = 0$ and $\lim\limits_{x\nearrow \infty} F_X(x) = 1$.*

3. right-continuous, i.e., $\lim_{\epsilon \searrow 0} F_X(x+\epsilon) = F_X(x)$*, for all* $x \in \mathbb{R}$*;*

Remark 2.1. In more formal treatise, one typically introduces a probability space, being a triplet $(\Omega, \mathcal{F}, \mathrm{P})$, consisting of a sample space Ω, a σ-algebra $\mathcal{F}$ (i.e., a collection of subsets of Ω including Ω itself, which is closed under complement and under countable unions) and a probability measure P. A random variable on a probability space is a measurable function from Ω to the real numbers: $X : \Omega \to \mathbb{R}$, $\omega \mapsto X(\omega)$. To indicate its dependence on elementary events, one often writes the argument explicitly, e.g., $\mathrm{P}(X(\omega) \leq x)$. ♣

For further characterizations of random variables, we need to differentiate according to the sample space of the random variables. The next two sections discuss the two essential settings.

2.2 Discrete Distributions

A random variable is called discrete when it can assume only a finite or countably infinite number of values, as illustrated in the following two examples.

Example 2.2. Let X be the sum of the roll of two dice. The random variable X assumes the values $2, 3, \dots, 12$, with probabilities $1/36, 2/36, \dots, 5/36, 6/36, 5/36, \dots, 1/36$. Hence, for example, $\mathrm{P}(X \leq 4) = 1/6$, $\mathrm{P}(X < 2) = 0$, $\mathrm{P}(X < 12.2) = \mathrm{P}(X \leq 12) = 1$. The left panel of Figure 2.2 illustrates the distribution function. This distribution function (as for all discrete random variables) is piece-wise constant with jumps equal to the probability of that value. ♣

Example 2.3. A boy practices *free throws*, i.e., foul shots to the basket standing at a distance of 15 ft to the board. Each shot is either a success or a failure, which can be coded as 1/0. He counts the number of attempts it takes until he has a successful shot. We let the random variable X be the number of throws that are necessary until the boy succeeds. Theoretically, there is no upper bound on this number. Hence X can take the values $1, 2, \dots$. ♣

Next to the cdf, another way of describing discrete random variables is the probability mass function, defined as follows.

Definition 2.2. The probability mass function (pmf) of a discrete random variable X is defined by $f_X(x) = \mathrm{P}(X = x)$. ♢

In other words, the pmf gives probabilities that the random variables takes a precise single value, whereas, as seen, the cdf gives probabilities that the random variables takes that or any smaller value.

Figure 2.2 illustrates the cumulative distribution and the probability mass function of the random variable X as given in Example 2.2. The jump locations and sizes (discontinuities) of the cdf correspond to probabilities given in the right panel. Notice that we have emphasized the right continuity of the cdf (see Proposition 2.1.3) with the additional dot.

It is important to note that we have a theoretical construct here. When tossing two dice several times and reporting the frequencies of their sum, the corresponding plot (bar plot or histogram with appropriate bins) does not exactly match the right panel of Figure 2.2. The more tosses we take, the better the match is which we will discuss further in the next chapter.

R-Code 2.1 Cdf and pmf of X as given in Example 2.2. (See Figure 2.2.)

```
plot.ecdf( outer(1:6, 1:6, "+"),     # generating all possible outcomes
         ylab=bquote(F[X](x)), main='', pch=20)  # `bquote` for subscripts
x <- 2:12                  #  possible outcomes
p <- c(1:6, 5:1)/36        #  corresponding probabilities
plot( x, p, type='h', xlim=c(1,13), ylim=c(0, .2),
     xlab=bquote(x[i]), ylab=bquote(p[i]==f[X](x[i])))
points( x, p, pch = 20) # adding points for clarity
```

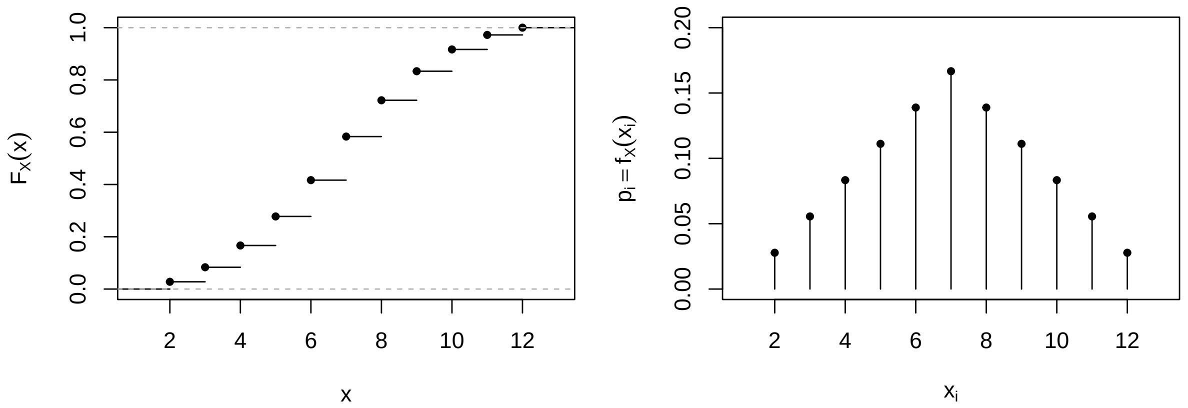


Figure 2.2: Cumulative distribution function (left) and probability mass function (right) of X = “the sum of the roll of two dice”. The two y-axes have a different scale. (See R-Code 2.1.)

Property 2.2. *Let X be a discrete random variable with probability mass function $f_X(x)$ and cumulative distribution function $F_X(x)$. Then:*

1. *The probability mass function satisfies $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.*
2. $\displaystyle\sum_i f_X(x_i) = 1.$
3. $\displaystyle F_X(x) = \sum_{i:x_i \leq x} f_X(x_i).$
4. *The values $f_X(x_i) > 0$ are the “jumps” in x_i of $F_X(x)$.*
5. *The cumulative distribution function is a right-continuous step function.*

Points 3 and 4 of the property show that there is a one-to-one relation (also called a bijection) between the cumulative distribution function and probability mass function. Given one, we can construct the other.

There is of course no limitation on the number of different random variables. In practice, we can often reduce our framework to some common distributions. We now look at two discrete ones.

2.2.1 Binomial Distribution

A random experiment with exactly two possible outcomes (for example: heads/tails, male/female, success/failure) is called a *Bernoulli trial* or Bernoulli random variable. For simplicity, we code the sample space with '1' (success) and '0' (failure). The probability mass function is determined by a single probability:

$$\mathrm{P}(X=1)=p, \qquad \mathrm{P}(X=0)=1-p, \qquad 0<p<1, \tag{2.8}$$

where the cases $p=0$ and $p=1$ are typically not relevant.

Example 2.4. A single three throw (see also Example 2.3) can be modeled as a Bernoulli experiment with success probability p. In practice, repeating throws might probably effect the success probability. For simplicity, one often keeps the probability constant nevertheless. ♣

If a Bernoulli experiment is repeated n times (resulting in an n-tuple of zeros and ones), the exact order of the successes are typically not important, only the total number. Hence, the random variable X = "number of successes in n trials" is intuitive. The distribution of X is called the *binomial distribution*, denoted with $X \sim \mathcal{B}in(n,p)$ and the following applies:

$$\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \qquad 0<p<1, \quad k=0,1,\dots,n. \tag{2.9}$$

Example 2.5. In this example we visualize a particular binomial random variable including a hypothetical experiment. Suppose we draw with replacement 12 times one card from a deck of 36. The probability of having a face card in a single draw is thus 3/9. If X denotes the total number of face cards drawn, we model $X \sim \mathcal{B}in(12,1/3)$. To calculate the probability mass or cumulative distribution function, R provides the construct of prefix and variate. Here, the latter is `binom`. The prefixes are `d` for the probability mass function, `p` for the cumulative distribution function, both illustrated in Figure 2.3.

When I have made the experiment, I've had three face cards; my son only had two. Instead of asking more persons to make the same experiment, we ask R to do so, using the prefix `r` with the variate `binom`. (This also implies that the deck is well mixed and all is added up correctly). Figure 2.3 shows the counts (left) for 10 and frequencies (right) for 100 experiments, i.e., realizations of the random variable. The larger the number of realizations, the closer the match to the probability mass function in the top left corner (again, further discussed in the next chapter). For example, $P(X=5)=0.19$, whereas 24 out of the 100 realizations "observed"

5 face cards and $24/100 = 0.24$ is much closer to 0.19, compared to $1/10 = 0.1$, when only ten realizations are considered.

We have initialized the random number generator or R with a function call *set.seed()* to obtain "repeatable" or reproducible results. ♣

R-Code 2.2 Density and distribution function of a Binomial random variable. (See Figure 2.3.)

```
plot( 0:12, dbinom(0:12, size=12, prob=1/3), type='h',
      xlab='x', ylab=bquote(f[X](x)))
plot( stepfun( 0:12, pbinom(-1:12, size=12, prob=1/3)),
           ylab=bquote(F[X](x)), verticals=FALSE,  main='', pch=20)
set.seed( 14)     # same results if the following lines are evaluated again
print( x10 <- rbinom(10, size=12, prob=1/3 ))    # printing the 10 draws

##  [1] 3 5 7 4 8 4 6 4 4 3

barplot( table( factor(x10, levels=0:12)))       # barplot of the 10 draws
x100 <- rbinom(100, size=12, prob=1/3 )          # 100 draws
barplot( table( factor(x100, levels=0:12))/100)  # frequency barplot
sum(x100==5)                     # how many times five face cards?

## [1] 24
```

2.2.2 Poisson Distribution

The Poisson distribution gives the probability of a given number of events occurring in a fixed interval of time if these events occur with a known and constant rate over time. One way to formally introduce such a random variable is by defining it through its probability mass function. Here and in other cases to follow, memorizing the exact form of the pdf is not necessary; recognizing the stated pmf as the Poisson one is.

Definition 2.3. A random variable X, whose probability mass function is given by

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!}\exp(-\lambda), \qquad 0 < \lambda, \quad k = 0, 1, \dots, \tag{2.10}$$

is said to follow a Poisson distribution with parameter λ, denoted by $X \sim \mathcal{P}ois(\lambda)$. ♢

The Poisson distribution is also a good approximation for the binomial distribution with large n and small p; as a rule of thumb if $n > 20$ and $np < 10$ (see Problem 3.3.a).

Example 2.6. Seismic activities are quite frequent in Switzerland, most of them are not of high strength, luckily. The webpage ecos09.seismo.ethz.ch provides a portal to download information about earthquakes in the vicinity of Switzerland along several other variables. From the page we have manually aggregated the number of earthquakes with a magnitude exceeding 4 between 1980

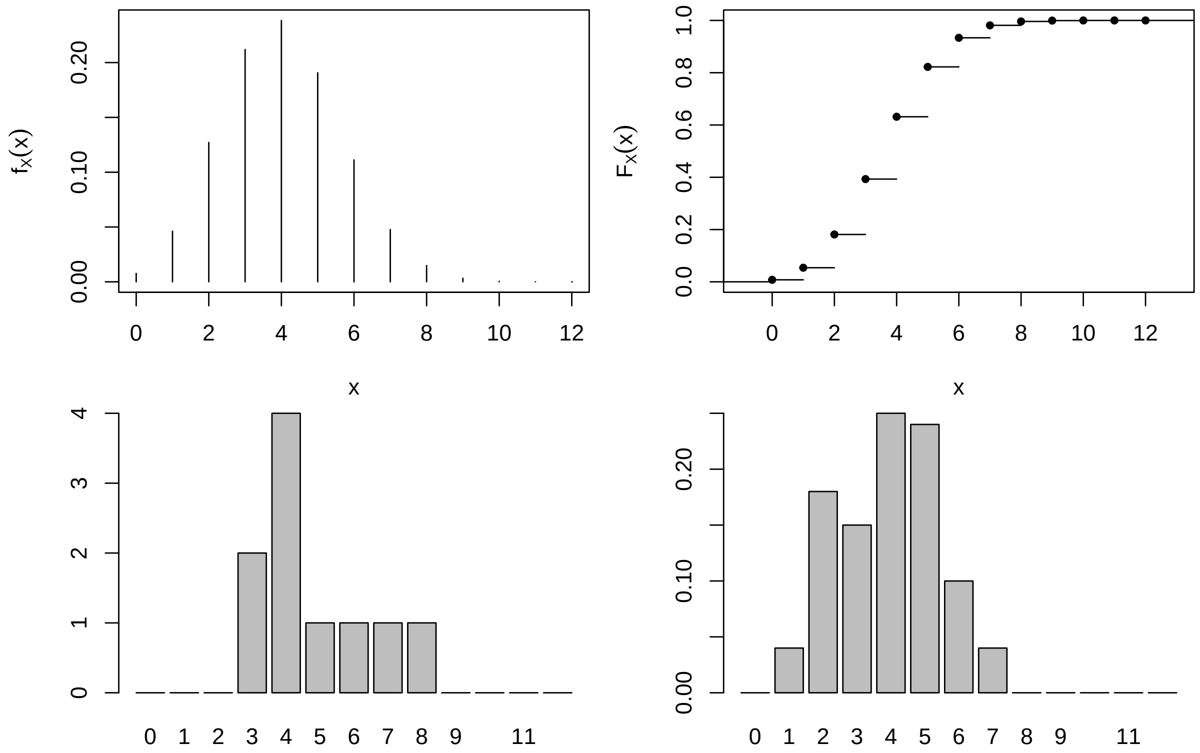


Figure 2.3: Top row: probability mass function and cumulative distribution function of the binomial random variable $X \sim \mathcal{B}in(12, 1/3)$. Bottom row: observed counts for 10 repetitions and observed frequencies for 100 repetitions of the experiment. (See R-Code 2.2.)

and 2005 (Richter magnitude, ML scale). Figure 2.4 (based on R-Code 2.3) shows a histogram of the data with superimposed probability mass function of a Poisson random variable with $\lambda = 2$. There is a surprisingly good fit. There are slightly too few years with a single event, with is offset by a few too many with zero and two events. The value of λ was chosen to match visually the data. In later chapters we see approaches to determine the best possible value (which would be $\lambda = 1.92$ here). ♣

R-Code 2.3 Number of earthquakes and Poisson random variable. (See Figure 2.4.)

```
mag4 <- c(2,0,0,2,5,0,1,4,1,2,1,3,3,2,2,3,0,0,3,1,2,2,2,4,3) # data
hist( mag4, breaks=seq(-0.5, to=10), prob=TRUE, main="", col='gray')
points( 0:10, dpois(0:10, lambda=2), pch=19, col=4)  # pmf of X~Pois(2)
```

2.3 Continuous Distributions

A random variable is called continuous if it can (theoretically) assume any value within one or several intervals. This means that the number of possible values in the sample space is uncount-

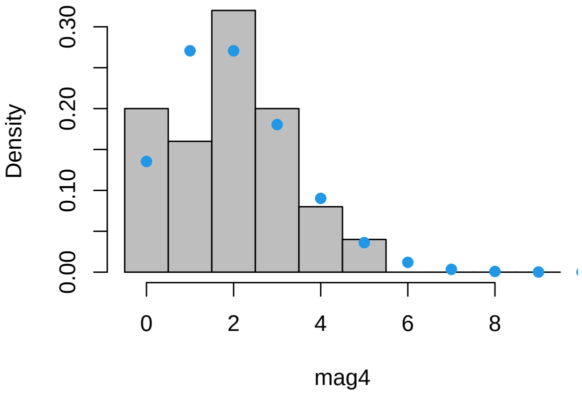


Figure 2.4: Number of earthquakes with a magnitude exceeding 4 between 1980 and 2005 as barplot with superimposed probabilities of the random variable $X \sim \mathcal{Pois}(2)$. (See R-Code 2.3.)

able infinite. Therefore, it is impossible to assign a positive probability value to (elementary) events. Or, in other words, given such an infinite amount of possible outcomes, the likeliness of one particular value being the outcome becomes zero. For this reason, we need to consider outcomes that are contained in a specific interval. Instead of a probability mass function we introduce the density function which is, loosely speaking, the theoretical counterpart to a histogram. The probability is described by an integral, as an area under the probability density function, which is formally defined as follows.

Definition 2.4. The *probability density function* (density function, pdf) $f_X(x)$, or density for short, of a continuous random variable X is defined by

$$\mathrm{P}(a < X \leq b) = \int_a^b f_X(x)dx, \qquad a < b. \tag{2.11}$$

$\diamondsuit$

The density function does not give directly a probability and thus cannot be compared to the probability mass function. The following properties are nevertheless similar to Property 2.2.

Property 2.3. *Let X be a continuous random variable with density function $f_X(x)$ and distribution function $F_X(x)$. Then:*

1. *The density function satisfies $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $f_X(x)$ is continuous almost everywhere.*
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x)dx = 1.$
3. $\displaystyle F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$
4. $\displaystyle f_X(x) = F'_X(x) = \frac{d}{dx}F_X(x).$
5. *The cumulative distribution function $F_X(x)$ is continuous everywhere.*

6. $\mathrm{P}(X = x) = 0$.

Example 2.7. The continuous uniform distribution $\mathcal{U}(a, b)$ is defined by a constant density function over the interval $[a, b]$, $a < b$, i.e., $f(x) = 1/(b - a)$, if $a \leq x \leq b$, and $f(x) = 0$, otherwise. Figure 2.5 shows the density and cumulative distribution function of the uniform distribution $\mathcal{U}(0, 1)$ (see also Problem 1.6).

The distribution function is continuous everywhere and the density has only two discontinuities at a and b. ♣

R-Code 2.4 Density and distribution function of a uniform distribution. (See Figure 2.5.)

```
plot( c(-0.5, 0, NA, 0, 1, NA, 1, 1.5), c(0, 0, NA, 1, 1, NA, 0, 0),
      type='l', xlab='x', ylab=bquote(f[X](x)))
# curve(dunif( x), -0.5, 1.5) # does not emphasize the discontinuity!!
curve(punif( x), -0.5, 1.5)
```

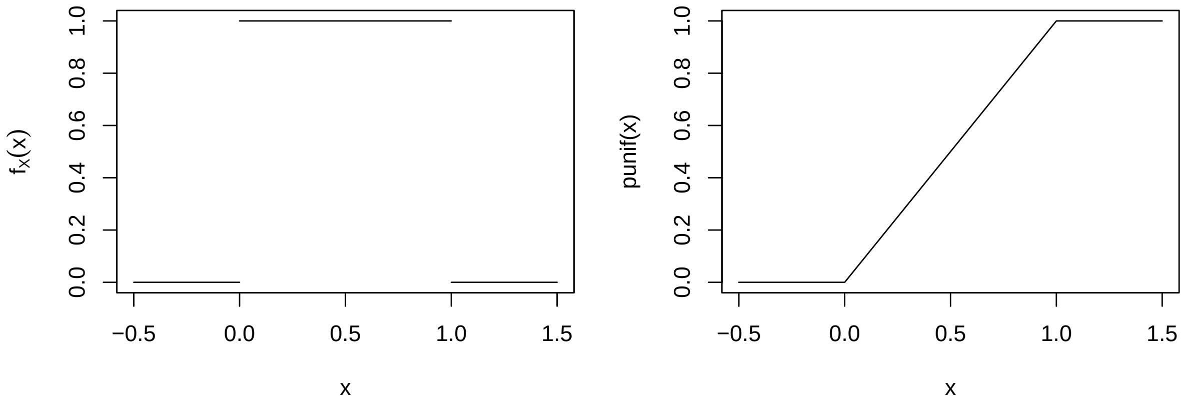


Figure 2.5: Density and distribution function of the uniform distribution $\mathcal{U}(0, 1)$. (See R-Code 2.4.)

As given by Property 2.3.3 and 4, there is again a bijection between the density function and the cumulative distribution function: if we know one we can construct the other. Actually, there is a third characterization of random variables, called the quantile function, which is essentially the inverse of the cdf. That means, we are interested in values x for which $F_X(x) = p$.

Definition 2.5. The quantile function $Q_X(p)$ of a random variable X with (strictly) monotone cumulative distribution function $F_X(x)$ is defined by

$$Q_X(p) = F_X^{-1}(p), \quad 0 < p < 1, \tag{2.12}$$

i.e., the quantile function is equivalent to the inverse of the distribution function. ♢

In R, the quantile function is specified with the prefix `q` and the corresponding variate. For example, `qunif` is the quantile function for the uniform distribution, which is $Q_X(p) = a+p(b-a)$ for $0 < p < 1$.

The quantile function can be used to define the theoretical counter part to the sample quartiles of Chapter 1 as illustrated next.

Definition 2.6. The median ν of a continuous random variable X with cumulative distribution function $F_X(x)$ is defined by $\nu = Q_X(1/2)$. Accordingly, the lower and upper quartiles of X are $Q_X(1/4)$ and $Q_X(3/4)$. ♢

The quantile function is essential for the theoretical quantiles of QQ-plots (Section 1.3.1). For example, the two left-most panels of Figure 1.6 are based essentially on the $i/(n+1)$-quantiles.

Remark 2.2. Instead of the simple $i/(n+1)$-quantiles, R uses the form $(i-a)/(n+1-2a)$, for a specific $a \in [0,1]$. The precise value can be specified using the argument `type` in `quantile()`, or `qtype` in `qqline` or `a` in `ppoints()`. Of course, different definitions of quantiles only play a role for very small sample sizes and in general we should not bother what approach has been taken. ♣

Remark 2.3. For discrete random variables the cdf is not continuous (see the plateaus in the left panel of Figure 2.2) and the inverse does not exist. The quantile function returns the minimum value of x from among all those values with probability $p \leq \mathrm{P}(X \leq x) = F_X(x)$, more formally,

$$Q_X(p) = \inf_{x\in\mathbb{R}} \{p \leq F_X(x)\}, \quad 0 < p < 1, \tag{2.13}$$

where for our level of rigor here, the *inf* can be read as *min*. ♣

2.4 Expectation and Variance

Density, cumulative distribution function or quantile function uniquely characterize random variables. Often we do not require such a complete definition and “summary” values are sufficient. We start introducing a measure of location (expectation) and spread (variance). More and alternative measures will be seen in Chapter 7.

Definition 2.7. The expectation of a discrete random variable X is defined by

$$\mathrm{E}(X) = \sum_i x_i \, \mathrm{P}(X = x_i)\,. \tag{2.14}$$

The expectation of a continuous random variable X is defined by

$$\mathrm{E}(X) = \int_{\mathbb{R}} x f_X(x) dx\,, \tag{2.15}$$

where $f_X(x)$ denotes the density of X. ♢

Remark 2.4. Mathematically, it is possible that the expectation is not finite: the random variable X may take very, very large values too often. In such cases we would say that the expectation does not exist. We see a single example in Chapter 8 where this is the case. In all other situations we assume a finite expectation and for simplicity, we do not explicitly state this. ♣

Many other "summary" values are reduced to calculate a particular expectation. The following property states how to calculate the expectation of a function of the random variable X, which is in turn used to summarize the spread of X.

Property 2.4. *For an "arbitrary" real function g we have:*

$$\mathrm{E}\big(g(X)\big) = \begin{cases} \displaystyle\sum_i g(x_i)\,\mathrm{P}(X = x_i), & \textit{if } X \textit{ discrete,} \\ \displaystyle\int_{\mathbb{R}} g(x) f_X(x) dx, & \textit{if } X \textit{ continuous.} \end{cases}$$

We often take for $g(x) = x^k$, $k = 2, 3 \dots$ and we call $\mathrm{E}(X^k)$ the *higher moments* of X.

Definition 2.8. The variance of X is the expectation of the squared deviation from its expected value:

$$\mathrm{Var}(X) = \mathrm{E}\big((X - \mathrm{E}(X))^2\big) \tag{2.16}$$

and is also denoted as the centered second moment, in contrast to the second moment $\mathrm{E}(X^2)$.

The standard deviation of X is $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$. ♢

The expectation is "linked" to the average (or sample mean, `mean()`) if we have a set of realizations thought to be from the particular random variable. Similarly, the variance is "linked" to the sample variance (`var()`). This link will be formalized in later chapters.

Example 2.8. 1. The expectation and variance of a Bernoulli trial are

$$\mathrm{E}(X) = 0 \cdot (1-p) + 1 \cdot p = p, \tag{2.17}$$

$$\mathrm{Var}(X) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p). \tag{2.18}$$

2. The expectation and variance of a Poisson random variable are (see Problem 3.3.a)

$$\mathrm{E}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda. \tag{2.19}$$

Property 2.5. *For random variables X and Y, regardless of whether discrete or continuous, and for a and b given constants, we have*

1. $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$;

2. $\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$;

3. $\mathrm{Var}(a + bX) = b^2 \,\mathrm{Var}(X)$ *,*

4. $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$*.*

The second but last property seems somewhat surprising. But starting from the definition of the variance, one quickly realizes that the variance is not a linear operator:

$$\mathrm{Var}(a + bX) = \mathrm{E}\Big(\big(a + bX - \mathrm{E}(a + bX)\big)^2\Big) = \mathrm{E}\Big(\big(a + bX - (a + b\,\mathrm{E}(X))\big)^2\Big), \tag{2.20}$$

followed by a factorization of b^2.

Example 2.9 (continuation of Example 2.2)**.** For X the sum of the roll of two dice, straightforward calculation shows that

$$\mathrm{E}(X) = \sum_{i=2}^{12} i\,\mathrm{P}(X = i) = 7, \qquad \text{by equation (2.14),} \tag{2.21}$$

$$= 2\sum_{i=1}^{6} i\frac{1}{6} = 2 \cdot \frac{7}{2}, \qquad \text{by using Property 2.5.4 first.} \quad ♣ \tag{2.22}$$

As a small note, recall that the first moment is a measure of location, the centered second moment a measure of spread. The centered third moment is a measure of asymmetry and can be used to quantify the skewness of a distribution. The centered forth moment is a measure of heaviness of the tails of the distribution.

2.5 The Normal Distribution

The normal or Gaussian distribution is probably the most known distribution, having the omnipresent "bell-shaped" density. Its importance is mainly due the fact that the sum of Gaussian random variables is distributed again as a Gaussian random variable. Moreover, the sum of many "arbitrary" random variables is distributed approximately as a normal random variable. This is due to the celebrated central limit theorem, which we see in details in the next chapter. As in the case of a Poisson random variable, we define the normal distribution by giving its density.

Definition 2.9. The random variable X is said to be normally distributed if the cumulative distribution function is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{2.23}$$

with density function

$$f(x) = f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right), \tag{2.24}$$

for all x ($\mu \in \mathbb{R}$, $\sigma_x > 0$). We denote this with $X \sim \mathcal{N}(\mu, \sigma^2)$.

The random variable $Z = (X - \mu)/\sigma$ (the so-called *z-transformation*) is standard normal and its density and distribution function are usually denoted with $\varphi(z)$ and $\Phi(z)$, respectively. ♢

While the exact form of the density (2.24) is not important, a certain recognizing factor will be very useful. Especially, for a standard normal random variable, the density is proportional to $\exp(-z^2/2)$. Figure 2.7 gives the density, distribution and the quantile function of a standard norm distributed random variable.

Property 2.6. *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$*, then* $\mathrm{E}(X) = \mu$ *and* $\mathrm{Var}(X) = \sigma^2$*.*

The following property is essentially a rewriting of the second part of the definition. We will state it explicitly because of its importance.

Property 2.7. *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$*, then* $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ *and* $F_X(x) = \Phi\left(\frac{X-\mu}{\sigma}\right)$*. Conversely, if* $Z \sim \mathcal{N}(0,1)$*, then* $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$*,* $\sigma > 0$*.*

The cumulative distribution function Φ has no closed form and the corresponding probabilities must be determined numerically. In the past, so-called "standard tables" summarized probabilities and were included in statistics books. Table 2.1 gives an excerpt of such a table. Now even "simple" pocket calculators have the corresponding functions to calculate the probabilities. It is probably worthwhile to remember $84\% = \Phi(1)$, $98\% = \Phi(2)$, $100\% \approx \Phi(3)$, as well as $95\% = \Phi(1.64)$ and $97.5\% = \Phi(1.96)$. Relevant quantiles have been illustrated in Figure 2.6 for a standard normal random variable. For arbitrary normal density, the density scales linearly with the standard deviation.

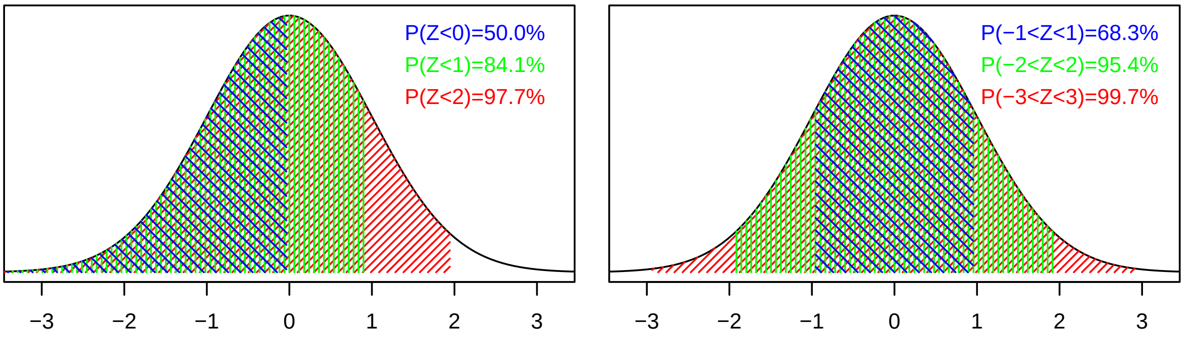


Figure 2.6: Different probabilities for the standard normal distribution.

Table 2.1: Probabilities of the standard normal distribution. The table gives the value of $\Phi(z_p)$ for selected values of z_p. For example, $\Phi(0.2 + 0.04) = 0.595$.

z_p	0.0	0.1	0.2	0.3	0.4	...	1	...	1.6	1.7	1.8	1.9	2	...	3
0.00	0.500	0.540	0.579	0.618	0.655		0.841		0.945	0.955	0.964	0.971	0.977		0.999
0.02	0.508	0.548	0.587	0.626	0.663		0.846		0.947	0.957	0.966	0.973	0.978		0.999
0.04	0.516	0.556	0.595	0.633	0.670		0.851		0.949	0.959	0.967	0.974	0.979		⋮
0.06	0.524	0.564	0.603	0.641	0.677		0.855		0.952	0.961	0.969	0.975	0.980		
0.08	0.532	0.571	0.610	0.648	0.684		0.860		0.954	0.962	0.970	0.976	0.981		

R-Code 2.5 Calculation of the "z-table" (see Table 2.1) and density, distribution, and quantile functions of the standard normal distribution. (See Figure 2.7.)

```
y <- seq( 0, by=.02, length=5)          # row values
x <- c( seq( 0, by=.1, to=.4), 1, seq(1.6, by=.1, to=2), 3) # column values
round( pnorm( outer( x, y, "+")), 3)

plot( dnorm, -3, 3, ylim=c(-.5,2), xlim=c(-2.6,2.6))
abline( c(0, 1), h=c(0,1), v=c(0,1), col='gray') # diag and horizontal lines
plot( pnorm, -3, 3, col=3, add=TRUE)
plot( qnorm, 0, 1, col=4, add=TRUE)
```

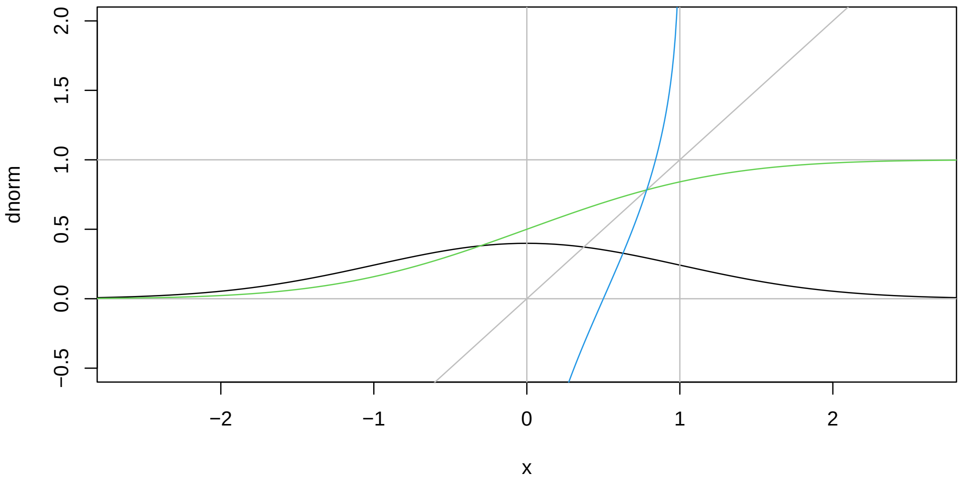


Figure 2.7: Probability density function (black), cumulative distribution function (green), and quantile function (blue) of the standard normal distribution. (See R-Code 2.5.)

We finish this section with two probability calculations that are found similarly in typical textbooks. We will, however, revisit the Gaussian distribution in virtually every following chapter.

Example 2.10. Let $X \sim \mathcal{N}(4, 9)$. Then

1. $\text{P}(X \leq -2) = \text{P}\left(\frac{X-4}{3} \leq \frac{-2-4}{3}\right)$
$$= \text{P}(Z \leq -2) = \Phi(-2) = 1 - \Phi(2) \approx 1 - 0.977 = 0.023\,.$$

2. $\text{P}(|X-3| > 2) = 1 - \text{P}(|X-3| \leq 2) = 1 - \text{P}(-2 \leq X - 3 \leq 2)$
$$= 1 - \big(\text{P}(X-3 \leq 2) - \text{P}(X-3 \leq -2)\big) = 1 - \Phi\left(\frac{5-4}{3}\right) + \Phi\left(\frac{1-4}{3}\right)$$
$$\approx 1 - \frac{0.626 + 0.633}{2} + (1 - 0.841) \approx 0.5295\,.$$ ♣

2.6 Bibliographic Remarks

There is an abundance list of books in probability, discussing the concepts of probabilities, random variables and properties thereof, e.g., Grinstead and Snell (2003) (PDF version under GNU FDL exists), Ross (2010) (or any older version), or advanced, but classical Feller (1968). In fact, the majority of textbooks in statistics have some introduction to probability, in a similar sprit as here.

The ultimate reference for (univariate) distributions is the encyclopedic series of Johnson *et al.* (2005, 1994, 1995). Figure 1 of Leemis and McQueston (2008) illustrates extensively the links between countless univariate distributions, a simplified version is available at https://www.johndcook.com/blog/distribution_chart/.

In general, wikipedia has nice summaries of many distributions. The page https://en.wikipedia.org/wiki/List_of_probability_distributions lists many thereof.

2.7 Exercises and Problems

Problem 2.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Show Property 2.5.

b) Show Property 2.6.

Hint: Let $Z \sim \mathcal{N}(0,1)$. Show that $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$, then use Property 2.5.

Problem 2.2 (Visualizing probabilities) For events $B_1, \dots, B_5$ with $B_i \cap B_j = \emptyset$, $i \neq j$ and $\bigcup_{i=1}^{5} B_i = \Omega$ visualize the law of total probability $\mathrm{P}(A) = \sum_i \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)$ using an Euler-diagram.

Problem 2.3 (Properties of probability measures) For events A and B show that:

a) $\mathrm{P}(A \backslash B) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$;

b) $\mathrm{P}(A \cup B) = 1 - \mathrm{P}(A^c \cap B^c)$;

c) $\mathrm{P}(A \cap B) = 1 - \mathrm{P}(A^c \cup B^c)$.

Problem 2.4 (Counting events) Place yourself in a sidewalk café during busy times and for several consecutive minutes, count the number of persons walking by. (If you do not enjoy the experimental setup, it is possible to count the number of vehicules in one of the live cams linked at https://www.autobahnen.ch/index.php?lg=001&page=017).

a) Visualize the data with a histogram and describe its form.

b) Try to generate realizations from a Poisson random variable whose histogram matches best the the one seen in **a**), i.e., use `rpois()` for different values of the argument `lambda`.

Would there be any other distribution that would yield a better match?

Problem 2.5 (Discrete uniform distribution) Let m be a positive integer. The discrete uniform distribution is defined by the pmf

$$\mathrm{P}(X = k) = \frac{1}{m}, \qquad k = 1, \dots, m.$$

a) Visualize the pmf and cdf of a discrete uniform random variable with $m = 12$.

b) Draw several realizations from X, visualize the results and compare to the pmf of **a**). What are sensible graphics types for the visualization?
Hint: the function `sample()` conveniently draws random samples without replacement.

c) Show that $\mathrm{E}(X) = \dfrac{m+1}{2}$ and $\mathrm{Var}(X) = \dfrac{m^2-1}{12}$.

Hint: $\displaystyle\sum_{k=1}^{m} k^2 = \frac{m(m+1)(2m+1)}{6}$.

Problem 2.6 (Uniform distribution) We assume $X \sim \mathcal{U}(0, \theta)$, for some value $\theta > 0$.

a) For all a and b, with $0 < a < b < \theta$ show that $\mathrm{P}(X \in [a, b]) = (b - a)/\theta$.

b) Calculate $\mathrm{E}(X)$, $\mathrm{SD}(X)$ and the quartiles of X.

c) Choose a sensible value for θ. In R, simulate $n = 10, 50, 10000$ random numbers and visualize the histogram as well as a QQ-plot thereof. Is it helpful to superimpose a smoothed density to the histograms (with `lines( density( ...))`)?

Problem 2.7 (Calculating probabilities) In the following settings, approximate the probabilities and quantiles q_1 and q_2 using a Gaussian “standard table”. Compare these values with the ones obtained with R.

$$X \sim \mathcal{N}(2, 16): \qquad \mathrm{P}(X < 4), \quad \mathrm{P}(0 \leq X \leq 4), \quad \mathrm{P}(X > q_1) = 0.95, \quad \mathrm{P}(X < -q_2) = 0.05.$$

If you do not have standard table, the following two R commands may be used instead: `pnorm(a)` and `qnorm(b)` for specific values `a` and `b`.

Problem 2.8 (Exponential Distribution) In this problem we get to know another important distribution you will frequently come across - the expontential distribution. Consider the random variable X with density

$$f(x) = \begin{cases} 0, & x < 0, \\ c \cdot \exp(-\lambda x), & x \geq 0, \end{cases}$$

with $\lambda > 0$. The parameter λ is called the rate. Subsequently, we denote an exponential random variable with $X \sim \mathcal{E}xp(\lambda)$.

a) Determine c such that $f(x)$ is a proper density.

b) Determine the cumulative distribution function (cdf) $F(x)$ of X.

c) Determine the quantile function $Q(p)$ of X. What are the quartiles of X?

d) Let $\lambda = 2$. Calculate:

$$\begin{array}{lll} \mathrm{P}(X \in \mathbb{R}) & \mathrm{P}(X \geq -10) & \mathrm{P}(X = 4) \\ \mathrm{P}(X \leq 4) & \mathrm{P}(X \leq \log(2)/2) & \mathrm{P}(3 < X \leq 5) \end{array}$$

e) Show that $\mathrm{E}(X) = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$.

f) Show that $\mathrm{P}(X > s + t \mid X > t) = \mathrm{P}(X > s)$, $s, t \geq 0$.
A random variable satisfying the previous equation is called *memoryless*. (Why?)

Problem 2.9 (BMJ Endgame) Discuss and justify the statements about 'The Normal distribution' given in doi.org/10.1136/bmj.c6085.

Chapter 3

Functions of Random Variables

Learning goals for this chapter:

- ⋄ Know the definition and properties of independent and identically distributed (iid) random variables
- ⋄ Know the distribution of the average of iid Gaussian random variables
- ⋄ Explain in own words how to construct chi-squared, t- and F-distributed random variables
- ⋄ Know the central limit theorem (CLT) and being able to approximate binomial random variables
- ⋄ Able to calculate the cdf of transformed random variable, and if applicable, the pdf as well
- ⋄ Able to approximate the mean and variance of transformed random variables

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter03.R.

The last chapter introduced "individual" random variables, e.g., a Poisson random variable, Normal random variable. In subsequent chapters, we need as many random variables as we have observations because we will match one random variable with each observation. Of course, these random variables may have the same distribution (may be identical).

3.1 Independent Random Variables

We considered the case of several trials and aggregated these to a binomial random variable. By definition, the trials are identical and independent of each other. We now formally introduce independence.

Definition 3.1. Two events A and B are independent if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)\,. \tag{3.1}$$

Two random variables X and Y are independent if for all A and B

$$\mathrm{P}(X \in A \cap Y \in B) = \mathrm{P}(X \in A)\,\mathrm{P}(Y \in B)\,. \tag{3.2}$$

The random variables $X_1, \dots, X_n$ are independent if for all $A_1, \dots, A_n$

$$\mathrm{P}\big(\bigcap_{i=1}^{n} X_i \in A_i\big) = \prod_{i=1}^{n} \mathrm{P}(X_i \in A_i)\,. \tag{3.3}$$

If this is not the case, they are dependent (all three cases). ♢

The concept of independence can be translated to information flow. Recall the formula of conditional probability (2.5), if we have independence, $\mathrm{P}(A \mid B) = \mathrm{P}(A)$: knowing anything about the event B does not change the probability (knowledge) of the event A. Knowing that I was born on a Sunday does not provide any information whether I have a driving license. But having a driving license increases the probability that I own a car.

Formally, to write Equation (3.2), we would have to introduce bivariate random variables. We will formalize these ideas in Chapter 8 but note here that the definition of independence also implies that the joint density and the joint cumulative distribution is simply the product of the individual ones, also called marginal ones. For example, if X and Y are independent, their joint density is $f_X(x)f_Y(y)$.

Example 3.1. The sum of two dice (Example 2.2) is not independent of the value of the first die: $\mathrm{P}(X \leq 4 \cap Y \leq 2) = 5/36 \neq \mathrm{P}(X \leq 4)\,\mathrm{P}(Y \leq 2) = 1/6 \cdot 1/3$. ♣

Example 3.2 (continuation of Example 2.3)**.** If we assume that each foul shot is independent, then the probability that 6 shots are necessary is $\mathrm{P}(X = 6) = (1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p)\cdot p = (1-p)^5 p$, where p is the probability that a shot is successful.

This independence also implies that the success of the next shot is independent of my previous one. In other words, there is no memory of how often I have missed in the past. See also Problem 3.2. ♣

We will often use several independent random variables with a common distribution function.

Definition 3.2. A random sample $X_1, \dots, X_n$ consists of n independent random variables with the same distribution, specified by, say, F. We write $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$ where iid stands for "independent and identically distributed". The number of random variables n is called the sample size. ♢

The iid assumption is very crucial and relaxing the assumptions to allow, for example, dependence between the random variables, has severe implications on the statistical modeling. Independence also implies a simple formula for the variance of the sum of two or several random variables, a formal justification will follow in Chapter 8.

Property 3.1. *Let X and Y be two independent random variables. Then*

1. $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)\,.$

Let $X_1, \dots, X_n \overset{iid}{\sim} F$ *with* $\mathrm{E}(X_1) = \mu$ *and* $\mathrm{Var}(X_1) = \sigma^2$*. Denote* $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$*. Then*

2. $\mathrm{E}(\overline{X}) = \mathrm{E}\big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\big) = \mathrm{E}(X_1) = \mu\,.$

3. $\mathrm{Var}(\overline{X}) = \mathrm{Var}\big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\big) = \dfrac{1}{n}\,\mathrm{Var}(X_1) = \dfrac{\sigma^2}{n}\,.$

Example 3.3. For $X \sim \mathcal{Bin}(n, p)$, we have

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p) \tag{3.4}$$

as we can write the binomial random variable as a sum of Bernoulli ones. ♣

The latter two points of Property 3.1 are used when we investigate statistical properties of the sample mean. This concept is quite powerful and works as follows. Consider the sample mean $\bar{x} = 1/n\sum_{i=1}^{n} x_i$ which can be seen as function of n arguments $f(x_1, \dots, x_n) = \bar{x}$. We then evaluate the function at the arguments corresponding to the random sample, $f(X_1, \dots, X_n)$, which is the random sample mean $\overline{X} = 1/n\sum_{i=1}^{n} X_i$, a random variable itself!

In subsequent chapters we encounter essentially the following two situations:

1. (practical context) We have n observations $x_1, \dots, x_n$ which we will model statistically and thus assume a distribution F such that $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$.

2. (theoretical approach) We start with a random sample and study the theoretical properties. Often these will be illustrated with realizations from the the random sample (generated using R).

3.2 Random Variables Derived from Gaussian Random Variables

Gaussian random variable have many appealing properties such that we often assume $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. In Chapter 1 we summarized data $x_1, \dots, x_n$ with statistical measures like sample mean, sample variance etc. Now using the approach outlined above by replacing the observations with the random variables (in equations (1.1), (1.3), for example) we may wonder what is the distribution of these statistics?

We start with the following property which is essential and will be consistently used throughout the work.

Property 3.2. *Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent and a and b arbitrary, then $aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.*

Hence, the density of the sum of two random variables is again unimodal and does not correspond to a simple "superposition" of the densities. The following property is essential and a direct consequence of the previous one.

Property 3.3. *Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.*

The next paragraphs discuss random variables that are derived (obtained as functions) from Gaussian random variables, going beyond the average of iid Gaussians.

More specifically, we will look at the distribution of $(\overline{X}-\mu)/\sqrt{\sigma^2/n}$, $S^2 = 1/(n-1)\sum_{i=1}^n (X_i - \overline{X})^2$ and other forms that are crucial in later chapters.

The expressions of the densities of the following random variables is not essential, they are complex and we do not use them subsequently. Similarly, the expectation and the variance are for reference only and might be helpful when scrutinizing some derivations or statistical approaches.

3.2.1 Chi-Square Distribution

The distribution of squared standard normal random variables is said to be chi-squared. Formally, let $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The distribution of the random variable

$$X = \sum_{i=1}^n Z_i^2 \tag{3.5}$$

is called the chi-square distribution ($\mathcal{X}^2$ distribution) with n degrees of freedom and denoted $X \sim \mathcal{X}_n^2$. The following applies:

$$\mathrm{E}(X) = n; \qquad \mathrm{Var}(X) = 2n. \tag{3.6}$$

We summarize below, the exact link between the distribution of S^2 and a chi-square distribution. Further, the chi-square distribution is used in numerous statistical tests that we see Chapter 5 and 6.

R-Code 3.1 Chi-square distribution for various degrees of freedom. (See Figure 3.1.)

```
x <- seq( 0, to=50, length=150)                 # x-values for which we plot
plot(x, dchisq( x, df=1), type='l', ylab='Density')  # plot first density
for (i in 1:6)                                  # different degrees of freedom
    lines( x, dchisq(x, df=2^i), col=i+1)
legend( "topright", legend=2^(0:6), col=1:7, lty=1, bty="n")
```

For small n the density of the chi-square distribution is quite right-skewed (Figure 3.1). For larger n the density is quite symmetric and has a bell-shaped form. In fact, for $n > 50$, we can approximate the chi-square distribution with a normal distribution, i.e., $\mathcal{X}_n^2$ is distributed approximately $\mathcal{N}(n, 2n)$ (a justification is given in Section 3.3).

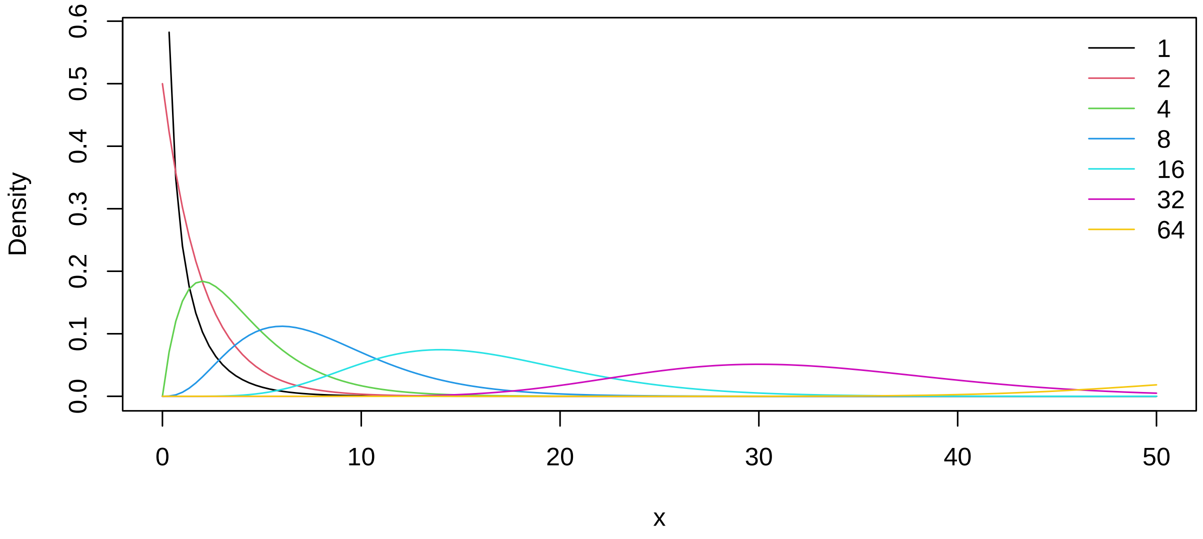


Figure 3.1: Densities of the chi-square distribution for various degrees of freedom. (See R-Code 3.1.)

Remark 3.1. A chi-square distribution is a particular case of a so-called Gamma distribution, which we will introduce in Chapter 13.

There are further transformations of a chi-square distribution that are approximately normal, e.g. for $\mathcal{X}_n^2$ with $n > 30$ the random variable $X = \sqrt{2\mathcal{X}_n^2}$ is approximately normally distributed with expectation $\sqrt{2n-1}$ and standard deviation of one. Similar approximations exist for $(\mathcal{X}_n^2/n)^{1/3}$ or $(\mathcal{X}_n^2/n)^{1/4}$, see, e.g., Canal (2005). ♣

3.2.2 Student's t-Distribution

We now introduce a distribution that is used when standardizing $\overline{X}$. This distribution is quite omnipresent and borrowed its name to further statistical approaches, e.g., the famous t-tests to compare sample means.

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{X}_m^2$ be two independent random variables. The distribution of the random variable

$$V = \frac{Z}{\sqrt{X/m}} \tag{3.7}$$

is called the t-distribution (or Student's t-distribution) with m degrees of freedom and denoted $V \sim t_n$. We have:

$$\mathrm{E}(V) = 0, \qquad \text{for } m > 1; \tag{3.8}$$

$$\mathrm{Var}(V) = \frac{m}{(m-2)}, \qquad \text{for } m > 2. \tag{3.9}$$

The density is symmetric around zero and as $m \to \infty$ the density converges to the standard normal density $\varphi(x)$ (see Figure 3.2, based on R-Code 3.2).

Remark 3.2. For $m = 1, 2$ the density is *heavy-tailed* and the variance of the distribution does not exist. Realizations of this random variable occasionally manifest with extremely large values.

R-Code 3.2 *t*-distribution for various degrees of freedom. (See Figure 3.2.)

```
x <- seq( -3, to=3, length=100)              # x-values for which we plot
plot( x, dnorm(x), type='l', ylab='Density') # Gaussian density as reference
for (i in 0:6)
    lines( x, dt(x, df=2^i), col=i+2)        # t-densities with different dofs
legend( "topright", legend=2^(0:6), col=2:8, lty=1, bty="n")
```

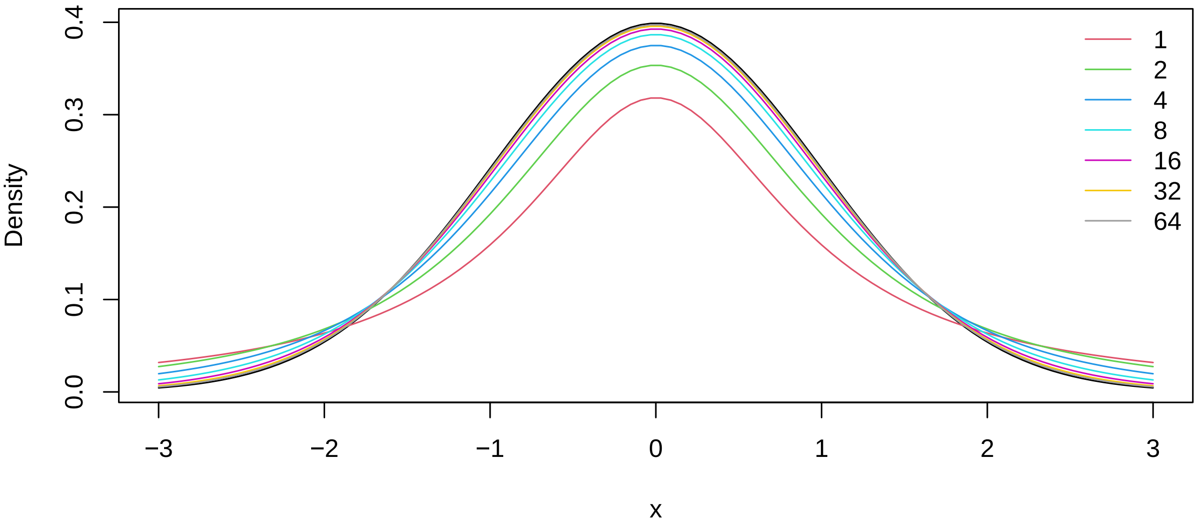


Figure 3.2: Densities of the *t*-distribution for various degrees of freedom. The normal distribution is in black. A density with $2^7 = 128$ degrees of freedom would make the normal density function appear thicker. (See R-Code 3.2.)

Of course, the sample variance can still be calculated (see R-Code 3.3). We come back to this issue in Chapter 6. ♣

R-Code 3.3 Sample variance of the *t*-distribution with one degree of freedom.

```
set.seed( 14)
tmp <- rt( 1000, df=1)   # 1000 realizations
var( tmp)                # variance is huge!!
## [1] 37391
sort( tmp)[1:7]      # many "large" values, but 2 exceptionally large
## [1] -190.929 -168.920  -60.603  -53.736  -47.764  -43.377  -36.252
sort( tmp, decreasing=TRUE)[1:7]
## [1] 5726.53 2083.68  280.85  239.75  137.36  119.16  102.70
```

The following property is fundamental in statistics.

Property 3.4. *Let* $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *Define* $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *and* $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

1. *(a)* $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ *(b)* $\dfrac{n-1}{\sigma^2}S^2 \sim \mathcal{X}^2_{n-1}$.

2. $\overline{X}$ *and* S^2 *are independent.*

3. $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Statement 1(a) is not surprising and is a direct consequence of Properties 2.7 and 3.3. Statement 1(b) is surprising insofar that centering the random variables with $\overline{X}$ amounts to reducing the degrees of freedom by one. Point 2 seems very surprising as the same random variables are used in $\overline{X}$ and S^2. A justification is that S^2 is essentially a sum of random variables which were corrected for $\overline{X}$ and thus S^2 does not contain information about $\overline{X}$ anymore. In Chapter 8 we give a more detailed explanation thereof. Point 3 is not surprising as we use the definition of the t-distribution and the previous two points.

3.2.3 F-Distribution

The F-distribution is mainly used to compare two sample variances with each other, as we will see in Chapters 5, 10 and ongoing.

Let $X \sim \mathcal{X}^2_m$ and $Y \sim \mathcal{X}^2_n$ be two independent random variables. The distribution of the random variable

$$W = \frac{X/m}{Y/n} \tag{3.10}$$

is called the F-distribution with m and n degrees of freedom and denoted $W \sim F_{m,n}$. It holds that:

$$\mathrm{E}(W) = \frac{n}{n-2}, \qquad \text{for } n > 2; \tag{3.11}$$

$$\mathrm{Var}(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \qquad \text{for } n > 4. \tag{3.12}$$

That means that if n increases the expectation gets closer to one and the variance to $2/m$, with m fixed. Figure 3.3 (based on R-Code 3.4) shows the density for various degrees of freedom.

3.3 Limit Theorems

The probability mass function of a binomial random variable with a large n and moderate p (i.e., p not too close to zero or one) has a bell shaped form. Similarly, the magenta density of a seemingly arbitrary random variable in Figure 3.1 looks very much like a Gaussian density. Coincidence?

No. The following theorem is of paramount importance in statistics and sheds some insight. In fact, the theorem gives us the distribution of $\overline{X}$ for virtually arbitrary distributions of X_i and goes much beyond Property 3.1!

R-Code 3.4 F-distribution for various degrees of freedom. (See Figure 3.3.)

```
x <- seq(0, to=4, length=500)            # x-values for which we plot
df1 <- c( 1, 2, 5, 10, 50, 50, 250)      # dof for the numerator
df2 <- c( 1, 50, 10, 50, 50, 250, 250)   # dof for the denumerator
plot( x, df( x, df1=1, df2=1), type='l', ylab='Density')
for (i in 2:length(df1))
    lines( x, df(x, df1=df1[i], df2=df2[i]), col=i)
legend( "topright", col=1:7, lty=1, bty="n",
        legend=parse(text=paste0("F['",df1,",",df2,"']")))
```

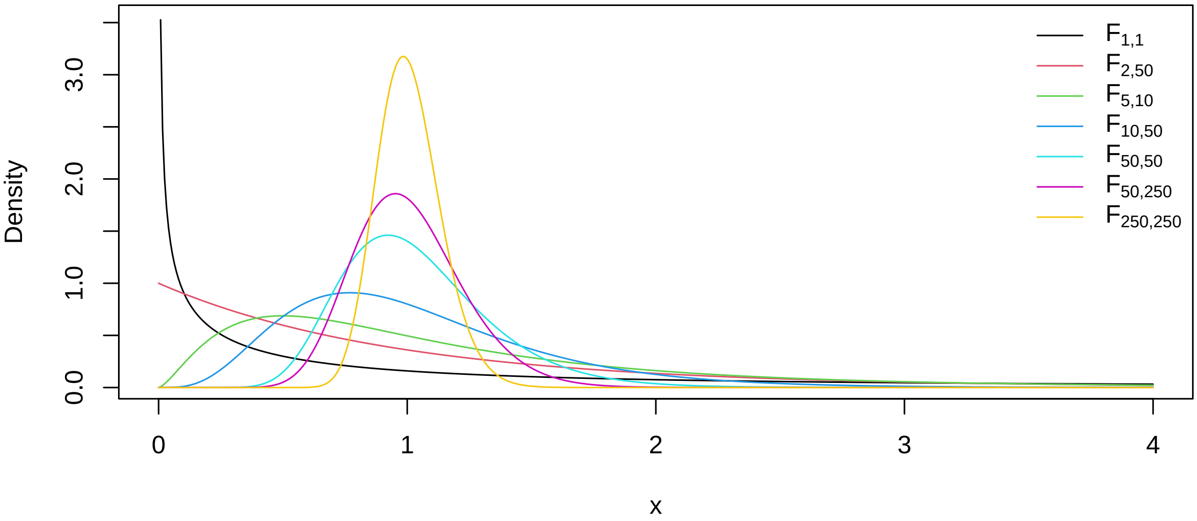


Figure 3.3: Density of the F-distribution for various degrees of freedom. (See R-Code 3.4.)

Property 3.5. *(Central Limit Theorem (CLT), classical version) Let $X_1, X_2, X_3, \dots$ an infinite sequence of iid random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then*

$$\lim_{n\to\infty} \mathrm{P}\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\Big) = \Phi(z) \tag{3.13}$$

where we kept the subscript n for the random sample mean to emphasis its dependence on n.

The proof of the CLT is a typical exercise in a probability theory lecture. Many extensions of the CLT exist, for example, the independence assumptions can be relaxed.

Using the central limit theorem argument, we can show that distribution of a binomial random variable $X \sim \mathcal{B}in(n, p)$ converges to a distribution of a normal random variable as $n \to \infty$. Thus, the distribution of a normal random variable $\mathcal{N}(np, np(1-p))$ can be used as an approximation for the binomial distribution $\mathcal{B}in(n, p)$. For the approximation, n should be larger than 30 for $p \approx 0.5$. For p closer to 0 and 1, n needs to be much larger.

For a binomial random variable, $\mathrm{P}(X \leq x) = \mathrm{P}(X < x+1)$, $x = 1, 2, \dots, n$, which motivates a a so-called *continuity correction* when calculating probabilities. Specifically, instead of $\mathrm{P}(X \leq x)$ we approximate $\mathrm{P}(X \leq x + 0.5)$ as illustrated in the following example.

Example 3.4. Let $X \sim \mathcal{B}in(30, 0.5)$. Then $\mathrm{P}(X \leq 10) = 0.0494$, "exactly". However,

$$\mathrm{P}(X \leq 10) \approx \mathrm{P}\Big(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{10 - np}{\sqrt{np(1-p)}}\Big) = \Phi\Big(\frac{10 - 15}{\sqrt{30/4}}\Big) = 0.0339\,, \tag{3.14}$$

$$\mathrm{P}(X \leq 10 + 0.5) \approx \mathrm{P}\Big(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{10 + 0.5 - np}{\sqrt{np(1-p)}}\Big) = \Phi\Big(\frac{10.5 - 15}{\sqrt{30/4}}\Big) = 0.05017\,. \tag{3.15}$$

The improvement of the continuity correction can be quantified by the reduction of the *absolute errors* $|0.0494 - 0.0339| = 0.0155$ vs $|0.0494 - 0.0502| = 0.0008$ or by the *relative errors* $|0.0494 - 0.0339|/0.0494 = 0.3138$ vs $|0.0494 - 0.0502|/0.0494 = 0.0162$. ♣

Another very important limit theorem is the *law of large numbers* (LLN) that essentially states that for $X_1, \dots, X_n$ iid with $\mathrm{E}(X_i) = \mu$, the average $\overline{X}_n$ converges to μ. We have deliberately used the somewhat ambiguous "convergence" statement, a more rigorous statement is technically a bit more involved. We will use the LLN in the next chapter, when we try to infer parameter values from data, i.e., say something about μ when we observe $x_1, \dots, x_n$.

Remark 3.3. There are actually two forms of the LLN theorem, the strong and the weak formulation. We do not not need the precise formulation later and thus simply state them here for the sole reason of stating them

$$\text{Weak LLN:} \ \lim_{n\to\infty} \mathrm{P}\big(|\overline{X}_n - \mu| > \epsilon\big) = 0 \text{ for all } \epsilon > 0, \tag{3.16}$$

$$\text{Strong LLN:} \ \mathrm{P}\big(\lim_{n\to\infty} \overline{X}_n = \mu\big) = 1. \tag{3.17}$$

The differences between both formulations are subtle. The weak version states that the average is close to the mean and excursions (for specific n) beyond $\mu \pm \epsilon$ can happen arbitrary often. The strong version states that there exists a large n such that the average is always within $\mu \pm \epsilon$.

The two forms represent fundamentally different notions of convergence of random variables: (3.17) is *almost sure convergence*, (3.16) is *convergence in probability*. The CLT represents *convergence in distribution*. ♣

We have observed several times that for increasing sample sizes, the discrepancy between the theoretical value and the sample diminishes. Examples include the CLT, LLN, but also our observation that the histogram of `rnorm(n)` looks like the normal density for large `n`.

It is possible to formalize this important concept. Let $X_1, \dots, X_n$ iid with distribution function $F(x)$. We define the *sample empirical distribution function*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{x_i \leq x}(x), \tag{3.18}$$

that is, a step function with jump size $1/n$ at the values $x_1, \dots, x_n$ (see also Equation 3.32 for the meaning of $\mathbb{I}$). As $n \to \infty$ the empirical distribution function $F_n(x)$ converges to the underlying distribution function $F(x)$. Because of this fundamental result we are able to work with specific distributions of random samples.

For discrete random variables, the previous convergence result can be written in terms of probabilities and observed proportions (see Problem 3.3.**b**). For continuous random variables, we would have to invoke binning to compare the histogram with the density. The binning adds another technical layer for the theoretical results. In practice, we simply compare the histogram (or a smoothed version thereof, Figure 1.3) with the theoretical density.

Remark 3.4. To show the convergence of the sample empirical distribution function, we consider the random version thereof, i.e., the *empirical distribution function* $1/n\sum_{i=1}^{n}\mathbb{I}_{X_i\leq x}(x)$ and invoke the LLN. This convergence is pointwise only, meaning it holds for all x. Stronger results hold and the *Glivenko–Cantelli theorem* states that the convergence is even uniform: $\sup_x\big|F_n(x)-F(x)\big|$ converges to zero almost surely. ♣

3.4 Functions of a Random Variable

In the previous sections we saw different examples of often used, classical random variables. These examples are often not enough and through a modeling approach, we need additional ones. In this section we illustrate how to construct the cdf and pdf of a random variable that is the square of one from which we know the density.

Let X be a random variable with distribution function $F_X(x)$. We define a random variable $Y = g(X)$, for a suitable chosen function $g(\cdot)$. The cumulative distribution function of Y is written as

$$F_Y(y) = P(Y \leq y) = P\big(g(X)\leq y\big). \tag{3.19}$$

In many cases $g(\cdot)$ is invertable (and differentiable) and we obtain

$$F_Y(y) = \begin{cases} \mathrm{P}\left(X\leq g^{-1}(y)\right) = F_X\big(g^{-1}(y)\big), & \text{if } g^{-1} \text{ monotonically increasing,}\\ \mathrm{P}\left(X\geq g^{-1}(y)\right) = 1-F_X\big(g^{-1}(y)\big), & \text{if } g^{-1} \text{ monotonically decreasing.}\end{cases} \tag{3.20}$$

To derive the probability mass function we apply Property 2.2.4. In the more interesting setting of continuous random variables, the density function is derived by Property 2.3.4 and is thus

$$f_Y(y) = \Big|\frac{d}{dy}g^{-1}(y)\Big| f_X(g^{-1}(y)). \tag{3.21}$$

Example 3.5. Let X be a random variable with cdf $F_X(x)$ and pdf $f_X(x)$. We consider $Y = a+bX$, for $b>0$ and a arbitrary. Hence, $g(\cdot)$ is a linear function and its inverse $g^{-1}(y) = (y-a)/b$ is monotonically increasing. The cdf of Y is thus $F_X\big((y-a)/b\big)$ and, for a continuous random variable X, the pdf is $f_X\big((y-a)/b\big)\cdot 1/b$. This fact has already been stated in Property 2.7 for the Gaussian random variables. ♣

This last example also motivates a formal definition of a location parameter and a scale parameter.

Definition 3.3. For a random variable X with density $f_X(x)$, we call θ a *location parameter* if the density has the form $f_X(x) = f(x-\theta)$ and call θ a *scale parameter* if the density has the form $f_X(x) = f(x/\theta)/\theta$. ♢

In the previous definition θ stands for an arbitrary parameter. The parameters are often chosen such that the expectation of the random variable is equal to the location parameter and the variance equal to the squared scale parameter. Hence, the following examples are no surprise.

Example 3.6. For a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, μ is the location parameter, σ a scale parameter, as seen from the definition of the density (2.24). The parameter λ of the Poisson distribution $\mathcal{P}ois(\lambda)$ is neither a location nor a scale parameter. ♣

We consider a second example of a transformation, closely related to Problem 2.8.

Example 3.7. Let $X \sim \mathcal{U}(0,1)$ and for $0 < x < 1$, we set $g(x) = -\log(1-x)$, where $\log()$ is the natural logarithm, i.e., the logarithm to the base e. Thus, $g^{-1}(y) = 1 - \exp(-y)$ and the distribution and density function of $Y = g(X)$ is

$$F_Y(y) = F_X(g^{-1}(y)) = g^{-1}(y) = 1 - \exp(-y), \tag{3.22}$$

$$f_Y(y) = \left|\frac{d}{dy} g^{-1}(y)\right| f_X\big(g^{-1}(y)\big) = \exp(-y), \tag{3.23}$$

for $y > 0$. This random variable is called the exponential random variable (with rate parameter one), denoted by $X \sim \mathcal{E}xp(\lambda)$ with $\lambda = 1$.

The random variable $V = Y/\lambda$, $\lambda > 0$, has the density $f_V(v) = \exp(-x/\theta)/\theta$ (by (3.21) or Example 3.5) and is denoted $V \sim \mathcal{E}xp(\lambda)$. Hence, the parameter $\theta = 1/\lambda$ of an exponential distribution is a scale parameter.

Notice further that $g(x)$ is the quantile function of this random variable. ♣

This last example gives rise to the so-called *inverse transform sampling* method to draw realizations from a random variable X with a closed form quantile function $Q_X(p)$. In more details, assume an arbitrary cumulative distribution function $F_X(x)$ with a closed form inverse $F_X^{-1}(p) = Q_X(p)$. For $U \sim \mathcal{U}(0,1)$, the random variable $X = F^{-1}(U)$ has cdf $F_X(x)$:

$$\text{P}(X \leq x) = \text{P}\big(F_X^{-1}(U) \leq x\big) = \text{P}\big(U \leq F_X(x)\big) = F_X(x). \tag{3.24}$$

Hence, based on Example 3.7, the R expression `-log(1- runif(n))/lambda` draws a realization from $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{E}xp(1)$.

Remark 3.5. For $U \sim \mathcal{U}(0,1)$, $1 - U \sim \mathcal{U}(0,1)$, and thus it is possible to further simplify the sampling procedure. Interestingly, R uses a seemingly more complex algorithm to sample. The algorithm is however fast and does not require a lot of memory (Ahrens and Dieter, 1972); properties that were historically very important. ♣

In the case, where we cannot invert the function g, we can nevertheless use the concept of the approach by starting with (3.19), followed by simplification and use of Property 2.3.4, as illustrated in the following example.

Example 3.8. The density of a chi-squared distributed random variable with one degree of freedom is calculated as follows. Let $X \sim \mathcal{N}(0,1)$ and $Y = Z^2$.

$$F_Y(y) = \mathrm{P}(Y = Z^2 \leq y) = \mathrm{P}\big(\mid Z \mid \leq \sqrt{y}\big) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1 \tag{3.25}$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = 2\phi(\sqrt{y})\frac{d}{dy}\sqrt{y}\frac{2}{\sqrt{2\pi}}\exp\Big(-\frac{y}{2}\Big)\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}\exp\Big(-\frac{y}{2}\Big). \tag{3.26}$$

♣

As we are often interested in summarizing a random variable by its mean and variance, we now introduce a very convenient approximation for transformed random variables $Y = g(X)$ by the so-called *delta method*. The idea thereof consists of a Taylor expansion around the expectation $\mathrm{E}(X)$:

$$Y = g(X) \approx g\big(\mathrm{E}(X)\big) + g'\big(\mathrm{E}(X)\big) \cdot \big(X - \mathrm{E}(X)\big) \tag{3.27}$$

(first two terms of the *Taylor series*). Thus

$$\mathrm{E}(Y) \approx g\big(\mathrm{E}(X)\big), \tag{3.28}$$

$$\mathrm{Var}(Y) \approx g'\big(\mathrm{E}(X)\big)^2 \cdot \mathrm{Var}(X). \tag{3.29}$$

The approach is illustrated with the following two examples.

Example 3.9. Let $X \sim \mathcal{B}(1,p)$ and $Y = X/(1-X)$. Thus,

$$\mathrm{E}(Y) \approx \frac{p}{1-p}; \qquad \mathrm{Var}(Y) \approx \Big(\frac{1}{(1-p)^2}\Big)^2 \cdot p(1-p) = \frac{p}{(1-p)^3}. \quad ♣ \tag{3.30}$$

Example 3.10. Let $X \sim \mathcal{B}(1,p)$ and $Y = \log(X)$. Thus

$$\mathrm{E}(Y) \approx \log(p), \qquad \mathrm{Var}(Y) \approx \Big(\frac{1}{p}\Big)^2 \cdot p(1-p) = \frac{1-p}{p}. \quad ♣ \tag{3.31}$$

Of course, in the case of a linear transformation (as, e.g., in Example 3.5), equation (3.27) is an equality and thus relations (3.28) and (3.29) are exact, which is in sync with Property 2.7.

Consider $g(X) = X^2$, by Property 2.5.1, $\mathrm{E}(X^2) = \mathrm{E}(X)^2 + \mathrm{Var}(X)$ and thus $\mathrm{E}(X^2) > \mathrm{E}(X)^2$, refining the approximation (3.28). This result can be generalized and states that for every convex function $g(\cdot)$, $\mathrm{E}\big(g(X)\big) \leq g\big(\mathrm{E}(X)\big)$. For concave functions, the inequality is reversed. For strictly convex or strictly concave functions, we have strict inequalities. Finally, a linear function is not concave or convex and we have equality, as given in Property 2.5.4. These inequalities run under *Jensen's inequality*.

Remark 3.6. The following results are for completeness only. They are nicely elaborated in Rice (2006).

1. For continuous random variables X_1 and X_2 it is possible to calculate probability density function of $X_1 + X_2$. The formula is the so-called *convolution*. Same holds for discrete random variables.

2. It is also possible to construct random variables based on an entire random sample, say $Y = g(X_1, \dots, X_n)$. Property 3.5 uses exactly such an approach, where $g(\cdot)$ is given by $g(X_1, \dots, X_n) = \big(\sum_i X_i - \mu\big)\Big/\big(\sigma/\sqrt{n}\big)$.

3. Starting with two random continuous variables X_1 and X_2, and two bijective functions $g_1(X_1, X_2)$ and $g_2(X_1, X_2)$, there exists a closed form expression for the (joint) density of $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$.

♣

We finish the chapter with introducing a very simple but convenient function, the so-called *indicator function*, defined by

$$\mathbb{I}_{x\in A}(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases} \tag{3.32}$$

The function argument '(x)' is redundant but often helps to clarify. In the literature, one often finds the concise notation $\mathbb{I}_A$. Here, let X be a random variable then we define a random variable $Y = \mathbb{I}_{X\in A}(X)$, i.e., Y 'specifies' if the random variable X is in the set A.

The indicator function is often used to calculate expectations. For example, let Y be defined as above. Then

$$\mathrm{E}(Y) = \mathrm{E}\big(\mathbb{I}_{X>0}(X)\big) = 0 \cdot \mathrm{P}(X \le 0) + 1 \cdot \mathrm{P}(X > 0) = \mathrm{P}(X > 0) \tag{3.33}$$

where we used Property 2.7 with $g(x) = \mathbb{I}_{x\in A}(x)$. In other words, we see $\mathbb{I}_{X>0}(X)$ as a Bernoulli random variable with success probability $P(X > 0)$.

3.5 Bibliographic Remarks

Limit theorems are part of any mathematical statistics or probability book and similar references as in Section 2.6 can be added here.

Needham (1993) gives a graphical justification of Jensen's inequality.

The derivation of the chi-square, t- and F-distribution is accessibly presented in Rice (2006) and their properties extensively discussed in Johnson *et al.* (1994, 1995).

3.6 Exercises and Problems

Problem 3.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Show that for $X \sim \mathcal{Pois}(\lambda)$, $\mathrm{E}(X) = \lambda$, and $\mathrm{Var}(X) = \lambda$.

b) Starting from the pmf of a Binomial random variable, derive the pmf of the Poisson random variable when $n \to \infty$, $p \to 0$ but $\lambda = np$ constant.

Problem 3.2 (Geometric distribution) In the setting of Examples 2.3 and 3.2, denote $p = \mathrm{P}(\text{shot is successful})$. Assume that the individual shots are independent. Show that

a) $\mathrm{P}(X \leq k) = 1 - (1-p)^k$, $k = 1, 2, \dots$.

b) $\mathrm{E}(X) = 1/p$ and $\mathrm{Var}(X) = (1-p)/p^2$.

c) $\mathrm{P}(X = k + j \mid X > j) = \mathrm{P}(X = k)$.

Problem 3.3 (Poisson distribution) In this problem we visualize and derive some properties of the Poisson random variable with parameter $\lambda > 0$.

a) Visualize in R the cdf and pmf of $X \sim \mathcal{Pois}(\lambda)$, for $\lambda = 0.2$ and $\lambda = 2$.

b) For $\lambda = 0.2$ and $\lambda = 2$, sample from $X_1, \dots, X_n \sim \mathcal{Pois}(\lambda)$ with $n = 200$ and draw histograms. Compare the histograms with **a**). What do you expect to happen when n increases?

c) Let $X \sim \mathcal{Pois}(\lambda)$, with $\lambda = 3$: calculate $\mathrm{P}(X \leq 2)$, $\mathrm{P}(X < 2)$ and $\mathrm{P}(X \geq 3)$.

d) Plot the pmf of $X \sim \mathcal{Pois}(\lambda)$, $\lambda = 5$, and $Y \sim \mathcal{Bin}(n, p)$ for $n = 10, 100, 1000$ with $\lambda = np$. What can you conclude?

Problem 3.4 (Approximating probabilities) In the following settings, approximate the probabilities and quantiles by reducing the problem to calculating probabilities from a standard normal, i.e., using a "standard table". Compare these values with the ones obtained with R.
If you do not have standard table, the following two R commands may be used instead: `pnorm(a)` and `qnorm(b)` for specific values `a` and `b`.

a) $X \sim \mathcal{Bin}(42, 5/7)$: $\mathrm{P}(X = 30)$, $\mathrm{P}(X < 21)$, $\mathrm{P}(24 \leq X \leq 35)$.

b) $X \sim t_{49}$: $\mathrm{P}(-1 < X < 1)$, $\mathrm{P}(X < q) = 0.05$.

c) $X \sim \chi^2_{68}$: $\mathrm{P}(X \leq 68)$, $\mathrm{P}(47 \leq X \leq 92.7)$

Problem 3.5 (Random numbers from a t-distribution) We sample random numbers from a t-distribution by drawing repeatedly $z_1, \dots, z_m$ from a standard normal and setting $y_j = \bar{z}/\sqrt{s^2/m}$, $j = 1, \dots, n$, where $\bar{z} = 1/m \sum_i z_i$ and $s^2 = 1/(m-1) \sum_i (z_i - \bar{z})^2$.

a) For $m = 6$ and $n = 500$ construct Q-Q plots based on the normal and appropriate t-distribution. Based on the plots, is it possible to discriminate which of the two distributions is more appropriate? Is this question getting more difficult if m and n are chosen larger or smaller?

b) Suppose we receive the sample $y_1, \dots, y_n$ constructed as above but the value m is not disclosed. Is it possible to determine m based on Q-Q plots? What general implication can be drawn, especially for small samples?

Problem 3.6 (Exponential Distribution in R) Let $X_1, \dots, X_n$ be independent and identically distributed (iid) random variables following $\mathcal{E}xp(\lambda)$. Assume $\lambda = 2$ for the following.

a) Sample $n = 100$ random numbers from $\mathcal{E}xp(\lambda)$. Visualize the data with a histogram and superimpose the theoretical density.

b) Derive theoretically the distribution of $\min(X_1, \dots, X_n)$.

c) Draw a histogram of $\min(X_1, \dots, X_n)$ from 500 realizations and compare it to the theoretical result from part b).

Problem 3.7 (Inverse transform sampling) The goal of this exercise is to implement your own R code to simulate from a continous random variable X with the following probability density function (pdf):

$$f_X(x) = c\,|x|\,\exp\left(-x^2\right), \quad x \in \mathbb{R}.$$

We use inverse transform sampling, which is well suited for distributions whose cdf is easily invertible.

Hint: you can get rid of the absolute value by defining

$$f_X(x) = \begin{cases} c\,x\,\exp\left(-x^2\right), & \text{if } x \geq 0, \\ -c\,x\,\exp\left(-x^2\right), & \text{if } x < 0. \end{cases}$$

a) Find c such that $f_X(x)$ is an actual pdf. Show that the quantile function is

$$F_X^{-1}(p) = Q_X(p) = \begin{cases} \sqrt{-\log\left(2(1-p)\right)}, & p \geq 0.5, \\ -\sqrt{-\log\left(2p\right)}, & p < 0.5. \end{cases}$$

b) Write your own code to sample from X. Check the correctness of your sampler.

Problem 3.8 (Geometric mean) The geometric mean of $x_1, \dots, x_n$ is defined as

$$m_g = m_g(x_1, \dots, x_n) = \sqrt[n]{x_1 \cdot \dots \cdot x_n}$$

and is often used to summarize rates or to summarize different items that are rated on different scales.

a) For $x_1, \dots, x_n$ and $y_1, \dots, y_n$, show that $m_g(x_1/y_1, \dots, x_n/y_n) = \dfrac{m_g(x_1, \dots, x_n)}{m_g(y_1, \dots, y_n)}$. What is a profound consequence of this property?

b) Using Jensen's inequality, show that $\sqrt[n]{x_1 \cdot \dots \cdot x_n} \leq \frac{1}{n}(x_1 + \dots + x_n)$, i.e., $m_g \leq \bar{x}$.

Chapter 4

Estimation of Parameters

Learning goals for this chapter:

- Explain what a simple statistical model is (including the role of parameters)
- Describe the concept of point estimation and interval estimation
- Interpret point estimates and confidence intervals
- Describe the concept of method of moments, least squares and likelihood estimation
- Construct theoretically and using R confidence intervals

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter04.R.

A central point of statistics is to draw information from observations (in form of data from measurements of an experiment or from a subset of a population) and to infer from these towards hypotheses in form of scientific questions or general statements of the population. Inferential statistics is often explained by "inferring" conclusions from the sample toward the "population". This step requires foremost data and an adequate *statistical model*. Such a statistical model describes the data through the use of random variables and their distributions, by means the data is considered a realization of the corresponding distribution. The model comprises of unknown quantities, so-called parameters. The goal of a *statistical estimation* is to determine plausible values for the parameters of the model using the observed data.

4.1 Linking Data with Parameters

In this section we first introduce the concept of statistical models and the associated parameters. In the second step we introduce the idea of estimation.

4.1.1 Statistical Models and Parameters

We now discuss the box 'Propose a statistical model' in our statistical workflow (Figure 1.1). To introduce a first statistical model we consider a specific example.

Example 4.1. Antimicrobial susceptibility tests (ASTs) are used to identify the level of resistance of a bacterial strain to a specific antibiotic. The procedure can be summarized by the following steps. The bacteria are isolated and a resulting bacterial suspension is applied (streaked) on a Petri dish that contains a growth medium and nutrients for the bacteria. The antibiotics sources are added and the dish is then incubated, that means kept for several hours at a favorable temperature to allow bacterial growth. If the antibiotic works well, the sources inhibit the growth and create small circular areas that do not contain any bacterial load. The larger the disk, the more effective the bacteria.

To assess the variability and errors of the AST testing procedure, 100 measurements based on 100 suspensions have been evaluated for different antibiotics Hombach *et al.* (2016). Table 4.1 reports the observed diameters with frequencies by E. coli and the antibiotics imipenem and meropenem.

Natural questions that arise are: What are plausible or representative values of the inhibition diameter? How much do the individuals diameters deviate around the mean?

The data is visualized in Figure 4.1 (based on R-Code 4.1). ♣

Table 4.1: Inhibition diameters by E. coli and the antibiotics imipenem and meropenem.

Diameter (mm)	28	29	30	31	32	33	34	35	36	37	38	39	40
Imipenem	0	3	7	14	32	20	18	4	1	1	0	0	0
Meropenem	0	0	0	0	2	9	33	20	17	9	6	4	0

Although the data of the previous example is rounded to the nearest millimeter, it would be reasonable to assume that the diameters are real valued. Each measurement being identical to others and thus fluctuating naturally around a common mean. The following is a very simple example of a statistical model adequate here:

$$Y_i = \mu + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4.1}$$

where Y_i are the observations, μ is an unknown diameter and ε_i are random variables representing measurement error. It is often reasonable to assume $\mathrm{E}(\varepsilon_i) = 0$ with a symmetric density. Here, we further assume $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Thus, $Y_1, \dots, Y_n$ are normally distributed with mean μ and variance σ^2. As typically, both parameters μ and σ^2 are unknown and we need to determine plausible values for these parameters from the available data.

Such a statistical model describes the data $y_i, \dots, y_n$ through random variables $Y_i, \dots, Y_n$ and their distributions. In other words, the data is a realization of the random sample given by the statistical model.

For both antibiotics, a slightly more evolved example consists of

$$Y_i = \mu_{\text{imi}} + \varepsilon_i, \qquad i = 1, \dots, n_{\text{imi}}, \tag{4.2}$$

$$Y_i = \mu_{\text{mero}} + \varepsilon_i, \qquad i = n_{\text{imi}} + 1, \dots, n_{\text{imi}} + n_{\text{mero}}, \tag{4.3}$$

We assume $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Thus the model states that both diseases have a different mean but the same variability. This assumption pools information from both samples to estimate the variance σ^2. The parameters of the model are μ_{imi}, μ_{mero} and to a lesser extend σ^2.

The statistical model represents the population with a seamingly infinite size. The data is the realization of a subset of our population. And thus based on the data we want to answer questions like: What are plausible values of the population levels? How much do the individuals deviate around the population mean?

R-Code 4.1 states the means which we might consider as reasonable values for μ_{imi}, μ_{mero}, similar for their variances. Of course we need to formalize that the mean of the sample can be taken as a representative value of the entire population. For the variance parameter σ^2, slightly more care is needed as neither *var(imipDat)* nor *var(meropDat)* are fully satisfying.

The questions if the inhibition diameter of both antibiotics are comparable or if the inhibition diameter from meropenem is (statistically) larger than 33 mm are of completely different nature and will be discussed in the next chapter, where we formally discuss statistical tests.

R-Code 4.1 Inhibition diameters (See Figure 4.1.)

```
diam <- 28:40     # diameters
imi  <- c(0, 3, 7, 14, 32, 20, 18, 4, 1, 1, 0, 0 ,0)  # frequencies
mero <- c(0, 0, 0, 0,  2, 9, 33, 20, 17, 9, 6, 4, 0)
barplot( imi,  names.arg=paste(diam), main="Imipenem")
barplot( mero, names.arg=paste(diam), main="Meropenem")
imiDat <- rep(diam, imi)    # now a vector with the 100 diameters
c( mean(imiDat), sum( mero*diam)/100)  # means for both, then spread
## [1] 32.40 35.12
c( var( imiDat), sum( (imiDat-mean(imiDat))^2)/(length(imiDat)-1) )
## [1] 2.2424 2.2424
c( sd( imiDat), sqrt( var( imiDat)))
## [1] 1.4975 1.4975
```

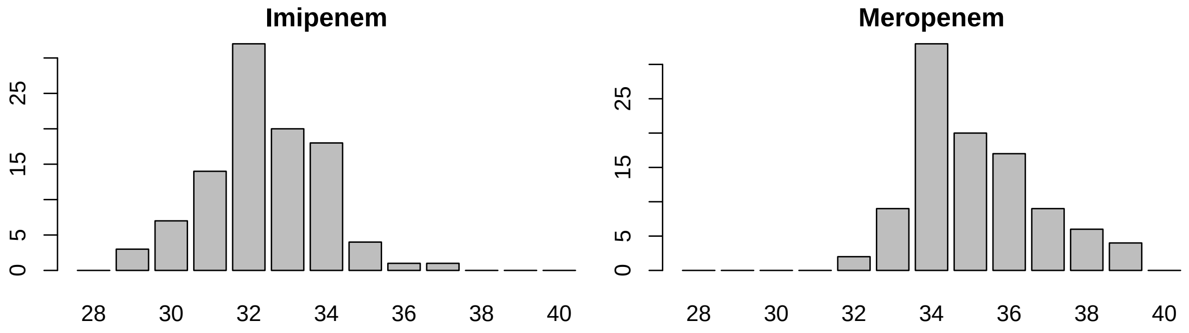


Figure 4.1: Frequencies of inhibition diameters (in mm) by E. coli and imipenem and meropenem (total 100 measurements). (See R-Code 4.1.)

4.1.2 Point Estimation

We assume now that an appropriate statistical model parametrized by one or several parameters has been proposed for a dataset. For estimation, the data is seen as a realization of a random sample where the distribution of the latter is given by the statistical model. The goal of point estimation is to provide a plausible value for the parameters of the distribution based on the data at hand.

Definition 4.1. A *statistic* is an arbitrary function of a random sample $Y_1, \dots, Y_n$ and is therefore also a random variable.

An *estimator* for a particular parameter is a statistic used to obtain a plausible value for this parameter, based on the random sample.

A *point estimate* is the value of the estimator evaluated at $y_1, \dots, y_n$, the realizations of the random sample.

Estimation (or *estimating*) is the process of finding a (point) estimate. ♢

Hence, in order to estimate a parameter, we start from an estimator for that particular parameter and evaluate the estimator at the available data. The estimator may depend on the underlying statistical model and typically depends on the sample size.

Example 4.2. 1. The numerical values shown in R-Code 4.1 are estimates.

2. $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ is an estimator.

   $\bar{y} = \dfrac{1}{100}\sum_{i=1}^{100} y_i = 32.40$ is a point estimate.

3. $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is an estimator.

   $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{100}(y_i - \bar{y})^2 = 2.242$ or $s = 1.498$ are a point estimates. ♣

Often, we denote parameters with Greek letters (μ, σ, λ, ...), with θ being the generic one. The estimator and estimate of a parameter θ are denoted by $\widehat{\theta}$. Context makes clear which of the two cases is meant.

4.2 Construction of Estimators

We have already encountered several estimators for location and spred, cf. Equations (1.1) to (1.4). In fact, there are an abundance of different estimators for a specific parameter. We now consider three approaches to construct estimators. Not all three work in all settings, more so, depending on the situations a particular approach might be preferable.

4.2.1 Ordinary Least Squares

The first approach is intuitive and it is straightforward to construct estimators for location parameters or, more generally, for parameters linked to the expectation of the random sample.

The ordinary least squares method of parameter estimation minimizes the sum of squares of the differences between the random variables and the location parameter. More formally, let $Y_1, \dots, Y_n$ be iid with $\mathrm{E}(Y_i) = \mu$. The least squares estimator for μ is

$$\widehat{\mu} = \widehat{\mu}_{\mathrm{LS}} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mu)^2, \tag{4.4}$$

and thus after minimizing the sums of squares we get the estimator $\widehat{\mu}_{\mathrm{LS}} = \overline{Y}$ and the estimate $\widehat{\mu}_{\mathrm{LS}} = \bar{y}$.

Often, the parameter θ is linked to the expectation $\mathrm{E}(Y_i)$ through some function, say g. In such a setting, we have

$$\widehat{\theta} = \widehat{\theta}_{\mathrm{LS}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \big(Y_i - g(\theta)\big)^2. \tag{4.5}$$

and $\widehat{\theta}_{\mathrm{LS}}$ solves $g(\widehat{\theta}) = \overline{Y}$.

In linear regression settings, the ordinary least squares method minimizes the sum of squares of the differences between observed responses and those predicted by a linear function of the explanatory variables. Due to the linearity, simple and close form solutions exist (see Chapters 9ff).

4.2.2 Method of Moments

The method of moments is based on the following idea. The parameters of the distribution are expressed as functions of the moments, e.g., $\mathrm{E}(Y)$, $\mathrm{E}(Y^2)$. The random sample moments are then plugged into the theoretical moments of the equations in order to obtain the estimators:

$$\mu := \mathrm{E}(Y)\,, \qquad\qquad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}\,, \tag{4.6}$$

$$\mu_2 := \mathrm{E}(Y^2)\,, \qquad\qquad \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2\,. \tag{4.7}$$

By using the observed values of a random sample in the method of moments estimator, the estimates of the corresponding parameters are obtained. If the parameter is a function of the moments, we need to additionally solve the corresponding equation, as illustrated in the following two examples.

Example 4.3. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$

$$\mathrm{E}(Y) = 1/\lambda, \qquad \overline{Y} = 1\Big/\widehat{\lambda} \qquad \widehat{\lambda} = \widehat{\lambda}_{\mathrm{MM}} = \frac{1}{\overline{Y}}\,. \tag{4.8}$$

Thus, the method of moment estimate of λ is the value $1/\bar{y}$. ♣

Example 4.4. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} F$ with expectation μ and variance σ^2. Since $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2$ (Property 2.5.1), we can write $\sigma^2 = \mu_2 - (\mu)^2$ and we have the estimator

$$\widehat{\sigma^2_{\mathrm{MM}}} = \widehat{\mu_2} - (\widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{4.9}$$

♣

4.2.3 Likelihood Method

The likelihood method chooses as estimate the value such that the observed data is most likely to stem from the model (using the estimate). To derive the method, we consider the probability density function (for continuous random variables) or the probability mass function (for discrete random variables) to be a function of parameter θ, i.e.,

$$f_Y(y) = f_Y(y;\theta) \qquad \longrightarrow \qquad L(\theta) := f_Y(y;\theta) \tag{4.10}$$

$$p_i = \mathrm{P}(Y = y_i) = \mathrm{P}(Y = y_i;\theta) \qquad \longrightarrow \qquad L(\theta) := \mathrm{P}(Y = y_i;\theta). \tag{4.11}$$

For a given distribution, we call $L(\theta)$ the likelihood function, or simply the likelihood.

Definition 4.2. The maximum likelihood estimate $\widehat{\theta}_{\mathrm{ML}}$ of the parameter θ is based on maximizing the likelihood, i.e.

$$\widehat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} L(\theta). \qquad \diamondsuit \tag{4.12}$$

By definition of a random sample, the random variables are independent and identically distributed and thus the likelihood is the product of the individual densities $f_Y(y_i;\theta)$ (see Section 3.1). To simplify the notation, we have omitted the index of Y. Since $\widehat{\theta}_{\mathrm{ML}} = \operatorname{argmax}_\theta L(\theta) = \operatorname{argmax}_\theta \log\big(L(\theta)\big)$, the log-likelihood $\ell(\theta) := \log\big(L(\theta)\big)$ can be maximized instead. The log-likelihood is often preferred because the expressions simplify more and maximizing sums is much easier than maximizing products.

Example 4.5. Let $Y_1,\dots,Y_n \overset{\text{iid}}{\sim} \mathcal{E}xp(\lambda)$, thus

$$L(\lambda) = \prod_{i=1}^n f_Y(y_i) = \prod_{i=1}^n \lambda \exp(-\lambda y_i) = \lambda^n \exp(-\lambda \sum_{i=1}^n y_i)\,. \tag{4.13}$$

Then

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d\log(\lambda^n \exp(-\lambda \sum_{i=1}^n y_i)}{d\lambda} = \frac{d(n\log(\lambda) - \lambda \sum_{i=1}^n y_i)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n y_i \overset{!}{=} 0 \tag{4.14}$$

$$\widehat{\lambda} = \widehat{\lambda}_{\mathrm{ML}} = \frac{n}{\sum_{i=1}^n y_i} = \frac{1}{\bar{y}}. \tag{4.15}$$

In this case (as in some others), $\widehat{\lambda}_{\mathrm{ML}} = \widehat{\lambda}_{\mathrm{MM}}$. ♣

In a vast majority of cases, maximum likelihood estimators posses very nice properties. Intuitively, because we use information about the density and not only about the moments, they are "better" compared to method of moment estimators and to least squares based estimators. Further, for many common random variables, the likelihood function has a single optimum, in fact a maximum, for all permissible θ.

In our daily live we often have estimators available and thus we rarely need to rely on the approaches presented in this section.

4.3 Comparison of Estimators

The variance estimate based on the method of moment estimator in Example 4.4 divides the sum of the squared deviances by n, whereas Equation (1.3) and and R use the denominator $n-1$ (see R-Code 4.1 or Example 4.1). There are different estimators for a particular parameter possible and we now introduce two measures to compare them.

Definition 4.3. An estimator $\widehat{\theta}$ of a parameter θ is *unbiased* if

$$\mathrm{E}(\widehat{\theta}) = \theta, \tag{4.16}$$

otherwise it is biased. The value $\mathrm{E}(\widehat{\theta}) - \theta$ is called the *bias*. ♢

Simply put, an unbiased estimator leads to estimates that are on the long run correct.

Example 4.6. $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

1. $\overline{Y}$ is unbiased for μ, since

$$\mathrm{E}(\overline{Y}) = \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i) = \frac{1}{n} n\,\mathrm{E}(Y_1) = \mu\,. \tag{4.17}$$

2. $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is unbiased for σ^2. To show this, we expand the square terms and simplify the sum of these cross-terms:

$$(Y_i - \overline{Y})^2 = (Y_i - \mu + \mu - \overline{Y})^2 = (Y_i - \mu)^2 + 2(Y_i - \mu)(\mu - \overline{Y}) + (\mu - \overline{Y})^2, \tag{4.18}$$

$$\begin{aligned}\sum_{i=1}^{n} 2(Y_i - \mu)(\mu - \overline{Y}) &= 2(\mu - \overline{Y})\sum_{i=1}^{n}(Y_i - \mu) = 2(\mu - \overline{Y})(n\overline{Y} - n\mu)\\ &= -2n(\mu - \overline{Y})^2,\end{aligned} \tag{4.19}$$

Collecting the terms yields

$$(n-1)S^2 = \sum_{i=1}^{n}(Y_i - \mu)^2 - 2n(\mu - \overline{Y})^2 + n(\mu - \overline{Y})^2. \tag{4.20}$$

We now use $\mathrm{E}(Y) = \mathrm{E}(\overline{Y}) = \mu$, thus $\mathrm{E}\big((Y_i - \mu)^2\big) = \mathrm{Var}(Y_i) = \sigma^2$, and similarly, $\mathrm{E}\big((\mu - \overline{Y})^2\big) = \mathrm{Var}(\overline{Y}) = \sigma^2/n$, by Property 3.1.3. Finally,

$$(n-1)\,\mathrm{E}(S^2) = \sum_{i=1}^{n}\mathrm{Var}(Y_i) - n\,\mathrm{Var}(\overline{Y}) = n\sigma^2 - n\cdot\frac{\sigma^2}{n} = (n-1)\sigma^2\,. \tag{4.21}$$

3. $\widehat{\sigma^2} = \dfrac{1}{n}\displaystyle\sum_i (Y_i - \overline{Y})^2$ is biased for σ^2, since

$$\mathrm{E}(\widehat{\sigma^2}) = \frac{1}{n}(n-1)\underbrace{\mathrm{E}\Big(\frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2\Big)}_{\mathrm{E}(S^2) = \sigma^2} = \frac{n-1}{n}\sigma^2. \tag{4.22}$$

The bias is

$$\mathrm{E}(\widehat{\sigma^2}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2, \tag{4.23}$$

which amounts to a slight underestimation of the variance. ♣

Unbiasedness is a nice and often desired property of an estimator. If an estimator is biased and this bias is known, then it is often possible to correct for it. In the spirit of Example 4.6.3, suppose that we have a biased estimator $\widehat{\theta}$ with $\mathrm{E}(\widehat{\theta}) = a\theta$, leading to a bias $(a-1)\theta$. Then the estimator $\widehat{\theta}/a$ is unbiased.

A second possibility for comparing estimators is the *mean squared error*

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{E}\Big((\widehat{\theta} - \theta)^2\Big). \tag{4.24}$$

The mean squared error can also be written as $\mathrm{MSE}(\widehat{\theta}) = \mathrm{bias}(\widehat{\theta}\,)^2 + \mathrm{Var}(\widehat{\theta}\,)$.

Example 4.7. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Using the result (4.17) and Property 3.1.3, we have

$$\mathrm{MSE}(\overline{Y}) = \mathrm{bias}(\overline{Y})^2 + \mathrm{Var}(\overline{Y}) = 0 + \frac{\sigma^2}{n}. \tag{4.25}$$

Hence, the MSE vanishes as n increases. ♣

There is a another "classical" example for the calculation of the mean squared error, however it requires some properties of squared Gaussian variables.

Example 4.8. If $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ then $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$ (see Property 3.4.1). Using the variance expression from a chi-squared random variable (see Equation (3.6)), we have

$$\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) = \frac{\sigma^4}{(n-1)^2}\,\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^4}{(n-1)^2}\big(2(n-1)\big) = \frac{2\sigma^4}{n-1}. \tag{4.26}$$

Analogously, one can show that $\mathrm{MSE}(\widehat{\sigma^2_{\mathrm{MM}}})$ is smaller than Equation (4.26). Moreover, the estimator $(n-1)S^2/(n+1)$ possesses the smallest MSE (see Problem 4.1.**b**). ♣

Remark 4.1. In both examples above, the variance has order $\mathcal{O}(1/n)$. In practical settings, it is not possible to get a better rate. In fact, there is a lower bound for the variance that cannot be undercut (the bound is called the *Cramér–Rao lower bound*). Ideally, we aim to construct and use *minimal variance unbiased estimators* (MVUE). Such bounds and properties are studied in mathematical statistics lectures and treatise. ♣

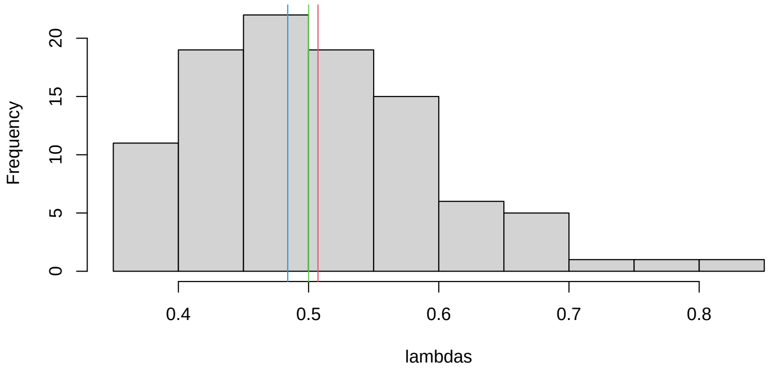


Figure 4.2: 100 estimates of the parameter $\lambda = 1/2$ based on a sample of size $n = 25$. The red and blue vertical line indicate the sample mean and median of the estimates respectively. (See R-Code 4.2.)

4.4 Interval Estimators

The estimates considered so far are point estimates as they are single values and do not provide us with uncertainties. For this we have to extend the idea towards interval estimates and interval estimators. We start with a motivating example followed by a couple analytic manipulations to finally introduce the concept of a confidence interval.

Example 4.9. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}xp(\lambda)$ and we take $n = 25$ and $\lambda = 1/2$. According to Example 4.3, $\widehat{\lambda} = 1/\overline{Y}$ is a legitimate estimator (possibly biased due to Jensen's inequality). If we have different samples, our estimates will not be exactly 1/2, but close to it, as nicely illustrated with R-Code 4.2. If we repeat and draw additional samples, we notice the variability in the estimates. Overall, we are quite close to the actual value, but individual estimates may be quite far from the truth, 90% of the estimates are within the interval $[0.39, 0.67]$. ♣

R-Code 4.2 Variability in estimates. (See Figure 4.2.)

```
set.seed(14)                # we work reproducible
n <- 25                     # sample size
R <- 100                    # number of estimates
lambda <- 1/2               # true value to estimate
samples <- matrix( rexp( n*R, rate=lambda), n, R)
lambdas <- 1/colMeans( samples)  # actual estimates
hist( lambdas, main='')          # unimodel, but not quite symmetric
abline( v=c(lambda, mean(lambdas), median(lambdas)), col=c(3,2,4))
quantile(lambdas, probs=c(.05,.95))
##     5%    95%
## 0.3888 0.6693
```

In practice we have one set of observations and thus we cannot repeat sampling to get a description of the variability in our estimate. Therefore we apply a different approach by attaching an uncertainty to the estimate itself, which we now illustrate in the Gaussian setting.

4.4.1 Confidence Intervals in the Gaussian Setting

Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2. Thus $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and, by standardizing, we have $(\overline{Y} - \mu)\big/\sqrt{\sigma^2/n} \sim \mathcal{N}(0,1)$. As a consequence,

$$1 - \alpha = \mathrm{P}\Big(z_{\alpha/2} \le \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2} \Big) = \mathrm{P}\Big(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{Y} - \mu \le z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Big) \tag{4.27}$$

$$= \mathrm{P}\Big(-\overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Big) \tag{4.28}$$

$$= \mathrm{P}\Big(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Big) \tag{4.29}$$

$$= \mathrm{P}\Big(\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Big), \tag{4.30}$$

where z_p is the p-quantile of the standard normal distribution $\mathrm{P}(Z \le z_p) = p$ for $Z \sim \mathcal{N}(0,1)$ (recall that $z_p = -z_{1-p}$).

The manipulations on the inequalities in the previous derivation are standard but the probabilities in (4.28) to (4.30) seem at first sight a bit strange, they should be read as $\mathrm{P}\big(\{\overline{Y} - a \le \mu\} \cap \{\mu \le \overline{Y} + a\}\big)$.

Based on Equation (4.30) we now define an interval estimator.

Definition 4.4. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2. The interval

$$\Big[\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Big] \tag{4.31}$$

is an exact $(1-\alpha)$ *confidence interval* for the parameter μ. $1-\alpha$ is the called the *level* of the confidence interval. ♢

If we evaluate the bounds of the confidence interval $\big[\, B_l,\, B_u \,\big]$ at a realization we denote $\big[\, b_l,\, b_u \,\big]$ as the sample confidence interval or the observed confidence interval.

The interpretation of an exact confidence interval is as follows: If a large number of realizations are drawn from a random sample, on average $(1-\alpha)\cdot 100\%$ of the confidence intervals will cover the true parameter μ.

Sample confidence intervals do not contain random variables and therefore it is not possible to make probability statements about the parameter.

If the standard deviation σ is unknown, the approach above must be modified by using a point estimate for σ, typically $S = \sqrt{S^2}$ with $S^2 = 1/(n-1)\sum_i (Y_i - \overline{Y})^2$. Since $(\overline{Y} - \mu)\big/\sqrt{S^2/n}$ has a t-distribution with $n-1$ degrees of freedom (see Property 3.4.3), the corresponding quantile must be modified:

$$1 - \alpha = \mathrm{P}\Big(t_{n-1,\alpha/2} \le \frac{\overline{Y} - \mu}{S/\sqrt{n}} \le t_{n-1,1-\alpha/2} \Big). \tag{4.32}$$

Here, $t_{n-1,p}$ is the p-quantile of a t-distributed random variable with $n-1$ degrees of freedom. The next steps are similar as in (4.28) to (4.30) and the result is summarized below.

Definition 4.5. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. The interval

$$\left[\overline{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right] \tag{4.33}$$

is an exact $(1-\alpha)$ confidence interval for the parameter μ. ◇

Example 4.10 (continuation of Example 4.1)**.** For the antibiotic imipenem, we do not know the underlying variability and hence we use (4.33). A sample 95% confidence interval for the mean inhibition diameter is $\big[\bar{y} - t_{99,.975}\, s/\sqrt{100}, \bar{y} + t_{99,.975}\, s/\sqrt{100}\,\big] = \big[32.40 - 1.98 \cdot 1.50/10, 32.40 + 1.98 \cdot 1.50/10\big] = \big[32.1, 32.7\big]$, where we used the information from R Code 4.1 and `qt(.975, 99)` (being 1.984).

Note we have deliberately rounded to a single digit here as the original data has been rounded to integer millimeters. ♣

Confidence intervals are, as shown in the previous two definitions, constituted by random variables (functions of $Y_1, \dots, Y_n$). Similar to estimators and estimates, sample confidence intervals are computed with the corresponding realization $y_1, \dots, y_n$ of the random sample. Subsequently, relevant confidence intervals will be summarized in the blue-highlighted text boxes, as shown here.

CI 1: Confidence interval for the mean μ

Under the assumption of a normal random sample,

$$\left[\overline{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right] \tag{4.34}$$

is an exact $(1-\alpha)$ confidence interval and

$$\left[\overline{Y} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right] \tag{4.35}$$

an approximate $(1-\alpha)$ confidence interval for μ.

Notice that both, the sample approximate and sample exact confidence intervals of the mean, are of the form

$$\text{estimate} \pm \text{quantile} \cdot \text{SE(estimate)}, \tag{4.36}$$

that is, symmetric intervals around the estimate. Here, $\text{SE}(\cdot)$ denotes the *standard error* of the estimate, that is, an estimate of the standard deviation of the estimator.

Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. The estimator S^2 for the parameter σ^2 is such, that $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$, i.e. a chi-square distribution with $n-1$ degrees of freedom (see Section 3.2.1).

Hence

$$1-\alpha = \mathrm{P}\Big(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\Big) \tag{4.37}$$

$$= \mathrm{P}\Big(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\Big), \tag{4.38}$$

where $\chi^2_{n-1,p}$ is the p-quantile of the chi-square distribution with $n-1$ degrees of freedom. The corresponding exact $(1-\alpha)$ confidence interval no longer has the form $\widehat{\theta} \pm q_{1-\alpha/2}\,\mathrm{SE}(\widehat{\theta})$, because the chi-square distribution is not symmetric.

For large n, the chi-square distribution can be approximated with a normal one (see also Section 3.2.1) with mean n and variance $2n$. Hence, a confidence interval based on a Gaussian approximation is reasonable, see CI 2.

CI 2: Confidence interval for the variance σ^2

Under the assumption of a normal random sample,

$$\Big[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\Big] \tag{4.39}$$

is an exact $(1-\alpha)$ confidence interval for σ^2.

For a random sample with $n > 50$

$$\Big[S^2 - z_{1-\alpha/2}\frac{\sqrt{2}S^2}{\sqrt{n}}, S^2 + z_{1-\alpha/2}\frac{\sqrt{2}S^2}{\sqrt{n}}\Big] \tag{4.40}$$

is an approximate $(1-\alpha)$ confidence interval for σ^2.

Example 4.11 (continuation of Example 4.1)**.** For the antibiotic imipenem, we have a sample variance of 2.24 leading to a 95%-confidence interval $[1.73, 3.03]$ for σ^2, computed with `(100-1)*var(imipDat)/qchisq(c(.975, .025), df=100-1)`. The confidence interval is slightly asymmetric: the center of the interval is $(1.73+3.03)/2 = 2.38$ compared to the estimate 2.24.

As the sample size is quite large, the Gaussian approximation yields a 95%-confidence interval $[1.62, 2.86]$ for σ^2, computed with `var(imipDat)+qnorm(c(.025, .975 ))*sqrt(2/100)* var(imipDat)`. The interval is symmetric, slightly narrower $2.86-1.62 = 1.24$ compared to $3.03-1.73 = 1.30$.

If we would like to construct a confidence interval for the standard deviation $\sigma = \sqrt{\sigma^2}$, we can use the approximation $[\sqrt{1.73}, \sqrt{3.03}\,] = [\,1.32, 1.74\,]$. That means, we have applied the same transformation for the bounds as for the estimate, a quite common approach. ♣

4.4.2 Interpretation of Confidence Intervals

The correct interpretation of sample confidence intervals is not straightforward and often causes confusion. An exact sample confidence interval $[\,b_l, b_u\,]$ for a parameter θ at level $1-\alpha$ means that when repeating the same experiment many times, on average, the fraction $1-\alpha$ of all confidence intervals contain the true parameter.

The sample confidence interval $[\, b_l, b_u \,]$ *does not* state that the parameter θ is in the interval with fraction $1-\alpha$. The parameter is not random and thus such a probability statement cannot be made.

Example 4.12. Let $Y_1, \dots, Y_4 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Figure 4.3 (based on R-Code 4.3) shows 100 sample confidence intervals based on Equation (4.31) (top), Equation (4.35) (middle) and Equation (4.33) (bottom). We color all intervals that do not contain zero, the true (unknown) parameter value $\mu = 0$, in red (if `ci[1]>mu | ci[2]<mu` is true).

On average we should observe 5% of the intervals colored red (five here) in the top and bottom panel because these are exact confidence intervals. Due to sampling variability we have for the specific simulation three and five in the two panels. In the middle panel there are typically more, as the normal quantiles are too small compared to the t-distribution ones (see Figure 3.2). Because n is small, the difference between the normal and the t-distribution is quite pronounced; here, there are eleven intervals that do not cover zero.

A few more points to note are as follows. As we do not estimate the variance, all intervals in the top panel have the same lengths. Further, the variance estimate shows a lot of variability (very different interval lengths in the middle and bottom panel). Instead of `for()`-loops, calls of the form `segments(1:ex.n, ybar + sigmaybar*qnorm(alpha/2)), 1:ex.n, ybar - sigmaybar*qnorm(alpha/2))` are possible. ♣

4.4.3 Comparing Confidence Intervals

Confidence intervals can often be constructed starting from an estimator $\widehat{\theta}$ and its distribution. In many cases it is possible to extract the parameter to get to $1-\alpha = \text{P}(B_l \leq \theta \leq B_u)$, often some approximations are necessary. For the variance parameter σ^2 in the framework of Gaussian random variables, CI 2 states two different intervals. These two are not the only ones and we can use further approximations (see derivation of Problem 4.7.**a** or Remark 3.1), all leading to slightly different confidence intervals.

Similar as with estimators, it is possible to compare different confidence intervals with equal level. Instead of bias and MSE, the criteria here are the width of a confidence interval and the so-called *coverage probability*. The latter is the probability that the confidence interval actually contains the true parameter. In case there is an exact confidence interval, the coverage probability is equal to the level of the interval.

For a specific estimator the width of a confidence interval can be reduced by reducing the level or increasing n. The former should be fixed at 90%, 95% or possibly 99% by convention. Increasing n after the experiment has been performed is often impossible. Therefore, the sample size should be choosen before the experiment, such that under the model assumptions the width of a confidence interval is below a certain threshold (see Chapter 12).

The following example illustrates the concept of coverage probability. A more relevant case will be presented in the next chapter.

R-Code 4.3 100 confidence intervals for the parameter μ, based on three different approaches (exact with known σ, approximate, and exact with unknown σ). (See Figure 4.3.)

```
set.seed( 1)     # important to reconstruct the same CIs
ex.n <- 100      # 100 confidence intervals
alpha <- .05     # 95\% confidence intervals
n <- 4           # sample size
mu <- 0          # mean
sigma <- 1       # standard deviation
sample <- matrix( rnorm( ex.n * n, mu, sigma), n, ex.n)  # sample used
yl <- mu + c( -6, 6)*sigma/sqrt(n)      # same y-axis for all 3 panels
ybar <- apply( sample, 2, mean)         # mean for each sample
# First panel: sigma known:
sigmaybar <- sigma/sqrt(n)
plot( 1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n', ylab='',
     main=bquote(sigma~known))
abline( h=mu)
for ( i in 1:ex.n){     # draw the individual CIs with appropriate color
  ci <- ybar[i] + sigmaybar * qnorm(c(alpha/2, 1-alpha/2))
  lines( c(i,i), ci, col=ifelse( ci[1]>mu|ci[2]<mu, 2, 1))
}
# Second panel: sigma unknown, normal approx:
sybar <- apply(sample, 2, sd)/sqrt(n)   # estimate the standard deviation
plot( 1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n', ylab='',
     main=bquote("Gaussian approximation"))
abline( h=mu)
for ( i in 1:ex.n){   # similar but with individual standard deviation
  ci <- ybar[i] + sybar[i] * qnorm(c(alpha/2, 1-alpha/2))
  lines( c(i,i), ci, col=ifelse( ci[1]>mu | ci[2]<mu, 2, 1))
}
# Third panel: sigma unknown, t-based:
plot(1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n', ylab='',
     main='t-distribution')
abline( h=mu)
for ( i in 1:ex.n){  # similar but with t-quantile
  ci <- ybar[i] + sybar[i] * qt(c(alpha/2, 1-alpha/2), n-1)
  lines( c(i,i), ci, col=ifelse( ci[1]>mu | ci[2]<mu, 2, 1))
}
```

Example 4.13. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. If σ^2 is known, the following confidence intervals

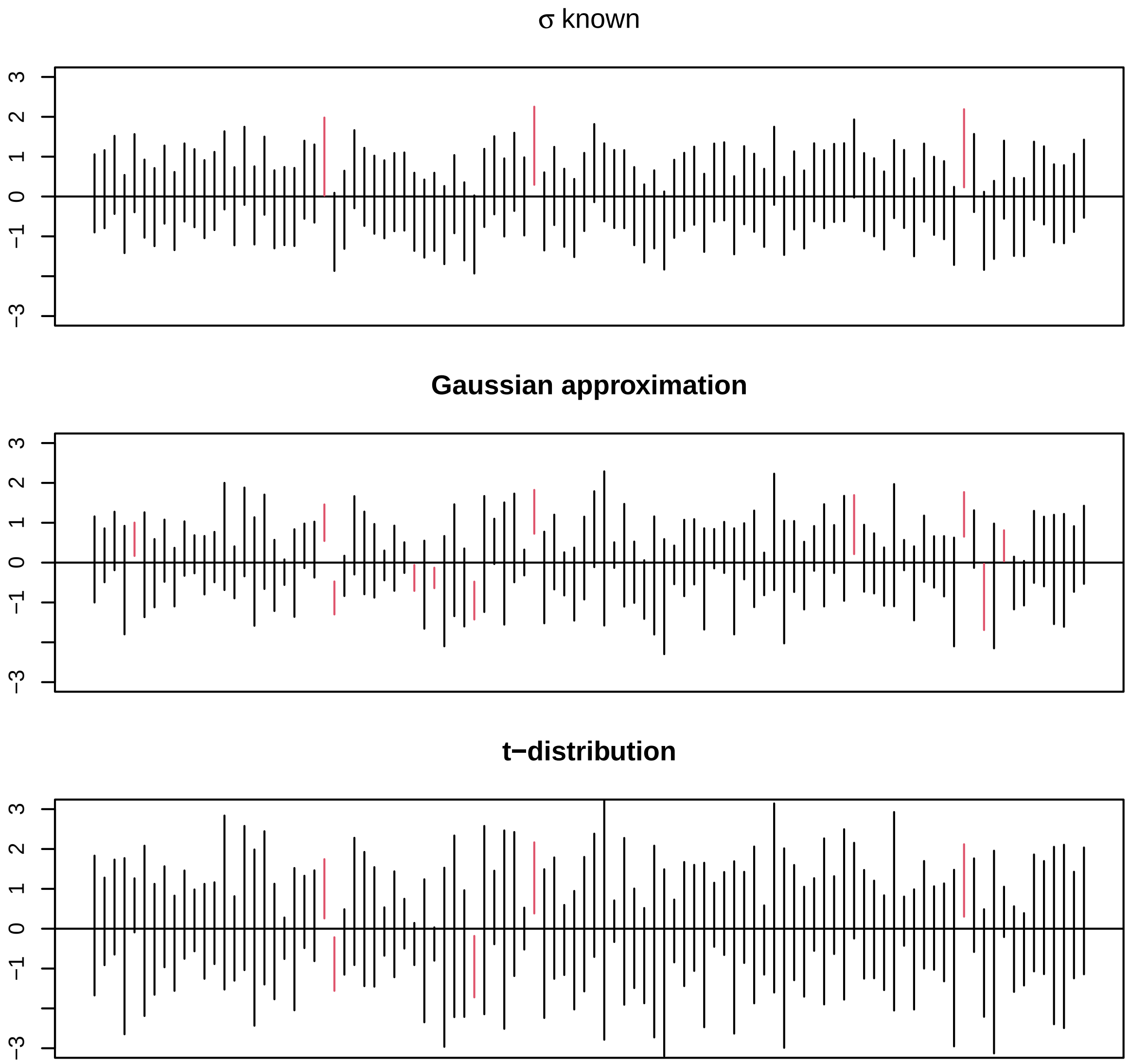


Figure 4.3: Normal and t-distribution based confidence intervals for the parameters $\mu = 0$ with $\sigma = 1$ (above) and unknown σ (middle and below). The sample size is $n = 4$ and confidence level is $(1 - \alpha) = 95\%$. Confidence intervals which do not cover the true value zero are shown in red. (See R-Code 4.3.)

have coverage probability exactly equal to the level $1 - \alpha$:

$$\Big[\overline{Y} + t_{n-1,3\alpha/4}\frac{S}{\sqrt{n}}, \overline{Y} + t_{n-1,1-\alpha/4}\frac{S}{\sqrt{n}}\Big] \qquad \Big[\overline{Y} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}, \infty\Big) \tag{4.41}$$

with (4.34) having the shortest possible width (for fixed level).

To calculate the coverage probability of (4.35), we would have to calculate probabilities of the form $\mathrm{P}\big(\overline{Y} + z_{1-\alpha/2}S/\sqrt{n} \geq \mu\big)$. This is not trival and we approach it differently. Suppose that $z_{1-\alpha/2} = t_{n-1,1-\alpha^\star/2}$, i.e., the standard normal p-quantile and the $1 - \alpha^\star/2$-quantile of the t-distribution with $n - 1$ degrees of freedom are equivalent. Thus the interval has coverage probability $\alpha^\star$ and in the case of Example 4.12 $\alpha^\star \approx 86\%$, based on `1-2*uniroot(function(p) qt(p, 3)-qnorm(.025),c(0,1))$root`. In Figure 4.3, we have 11 intervals marked compared to expected 14. ♣

4.5 Bibliographic Remarks

Statistical estimation theory is very classical and many books are available. For example, Held and Sabanés Bové (2014) (or it's German predecessor, Held, 2008) are written at an accessible level.

4.6 Exercises and Problems

Problem 4.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Derive the approximate confidence interval (4.40) for σ^2.

b) We assume $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Let $\widehat{\theta}_\rho = \dfrac{1}{\rho}\displaystyle\sum_{i=1}^{n}(Y_i - \overline{Y})^2$. Calculate $\text{bias}\big(\widehat{\theta}_\rho\big)$ and $\text{Var}\big(\widehat{\theta}_\rho\big)$, both as a function of ρ. Which value of ρ minimizes the mean squared error?

c) We assume $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Let $\widehat{\theta}_\rho = \dfrac{1}{\rho}\displaystyle\sum_{i=1}^{n} Y_i$ and calculate the MSE as a function of ρ. Argue that $\overline{Y}$ is the minimum variance unbiased estimator.

Problem 4.2 (Hemoglobin levels) The following hemoglobin levels of blood samples from patients with Hb SS and Hb S/β sickle cell disease are given (Hüsler and Zimmermann, 2010):

```
HbSS <- c( 7.2, 7.7, 8, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7, 9.1,
          9.1, 9.1, 9.8, 10.1, 10.3)
HbSb <- c( 8.1, 9.2, 10, 10.4, 10.6, 10.9, 11.1, 11.9, 12.0, 12.1)
```

a) Visualize the data with boxplots.

b) Propose a statistical model for Hb SS and for Hb S/β sickle cell diseases. What are the parameters? Indicate your random variables and parameters with subscripts $_{\text{SS}}$ and $_{\text{Sb}}$.

c) Estimate all parameters from your model proposed in part **b)**.

d) Propose a single statistical model for both diseases. What are the parameters? Estimate all parameters from your model based on intuitive estimators.

e) Based on boxplots and QQ-plots, is there coherence between your model and the data?

Problem 4.3 (Normal distribution with known σ^2) Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\sigma > 0$ assumed to be known.

a) What is the distribution of $\sqrt{n}(\overline{X}-\mu)/\sigma$? (No formal proofs required).

b) Let $n=9$. Calculate $\mathrm{P}(-1 \leq (\overline{X}-\mu)/\sigma \leq 1)$.

c) Determine the lower and upper bound of a confidence interval B_l and B_u (both functions of $\bar{X}$) such that

$$\mathrm{P}(-q \leq \sqrt{n}(\overline{X}-\mu)/\sigma \leq q) = \mathrm{P}(B_l \leq \mu \leq B_u)$$

d) Construct a sample 95%-confidence interval for μ.

e) Determine an expression of the width of the confidence interval? What "elements" appearing in a general $1-\alpha$-confidence interval for μ make the interval narrower?

f) Use the sickle-cell disease data from Problem 2 and construct 90%-confidence intervals for the means of HbSS and HbSβ variants (assume $\sigma = 1$).

g) Repeat problems **a**)–**f**) by replacing σ with S and making adequate and necessary changes.

Problem 4.4 (ASTs radii) We work with the inhibition diameters for imipenem and meropenem as given in Table 4.1.

a) The histograms of the inhibition diameters in Figure 4.1 are not quite symmetric. Someone proposes to work with square-root diameters.

Visualize the transformed data of Table 4.1 with histograms. Does such a transformation render the data more symmetric? Does such a transformation make sense? Are there other reasonable transformations?

b) What are estimates of the inhibition area for both antibiotics (in cm^2)?

c) Construct a 90% confidence interval for the inhibition area for both antibiotics?

d) What is an estimate of the variability of the inhibition area for both antibiotics? What is the uncertainty of this estimate.

Problem 4.5 (Geometric distribution) In the setting of Example 2.3, denote $p = \mathrm{P}(\text{shot is successful})$. Assume that it took the boy $k_1, k_2, \dots, k_n$ attempts for the 1st, 2nd, …, nth successful shot.

a) Derive the method of moment estimator of p. Argue that this estimator is also the least squares estimator.

b) Derive the maximum likelihood estimate of p.

c) The distribution of the estimator $\widehat{p} = n/\sum_{i=1}^{n} k_i$ is non-trivial. And thus we use a simulation approach to assess the uncertainty in the estimate. For $n=10$ and $p=0.1$ draw from the geometric distribution (`rgeom(...)+1)`) and report the estimate. Repeat $R = 500$ times and discuss the histogram of the estimates. What changes if $p=0.5$ or $p=0.9$?

d) Do you expect the estimator in **c**) to be biased? Can you support your claim with a simulation?

Problem 4.6 (Poisson Distribution) Consider $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{Pois}(\lambda)$ with a fixed $\lambda > 0$.

a) Let $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be an estimator of λ. Calculate $\text{E}(\widehat{\lambda})$, $\text{Var}(\widehat{\lambda})$ and the $\text{MSE}(\widehat{\lambda})$.

b) What is $\text{E}(\widehat{\lambda})$, $\text{Var}(\widehat{\lambda})$ and $\text{MSE}(\widehat{\lambda})$ when $n \to \infty$?

Problem 4.7 (Coverage probability) Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and consider the usual estimator S^2 for σ^2. Show that the coverage probability of (4.40) is given by

$$1 - \text{P}\Big(W \geq \frac{n-1}{1 - z_{1-\alpha/2}\frac{\sqrt{2}}{\sqrt{n}}}\Big) - \text{P}\Big(W \leq \frac{n-1}{1 + z_{1-\alpha/2}\frac{\sqrt{2}}{\sqrt{n}}}\Big)$$

with $W \sim \mathcal{X}^2_{n-1}$. Plot the coverage probability as a function of n.

Problem 4.8 (Germany cancer counts) The dataset `Oral` is available in the R package `spam` and contains oral cavity cancer counts for 544 districts in Germany.

a) Load the data and take a look at its help page using `?Oral`.

b) Compute summary statistics for all variables in the dataset.
Which of the 544 regions has the highest number of expected counts `E`?

c) Poisson distribution is common for modeling rare events such as death caused by cavity cancer (column `Y` in the data). However, the districts differ greatly in their populations. Define a subset from the data, which only considers districts with *expected* fatal casualties caused by cavity cancer between 35 and 45 (`subset`, column `E`). Perform a Q-Q Plot for a Poisson distribution.

Hint: use `qqplot()` from the `stats` package and define the theoretical quantiles with `qpois(ppoints( ...), lambda=...)`.

Simulate a Poisson distributed random variable with the same length and and the same lambda as your subset. Perform a QQ-plot of your simulated data. What can you say about the distribution of your subset of the cancer data?

d) Assume that the standardized mortality ratio $Z_i = Y_i/E_i$ is normally distributed, i.e., $Z_1, \dots, Z_{544} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimate μ and give a 95% (exact) confidence interval (CI). What is the precise meaning of the CI?

e) Simulate a 95% confidence interval based on the following bootstrap scheme (sampling with replacement):
Repeat $10'000$ times

- Sample 544 observations Z_i with replacement
- Calculate and store the mean of these sampled observations

Construct the confidence interval by taking the 2.5% and the 97.5% quantiles of the stored means.
Compare it to the CI from **d**).

Problem 4.9 (BMJ Endgame) Discuss and justify the statements about 'Describing the spread of data' given in doi.org/10.1136/bmj.c1116.

Chapter 5

Statistical Testing

Learning goals for this chapter:

- ⋄ Explain the concepts of hypothesis and significance test
- ⋄ Given a problem situation, state appropriate null and alternative hypotheses, perform a hypothesis test, interpret the results
- ⋄ Define p-value and the significance level
- ⋄ Know the difference between one-sample and two-sample t-tests
- ⋄ Explain and apply various classical tests
- ⋄ Understand the duality of tests and confidence intervals
- ⋄ Be aware of the multiple testing problem and know how to deal with it

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter05.R.

Recently I have replaced a LED light bulb that was claimed to last 20 000 hours. However, in less than 2 months (and only a fraction thereof in use) the bulb was already broken and I immediately asked myself if I am unlucky or is the claim simply exaggerated? A few moments later rational kicked back in and - being a statistician - I knew that one individual break should not be used to make too general statements

On a similar spirit, suppose we observe 13 heads in 17 tosses of a coin. When tossing a fair coin, I expect somewhere between seven to ten heads and the 13 observed ones representing seemingly an unusual case. We intuitively wonder if the coin is fair.

In this chapter we discuss a formal approach to answer if the observed data provides enough evidence against a hypothesis (against a claimed livetime or against a claimed fairness). We introduce two interlinked types of statistical testing procedures and provide a series of tests that can be used off-the-shelf.

5.1 The General Concept of Significance Testing

The idea of a statistical testing procedure is to formulate a statistical hypothesis and to draw conclusions from them based on the data. A testing procedure states at the beginning a null hypothesis, denoted with H_0, and we compare how compatible the data is with respect to this hypothesis.

Simply stated, starting from a statistical null hypothesis a statistical test calculates a value from the data and places that value in the context of the hypothetical distribution induced by the statistical null hypothesis. If the value from the data is unlikely to occur with respect to the hypothetical distribution, we argue that the data provides evidence against the null hypothesis. This coherence is typically quantified as a probability, i.e., the famous p-value, with formal definition as follows.

Definition 5.1. The p-value is the probability under the distribution of the null hypothesis of obtaining a result equal to or more extreme than the observed result. ♢

Example 5.1. We assume a fair coin is tossed 17 times and we observe 13 heads. Under the null hypothesis of a fair coin, each toss is a Bernoulli random variable and the 17 tosses can be modeled with a binomial random variable $\mathcal{B}in(n = 17, p = 1/2)$. Hence, the p-value is the probability of observing $0, 1, \dots, 4,\ 13, 14, \dots, 17$ heads (or by symmetry of observing $17, \dots, 13,\ 4, \dots, 0$), which can be calculated with `sum( dbinom(0:4, size=17, prob=1/2) + dbinom(13:17, size=17, prob=1/2))` and is 0.049. The p-value indicates that we observe such a seemingly unlikely event roughly every 20th time.

Note that because of the symmetry of the binomial distribution at $\mathcal{B}in(n, 1/2)$, we can alternatively calculate the p-value as `2*pbinom(4, size=17, prob=1/2)` or equivalently as `2*pbinom(12, size=17, prob=1/2, lower.tail=FALSE)`.

In this example, we have considered more extreme as very many or very few heads. There might be situations, where very few heads is not relevant or does not even make sense and thus "more extreme" corresponds only to observing $13, 14, \dots, 17$ heads. ♣

Figure 5.1 illustrates graphically the p-value in two hypothetical situations. Suppose that under the null hypothesis the hypothetical distribution of the observed result is Gaussian with mean zero and variance one and suppose that we observe a value of 1.8. If more extreme is considered on both sides of the tails of the density then the p-value consists of two probabilities (here because of the symmetry, twice the probability of either side). If more extreme is actually larger (possibly smaller in other situations), the p-value is calculated based on a one-sided probability. As the Gaussian distribution is symmetric around its mean, the two-sided p-value is twice the one-sided p-value, here `1-pnorm(1.8)`, or, equivalently, `pnorm(1.8, lower.tail=FALSE)`.

We illustrate the statistical testing approaches and the statistical tests with data introduced in the following example.

Example 5.2. In rabbits, pododermatitis is a chronic multifactorial skin disease that manifests mainly on the hind legs. This presumably progressive disease can cause pain leading to poor welfare. To study the progression of this disease on the level of individual animals, scientists

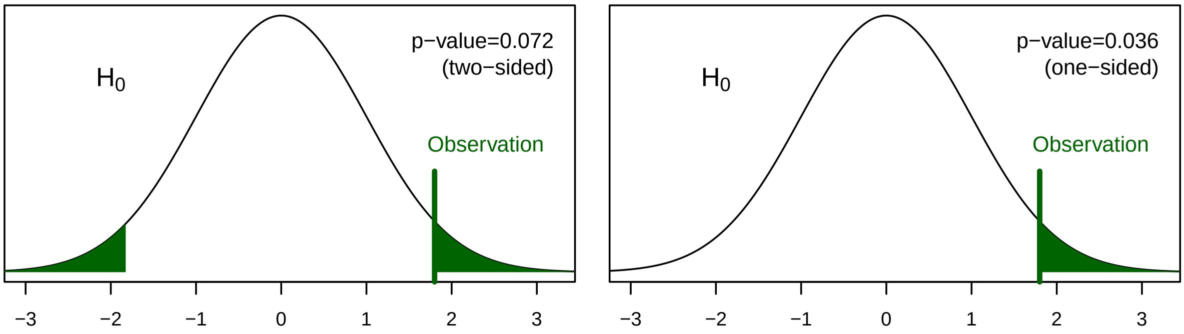


Figure 5.1: Illustration of the p-value in the case of a standard normal distribution with observed value 1.8. Two-sided (left panel) and one-sided setting (right panel).

assessed many rabbits in three farms over the period of an entire year (Ruchti *et al.*, 2019). We use a subset of the dataset in this and later chapters, consisting of one farm (with two barns) and four visits (between July 19/20, 2016 and June 29/30, 2017). The 6 stages from Drescher and Schlender-Böbbis (1996) were used as a tagged visual-analogue-scale to score the occurrence and severity of pododermatitis on 4 spots on the rabbits hind legs (left and right, heal and middle position), resulting in the variable `PDHmean` with range 0–10, for details on the scoring see Ruchti *et al.* (2018). We consider the visits in June 2017. R-Code 5.1 loads the dataset and subsets it correspondingly. ♣

In practice, we often start with a scientific hypothesis and subsequently collect data or perform an experiment to confirm the hypothesis. The data is then "modeled statistically", in the sense that we need to determine a theoretical distribution for which the data is a realization. In our discussion here, the distribution typically involves parameters that are linked to the scientific question (probability p in a binomial distribution for coin tosses, mean μ of a Gaussian distribution for testing differences pododermatitis scores). We then formulate the null hypothesis H_0 for which the data should provide evidence against. The calculation of a p-value can be summarized as follows. When testing about a certain parameter, say θ, we use an estimator $\widehat{\theta}$ for that parameter. We often need to transform the estimator such that the distribution thereof does not depend on (the) parameter(s). We call this random variable *test statistic* which is typically a function of the random sample. The test statistic evaluated at the observed data is then used to calculate the p-value based on the distribution of the test statistic. Based on the p-value we summarize the evidence *against* the null hypothesis. We cannot make any statement *for* the hypothesis.

Example 5.3 (continuation of Example 5.2)**.** For the visits in June 2017 we would like to asses if the score of the rabbits is comparable to $10/3 \approx 3.33$, representing a low-grade scoring (low-grade hyperkeratosis, hypotrichosis or alopecia). We have 17 observations and the sample mean is 3.87 with a standard deviation of 0.64. Is this enough evidence in the data to claim that the observed mean is different from low-grade scoring?

We postulate a Gaussian model for the scores. The observations are a realization of $X_1, \dots,$ $X_{17} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 0.8^2)$, i.e., $n = 17$ and the standard deviation is known (the latter will be relaxed

R-Code 5.1 Pododermatitis in rabbits, dataset `pododermatitis`.

```
str( podo <- read.csv('data/podo.csv'))

## 'data.frame': 67 obs. of  6 variables:
##  $ ID     : int  4 3 9 10 1 7 8 5 14 13 ...
##  $ Age    : int  12 12 14 12 17 14 14 12 6 6 ...
##  $ Weight : num  4.46 4.31 4.76 5.34 5.71 5.39 5.42 5.13 5.39 5.41 ...
##  $ Visit  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Barn   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PDHmean: num  4.47 3.62 4.28 4.2 1.77 ...

apply( podo, 2, function(x) length(unique(x)) )  # 4 visits, 2 barns

##      ID     Age  Weight   Visit    Barn PDHmean
##      17      21      63       4       2      49

PDHmean <- podo$PDHmean[podo$Visit==13]
length( PDHmean)

## [1] 17

print( me <- mean( PDHmean))

## [1] 3.8691

print( se <- sd( PDHmean))

## [1] 0.63753
```

soon, here we suppose that this value has been determined by another study or additional information). The scientific hypothesis results in the null hypothesis H_0 : "mean is low-grade scoring" being equivalent to $H_0 : \mu = 3.33$. Under the null hypothesis, we have $\overline{X} \overset{H_0:\mu=3.33}{\sim} \mathcal{N}(3.33, 0.8^2/17)$. Hence, we test about the parameter μ and our estimator is $\widehat{\theta} = \overline{X}$ with a known distribution under the null. With this information, the p-value in a two-sided setting is

$$\begin{aligned} p\text{-value} &= \text{P(under the null hypothesis we observe 3.87 or a more extreme value)} && (5.1)\\ &= 2\,\text{P}_{H_0}(|\overline{X}| \geq |\bar{x}|) = 2\big(1 - \text{P}_{H_0}(\overline{X} < 3.87)\big) && (5.2)\\ &= 2\Big(1 - \text{P}\Big(\frac{\overline{X} - 3.33}{0.8/\sqrt{17}} < \frac{3.87 - 3.33}{0.8/\sqrt{17}}\Big)\Big) = 2(1 - \varphi(2.76)) \approx 0.6\%. && (5.3)\end{aligned}$$

where we have used the subscript "H_0" to emphasis the calculation under the null hypothesis, i.e., $\mu = 3.33$. An alternative is to write it in a conditional form $\text{P}_{H_0}(\,\cdot\,) = \text{P}(\,\cdot \mid H_0)$. Hence, there is evidence in the data against the null hypothesis. ♣

In many cases we use a "known" statistical test, instead of a "manually" constructed test statistic. Specifically, we state the statistical model with a well-known and named test, as we shall see later.

Some authors summarize p-values in $[1, 0.1]$ as no evidence, in $[0.1, 0.01]$ as weak evidence, in $[0.01, 0.001]$ as substantial evidence, and smaller ones as strong evidence (see, e.g., Held and Sabanés Bové, 2014). For certain representations, R output uses symbols for similar ranges `␣`, `.` and `*`, `**`, and, `***`.

The workflow of a statistical significance test can be summarized as follows. The starting point is a scientific question or hypothesis and data that has been collected to support the scientific claim.

(i) Formulation of the statistical model and statistical assumptions. Formulate the scientific hypothesis in terms of a statistical one.

(ii) Selection of the appropriate test or test statistic and formulation of the null hypothesis H_0 with the parameters of the test.

(iii) Calculation of the p-value,

(iv) Interpretation of the results of the statistical test and conclusion.

Although the workflow is presented in a linear fashion, there are several dependencies. For example the interpretation depends not only on the p-value but also on the null hypothesis in terms of proper statistical formulation and, finally, on the scientific question to be answered, see Figure 5.2. The statistical test hypothesis essentially depends on the statistical assumptions, but need to be cast to answer the scientific questions, of course. The statistical assumptions may also determine the selection of statistical tests. This dependency will be taken up in Chapter 7.

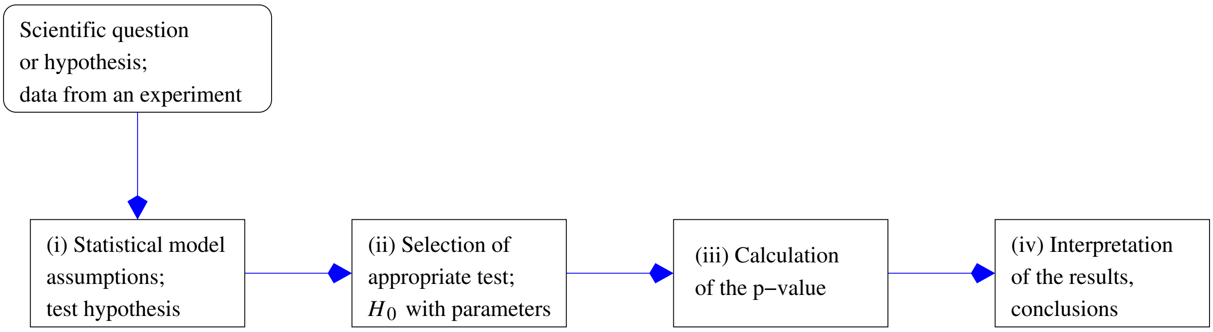


Figure 5.2: Workflow of statistical testing.

Example 5.4 (Revisiting Example 5.1 using the workflow of Figure 5.2)**.** The (scientific) claim is that the coin is biased and as data we have 17 coin tosses with 13 heads. (i) As the data is the number of successes among a fixed number of trials a binomal model is appropriate. The test hypothesis that we want to reject is equal chance of getting head or tail. (ii) We can work directly with $X \sim \mathcal{B}in(17, p)$ with the null hypothesis $H_0 : p = 1/2$. (iii) The p-value is 0.049. (iv) The p-value indicates that we observe such a seemingly unlikely event roughly every 20th time only weak evidence. ♣

Example 5.5 (Revisiting Example 5.2 using the workflow of Figure 5.2)**.** The scientific claim is that the observed score is different to low-grade. We have four scores from the hind legs from 17 animals. (i) We work with one average value per animal (`PDMean`). Thus we assume that $\overline{X} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 0.8^2)$. We want to quantify how much the observed mean differs from low-grade score. (ii) We have $\overline{X} \sim \mathcal{N}(\mu, 0.8^2/17)$, with the null hypothesis $H_0 : \mu = 3.33$. (iii) The p-value is $p\text{-value} = 2\,\mathrm{P}_{H_0}(|\overline{X}| \geq |\bar{x}|) \approx 0.6\%$. (iv) There is substantial evidence, that the pododermatitis scores are different than low-grade scoring. ♣

5.2 Hypothesis Testing

A criticism of significance testing is that we have only *one* hypotheses. It is often easier to choose between two alternatives and this is essentially the approach of hypothesis testing. However, as in the setting of significance testing, we will still not choose *for* one hypothesis but rather *reject* or *fail to reject* the null hypothesis.

Hypothesis testing starts with a null hypothesis H_0 and an alternative hypothesis, denoted by H_1 or H_A. These hypotheses are with respect to a parameter, say θ or some other choice specific to the situation.

Example 5.6. We revisit the light bulb situation elaborated at the beginning of the chapter. I postulate a null hypothesis H_0 : "median lifetime is 20 000 h" versus the alternative hypothesis H_1 : "median lifetime is 5 000 h". I only have one observation and thus I need external information about the distribution of light bulb lifetime. Although there is no consensus, some published literature claim that the cdf of certain types of light bulbs are given by $F(x) = 1 - \exp(-x/\lambda)^k$ for $x > 0$ and k between 4 and 5. For simplicity we take $k = 4$ and thus the median lifetime is $\lambda \log(2)^{1/4}$. The hypotheses are thus equivalent to $H_0 : \lambda = 20\,000/\log(2)^{1/4}$ h versus $H_1 : \lambda = 5\,000/\log(2)^{1/4}$ h. ♣

In the example above, I could have taken any other value for the alternative. Of course this is dangerous and very subjective (the null hypothesis is given by companies claim). Therefore, we state the alternative hypothesis as everything but the null hypothesis. In the example above it would be $H_1 : \lambda \neq 20\,000/\log(2)^{1/4}$ h.

In a similar fashion, we could state that the median lifetime is at least 20 000 h. In such a setting we would have a null hypothesis H_0 : "median lifetime is 20 000 h or larger" versus the alternative hypothesis H_1 : "median lifetime smaller than 20 000 h", which is equivalent to $H_0 : \lambda \geq 20\,000/\log(2)^{1/4}$ h versus $H_1 : \lambda <\ 20\,000/\log(2)^{1/4}$ h.

Hypotheses are classified as *simple* if parameter θ assumes only a single value (e.g., H_0: $\theta = 0$), or *composite* if parameter θ can take on a range of values (e.g., H_0: $\theta \leq 0$, H_1: $\mu \neq \mu_0$).

The case of a simple null hypothesis and composite alternative hypothesis is also called a *two-sided* setting. Whereas composite null hypothesis and composite alternative hypothesis is called *one-sided* or directional setting.

Note that for Example 5.2 a one-sided test is necessary for the hypothesis "there is a progression of the pododermatitis scores between two visits", but a two-sided test is needed for "the pododermatitis scores between two visits are different". We strongly recommend to always use two-sided tests (e.g. Bland and Bland, 1994; Moyé and Tita, 2002), not only in clinical studies where it is the norm but as Bland and Bland (1994) states "a one-sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all. Expectation of a difference in a particular direction is not adequate justification.". However to illustrate certain concepts, a one-sided setting may be simpler and more accessible.

In the case of a hypothesis test, we compare the value of the test statistic with the quantiles of the distribution of the null hypothesis. A predefined threshold determines if we *reject* H_0, if not, we *fail to reject* H_0.

Definition 5.2. The *significance level* α is a threshold determined before the testing, with $0 < \alpha < 1$ but it is often set to 5% or 1%.

The *rejection region* of a test includes all values of the test statistic for which we reject the null hypothesis. The boundary values of the rejection region are called *critical values*. ♢

Similar as for the significance test we reject for values of the test statistic that are in the tail of the density under the null hypothesis, e.g., that would lead to a small p-value. In fact, we can base our decision on whether the p-value is smaller than the significance level or not.

It is important to realize that the level α is set by the scientists, not by the experiment or the data. Therefore there is some "arbitrariness" to the value and thus whether we reject the H_0 or not. The level α may be imposed to other values in different scientific domains.

For one-sided hypotheses like $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ (similarly $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$), the "=" case is used in the null hypothesis, i.e., the most unfavorable case of the alternative point of view. That means, the null hypothesis is from a calculation perspective always simple and could be reduced to a simple hypothesis.

Figure 5.3 shows the rejection region of two-sided test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ (left panel), $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ and one-sided (right panel), for the specific case $\mu = 0$. For the left-sided case, $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$, the rejection area is on the left side, analogue to the right-sided case.

If we assume that the distribution of the test statistic is Gaussian and $\alpha = 0.05$, the critical values are ±1.96 and 1.64, respectively (`qnorm(c(0.05/2, 1 - 0.05/2)` and `qnorm(1 - 0.05)`). These critical values are typically linked with the so-called z-test.

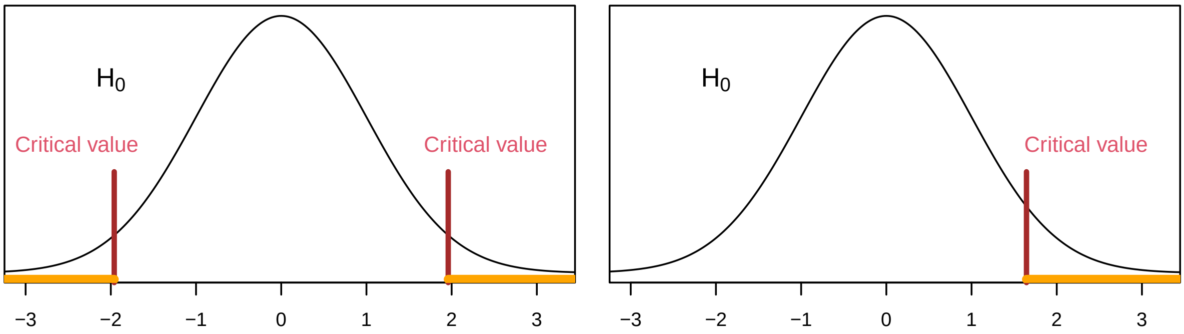


Figure 5.3: Critical values (red) and rejection regions (orange) for two-sided $H_0 : \mu = \mu_0 = 0$ (left) and one-sided $H_0 : \mu \leq \mu_0 = 0$ (right) hypothesis test with significance level $\alpha = 5\%$.

In significance testing two types of errors can occur. Type I errors: we reject H_0 if we should have not and Type II errors: we fail to reject H_0 if we should have. The framework of hypothesis testing allows us to express the probabilities of committing these two errors. The probability of committing a Type I error is exactly $\alpha = \text{P}(\text{reject } H_0 | H_0)$. This probability is often called the *size* of the test. To calculate the probability of committing a Type II error, we need to assume a specific value for our parameter within the alternative hypothesis, e.g., a simple alternative.

The probability of Type II error is often denoted with $\beta = \text{P}(\text{not rejecting } H_0 | H_1)$. Table 5.1 summarizes the errors in a classical 2×2 layout.

Table 5.1: Probabilities of Type I and Type II errors in the setting of a significance test.

		True but unknown state	
		H_0 true	H_1 true
Test result	do not reject H_0	$1-\alpha$	$\boldsymbol{\beta}$
	reject H_0	$\boldsymbol{\alpha}$	$1-\beta$

Example 5.7 (revisit Example 5.1)**.** The null hypothesis remains as *having a fair coin* and the alternative is simply *not having a fair coin*. Suppose we reject the null hypothesis if we observe $0, \dots, 4$ or $13, \dots, 17$ heads out of 17 tosses. The Type I error is `2*pbinom(4, size=17, prob=1/2)`, i.e., 0.049, and, if the coin has a probability of 0.7 for heads, the Type II error is `sum(dbinom(5:12, size=17, prob=0.7))`, i.e., 0.611. However, if the coin has a probability of 0.6 for heads, the Type II error increases to `sum(dbinom(5:12, size=17, prob=0.6))`, i.e., 0.871. ♣

Ideally we would like to use tests that have simultaneous small Type I and Type II errors. This is coneptually not possible as reducing one increases the other and one typically fixes the Type I error to some small value, say 5%, 1% or suchlike (committing a type one error has typically more severe consequences than a Type II error). Type I and Type II errors are shown in Figure 5.4 for two different alternative hypotheses. When reducing the significance level α, the critical values move further from the center of the density under H_0 and thus to an increase of the Type II error β. Additionally, the clearer the separation of the densities under H_0 and H_1, the smaller the Type II error β. This is intuitive, if the data stems from H_1 which is "far" from H_0, the chance that we reject is large.

As a summary, the Type I error

- is defined a priori by selection of the significance level (often 5%, 1%),
- is not influenced by sample size,
- is increased with multiple testing of the same data (we discuss this in Section 5.5.2)

and the Type II error

- depends on sample size and significance level α,
- is a function of the alternative hypothesis.

The value $1-\beta$ is called the *power of a test*. High power of a test is desirable in an experiment: we want to detect small effects with a large probability. R-Code 5.2 computes the power for a z-test (Gaussian random sample with known variance). More specifically, under the assumption of $\sigma/\sqrt{n} = 1$ we test H_0: $\mu_0 = 0$ versus H_1: $\mu_0 \neq 0$. Similarly to the probability of a Type II

R-Code 5.2 A one-sided and two-sided power curve for a z-test. (See Figure 5.5.)

```
alpha <- 0.05                            # significance level
mu0 <- 0                                 # mean under H_0
mu1 <- seq(-1.5, to=5, by=0.1)           # mean under H_1
power_onesided <- 1-pnorm( qnorm(1-alpha, mean=mu0), mean=mu1)
power_twosided <- pnorm( qnorm(alpha/2, mean=mu0), mean=mu1) +
    pnorm( qnorm(1-alpha/2, mean=mu0), mean=mu1, lower.tail=FALSE)
plot( mu1, power_onesided, type='l', ylim=c(0,1), xlim=c(-1, 4.25), las=1,
     xlab=bquote(mu[1]-mu[0]),  ylab="Power", col=4, yaxs='i', lwd=1)
axis(2, at=alpha, labels='')             # adding level
axis(2, at=1.4*alpha, labels=bquote(alpha), las=1, adj=0, tick=FALSE)
lines( mu1, power_twosided, lty=2)  # power curve for two-sided test
abline( h=alpha, col='gray')             # significance level
abline( v=c(2, 4), lwd=2, col=3)     # values from figure 4.3
```

error, the power can only be calculated for a specific assumption of the “actual” mean μ_1, i.e., of a simple alternative. Thus, as typically done, Figure 5.5 plots $power(\mu_1 - \mu_0)$.

For $\mu_1 = \mu_0$, the power is equivalent to the size of the test (significance level α). If $\mu_1 - \mu_0$

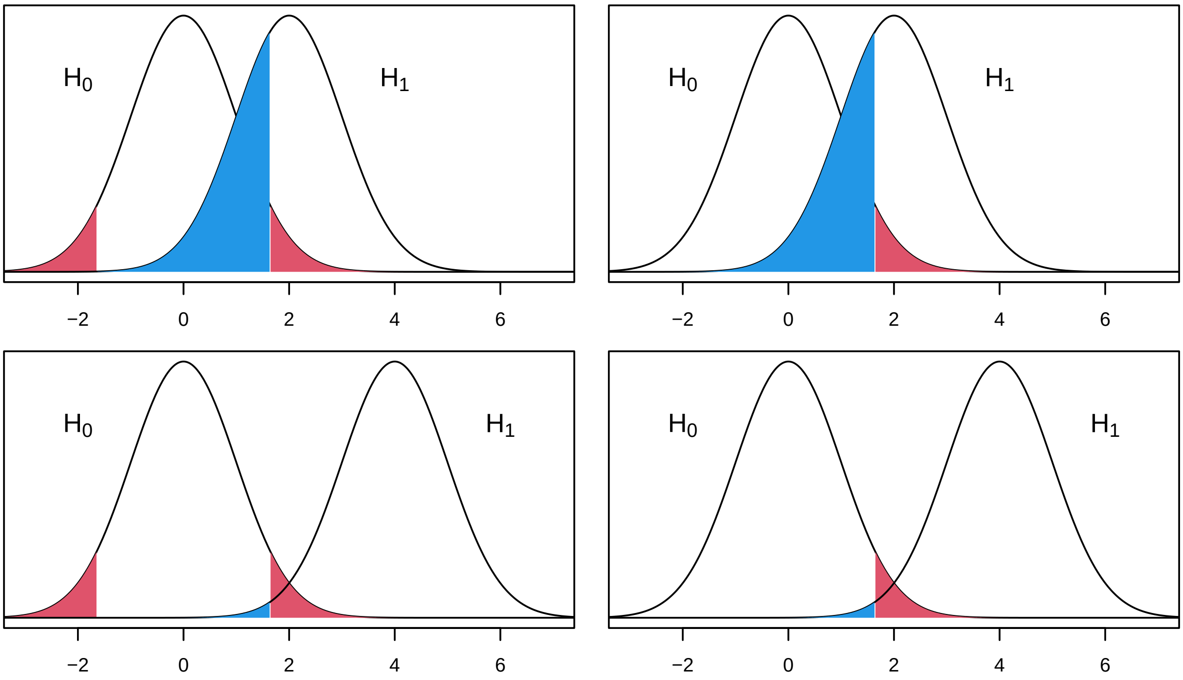


Figure 5.4: Type I error with significance level α (red) and Type II error with probability β (blue) for two different alternative hypotheses ($\mu = 2$ top row, $\mu = 4$ bottom row) with two-single hypothesis $H_0 : \mu = \mu_0 = 0$ (left column) and one-sided hypothesis $H_0 : \mu \leq \mu_0 = 0$ (right).

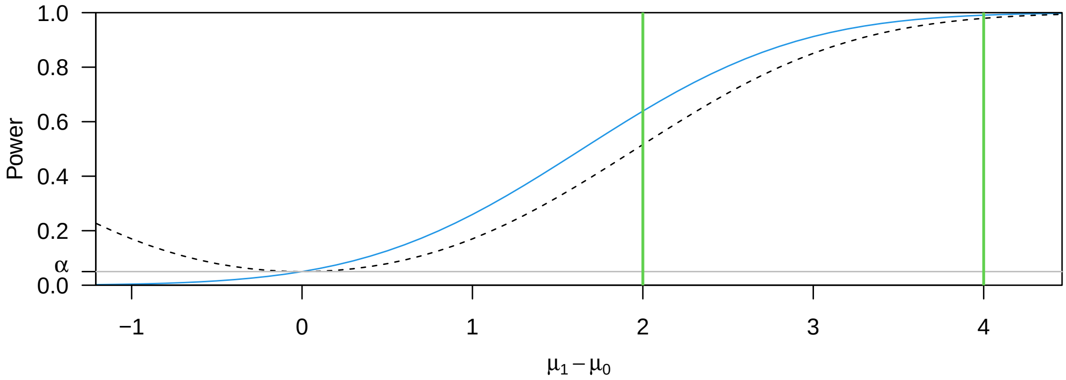


Figure 5.5: Power curves for a z-test: one-sided (blue solid line) and two-sided (black dashed line). The gray line represents the level of the test, here $\alpha = 5\%$. The vertical lines represent the alternative hypotheses $\mu = 2$ and $\mu = 4$ of Figure 5.4. (See R-Code 5.2.)

increases, the power increases in sigmoid-shaped form to reach asymptotically one. For a two-sided test, the power is symmetric around $\mu_1 = \mu_0$ (no direction preferred) and smaller than the power for a one sided test. The latter decreases further to zero for negative differences $\mu_1 - \mu_0$ although for the negative values the power curve does not make sense. In a similar fashion as to reduce the probability of a Type II error, it is possible to increase the power by increasing the sample. Note that in the illustrations above we work with $\sigma/\sqrt{n} = 1$, the dependence of the power on n is somewhat hidden by using the default argument for `sd` in the functions `pnorm()` and `qnorm()`.

Remark 5.1. It is impossible to reduce simultaneously both the Type I and II error probabilities, but it is plausible to consider tests that have the smallest possible Type II error (or the largest possible power) for a fixed significance level. More theoretical treatise discuss the existence and construction of *uniformly most powerful tests*. Note that the latter do not exist for two-sided settings or for tests involving more than one parameter. Most tests that we discuss in this book are “optimal” (the one-sided version being uniformly most powerful under the statistical assumptions). ♣

The workflow of a hypothesis test is very similar to the one of a statistical significance test and only point (ii) and (iii) need to be slightly modified:

(i) Formulation of the statistical model and statistical assumptions. Formulate the scientific hypothesis in terms of a statistical one.

(ii) Selection of the appropriate test or test statistic, significance level and formulation of the null hypothesis H_0 and the alternative hypothesis H_1 with the parameters of the test.

(iii) Calculation of the test statistic value (or p-value), comparison with critical value (or level) and decision

(iv) Interpretation of the results of the statistical test and conclusion.

The choice of test is again constrained by the assumptions. The significance level must, however, always be chosen before the computations.

The value of the test statistic, say t_{obs}, calculated in step (iii) is compared with critical values t_{crit} in order to reach a decision. When the decision is based on the calculation of the p-value, it consists of a comparison with α. The p-value can be difficult to calculate, but is valuable because of its direct interpretation as the strength (or weakness) of the evidence against the null hypothesis.

General remarks about statistical tests

In the following, we will present several statistical tests in text blocks like this one. In general, we denote

- n, n_x, n_y, ... the sample size;
- $x_1, \dots, x_{n_x}$, $y_1, \dots, y_{n_y}$, samples, independent observations;
- $\bar{x}$, $\bar{y}$, ... the sample mean;
- s^2, s_x^2, s_y^2, ... the sample variance, e.g., $s_x^2 = \dfrac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$ (in the tests under consideration, the variance is unknown);
- α the significance level ($0 < \alpha < 1$, but α typically small);
- t_{crit}, F_{crit}, ... the critical values, i.e., the quantiles according to the distribution of the test statistic and the significance level.

We forumlate our scientific question in general terms under *Question*. The statistical or formal assumptions summarizing the statistical model are given under *Assumptions*. Generally, two-sided tests are performed.
For most tests, there is a corresponding R function. The arguments `x`, `y` usually represent vectors containing the data and `alpha` the significance level. From the output, it is possible to get the p-value.

In this book we consider typical settings and discuss common and appropriate tests. The choice of test is primarily dependent on the quantity being tested (location, scale, frequencies, ...) and secondly on the statistical model and assumptions. The tests will be summarized in yellowish boxes similar as given here. The following list of tests can be used as a decision tree.

- Tests involving the location (mean, means, medians):
  - one sample (Test 1 in Section 5.3.1)
  - two samples:
    - two paired/dependent samples (Test 3 in Section 5.3.3 and Test 8 in Chapter 7)
    - two independent samples (Test 2 in Section 5.3.1 and Test 9 in Chapter 7)
  - several samples (Test 11 in Chapter 7 and Test 16 in Chapter 12)

- Tests involving variances:
  - one sample (Problem 5.4)
  - two samples (Test 4 in Section 5.3.4)
- Tests to compare frequencies:
  - one proportions
  - two proportions (Test 5 in Chapter 6)
  - distributions (Test 6 in Chapter 6)

Of course this list is not exhaustive and many additional possible tests exists and are frequently used. Moreover, the approaches described in the first two sections allow to construct arbitrary tests.

We present several of these tests in more details by motivating the test statistic, giving an explicit example and by summarizing the test in yellow boxes. Ultimately, we perform test with a single call in R. However, the underlying mechanism has to be understood, it would be too dangerous using statistical tests as black-box tools only.

5.3 Testing Means and Variances in Gaussian Samples

In this section, we compare means and variances from normally distributed samples. Formally, we assume that the data $y_1, \dots, y_n$ is a realization of $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. This distributional assumption is required to use the results of Section 3.2 and implies that our results (p-values, Type I and II error probabilities) are exact. If there is a discrepancy between the data and the statistical model, the results are at most approximate. In many situations, the approximation is fairly good because of the central limit theorem and we can proceed nevertheless but interpret the results accordingly. In Chapter 7, we see an alternative and we will relax the Gaussian assumption entirely.

5.3.1 Comparing a Sample Mean with a Theoretical Value

We revisit the setting when comparing the sample mean with a hypothesized value (e.g., observed pododermatitis score with the value 3.33). As stated above, $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with parameter of interest μ but now unknown σ^2. Thus, from

$$\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n) \qquad \Longrightarrow \qquad \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \overset{\sigma \text{ unknown}}{\Longrightarrow} \qquad \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \tag{5.4}$$

The null hypothesis $H_0 : \mu = \mu_0$ specifies the hypothesized mean and the distribution in (5.4).

This test is tyically called the “one-sample t-test”, for obvious reasons. To calculate p-values the function `pt(..., df=n-1)` (for sample size `n`) is used.

The box Test 1 summarizes the test and Example 5.8 illustrates the test based on the pododermatitis data.

Test 1: Comparing a sample mean with a theoretical value

Question: Does the sample mean deviate significantly from the postulated but unknown mean?

Assumptions: The observed data is a realization of a Gaussian random sample with unknown mean and variance.

Calculation: $t_{\text{obs}} = \dfrac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$.

Decision: Reject $H_0 : \mu = \mu_0$, if $t_{\text{obs}} > t_{\text{crit}} = t_{n-1,1-\alpha/2}$.

Calculation in R: `t.test( x, mu=mu0, conf.level=1-alpha)`

Example 5.8 (continuation of Example 5.2)**.** We test the hypothesis that the animals have a different pododermatitis score than low-grade hyperkeratosis, corresponding to 3.333. The sample mean is larger and we want to know if the difference is large enough for a statistical claim.

The statistical null hypothesis is that the mean score is equal to 3.333 and we want to know if the mean of the (single) sample deviates from a specified value, sufficiently for a statistical claim. Although there might be a preferred direction of the test (higher score than 3.333), we perform a two-sided hypothesis test. From R-Code 5.1 we have for the sample mean $\bar{x} = 3.869$, sample standard deviation $s = 0.638$ and sample size $n = 17$. Thus,

$$H_0 : \mu = 3.333 \text{ versus } H_1 : \mu \neq 3.333;$$

$$t_{\text{obs}} = \frac{|3.869 - 3.333|}{0.638/\sqrt{17}} = 3.467;$$

$$t_{\text{crit}} = t_{16,1-0.05/2} = 2.120 \qquad p\text{-value: } 0.003.$$

Formally, we can reject our H_0 because $t_{\text{obs}} > t_{\text{crit}}$. The p-value can be calculated with `2*(1-pt( tobs, n-1))` with `tobs` defined as 3.467. This calculation is equivalent to `2*pt( -tobs, n-1)`. The p-value is low and hence there is substantial evidence against the null hypothesis.

R-Code 5.3 illustrates the direct testing in R with the function `t.test()` and subsequent extraction of the p-value. ♣

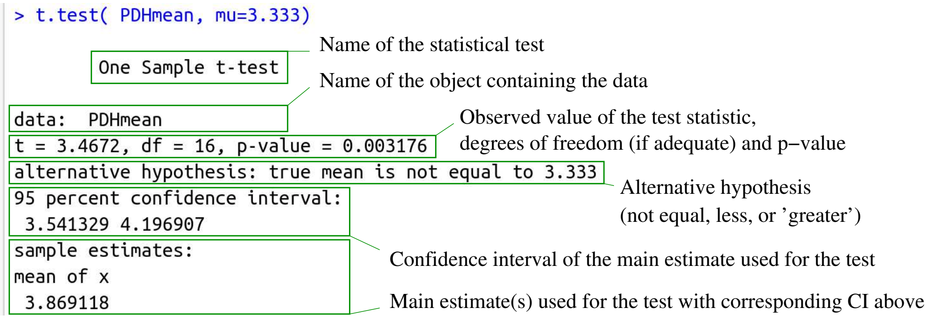
The returned object of the function `t.test()` (as well as of virtually all other test functions we will see) is of class `htest`. Hence, the output looks always similar and is summarized in Figure 5.6 for the particular case of the example above.

5.3.2 Comparing Sample Means of two Samples

Comparing means of two different samples is probably the most often used statistical test. To introduce this test, we assume that both random samples are normally distributed with equal sample size and variance, i.e., $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_x, \sigma^2)$, $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_y, \sigma^2)$. Further, we

R-Code 5.3 One sample t-test, *pododermatitis* (see Example 5.8 and Test 1)

```
print( out <- t.test( PDHmean, mu=3.333))    # print the result of the test
##
##  One Sample t-test
##
## data:  PDHmean
## t = 3.47, df = 16, p-value = 0.0032
## alternative hypothesis: true mean is not equal to 3.333
## 95 percent confidence interval:
##  3.5413 4.1969
## sample estimates:
## mean of x
##    3.8691
out$p.val                                    # printing only the p-value
## [1] 0.0031759
```



```
> t.test( PDHmean, mu=3.333)

        One Sample t-test

data:  PDHmean
t = 3.4672, df = 16, p-value = 0.003176
alternative hypothesis: true mean is not equal to 3.333
95 percent confidence interval:
 3.541329 4.196907
sample estimates:
mean of x
 3.869118
```

Name of the statistical test
Name of the object containing the data
Observed value of the test statistic,
degrees of freedom (if adequate) and p−value
Alternative hypothesis
(not equal, less, or 'greater')
Confidence interval of the main estimate used for the test
Main estimate(s) used for the test with corresponding CI above

Figure 5.6: R output of an object of class *htest*.

assume that both random samples are independent. Under these assumptions we have

$$\overline{X} \sim \mathcal{N}\Big(\mu_x, \frac{\sigma^2}{n}\Big), \qquad \overline{Y} \sim \mathcal{N}\Big(\mu_y, \frac{\sigma^2}{n}\Big) \qquad \Longrightarrow \qquad \overline{X} - \overline{Y} \sim \mathcal{N}\Big(\mu_x - \mu_y, \frac{2\sigma^2}{n}\Big) \tag{5.5}$$

$$\Longrightarrow \quad \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sigma/\sqrt{n/2}} \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad \frac{\overline{X} - \overline{Y}}{\sigma/\sqrt{n/2}} \overset{H_0: \mu_x = \mu_y}{\sim} \mathcal{N}(0,1). \tag{5.6}$$

As often in practice, we do not know σ and we have to estimate it. One possible estimate is so-called pooled estimate $s_p^2 = (s_x^2 + s_y^2)/2$, where s_x^2 and s_x^2 are the variance estimates of the two samples. When using the estimator S_p^2 in the righ-hand expression of (5.6), the distribution of $(\overline{X} - \overline{Y}) \Big/ (S_p/\sqrt{n/2})$ is a t-distribution with $2n - 2$ degrees of freedom. This result is not surprising (up to the degrees of freedom) but somewhat difficult to show formally.

If the sample sizes are different, we need to adjust the pooled estimate and the form is slightly more complicated (see Test 2). As the calculation s_p^2 requires the estimates μ_x and μ_y, we adjust the degrees of freedom to $2n - 2$ or $n_x + n_y - 2$ in case of different sample sizes.

The following example revisits the pododermatitis data again and compares the scores between the two different barns.

Test 2: Comparing means from two independent samples

Question: Are the means and of two samples significantly different?

Assumptions: Both samples are normally distributed with the same unknown variance. The samples are independent.

Calculation: $t_{\text{obs}} = \dfrac{|\bar{x} - \bar{y}|}{s_p \big/ \sqrt{1/n_x + 1/n_y}} = \dfrac{|\bar{x} - \bar{y}|}{s_p} \cdot \sqrt{\dfrac{n_x \cdot n_y}{n_x + n_y}},$

where $s_p^2 = \dfrac{1}{n_x + n_y - 2} \cdot \big((n_x - 1) s_x^2 + (n_y - 1) s_y^2 \big)\,.$

Decision: Reject $H_0 : \mu_x = \mu_y$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n_x+n_y-2,1-\alpha/2}\,.$

Calculation in R: `t.test( x, y, var=TRUE, conf.level=1-alpha)`

Example 5.9 (continuation of Example 5.2)**.** We question if the pododermatitis scores of the two barns are significantly different (means 3.83 and 3.67; standard deviations: 0.88 and 0.87; sample sizes: 20 and 14). Hence, using the formulas given in Test 2, we have

$$H_0 : \mu_x = \mu_y \text{ versus } H_1 : \mu_x \neq \mu_y$$

$$s_p^2 = \frac{1}{20 + 14 - 2}(19 \cdot 0.884^2 + 13 \cdot 0.868^2) = 0.878$$

$$t_{\text{obs}} = \frac{|3.826 - 3.675|}{\sqrt{0.878}} \sqrt{\frac{20 \cdot 14}{20 + 14}} = 0.494$$

$$t_{\text{crit}} = t_{32,1-0.05/2} = 2.037 \qquad p\text{-value: } 0.625.$$

Hence, 3.826 and 3.675 are not statistically different. See also R-Code 5.4, were we use again the function `t.test()` but with two data vectors and the argument `var.equal=TRUE`. ♣

In practice, we often have to assume that the variances of both samples are different, say σ_x^2 and σ_y^2. In such a setting, we have to normalize the mean difference by $\sqrt{s_x^2/n_x + s_y^2/n_y}$. While this estimate seems simpler than the pooled estimate s_p, the degrees of freedom of the resulting t-distribution is difficult to derive, and we refrain to elaborate it here (Problem 5.1.**d** gives some insight). In the literature, this test is called *Welch's two sample t-test* and is actually the default choice of `t.test( x, y)`.

5.3.3 Comparing Sample Means from two Paired Samples

In many situations we have paired measurements at two different time points, before and after a treatment or intervention, from twins, etc. Analyzing the different time points should be done on an individual level and not on the difference of the sample means of the paired samples.

The assumption of independence of both samples in the previous Test 2 may not be valid if the two samples consist of two measurements of the same individual, e.g., observations over two different instances of time. In such settings, were we have a “before” and “after” measurement, it

R-Code 5.4 Two-sample t-test with independent samples, `pododermatitis` (see Example 5.9 and Test 2).

```
t.test( PDHmeanB1, PDHmeanB2, var.equal=TRUE)
##
##  Two Sample t-test
##
## data:  PDHmeanB1 and PDHmeanB2
## t = 0.495, df = 32, p-value = 0.62
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4717  0.7742
## sample estimates:
## mean of x mean of y
##    3.8262    3.6750
```

would be better to take this pairing into account, by considering differences only instead of two samples. Hence, instead of constructing a test statistic based on $\overline{X} - \overline{Y}$ we consider

$$X_1 - Y_1, \dots, X_n - Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_x - \mu_y, \sigma_d^2) \quad\Longrightarrow\quad \overline{X} - \overline{Y} \sim \mathcal{N}\Big(\mu_x - \mu_y, \frac{\sigma_d^2}{n}\Big) \tag{5.7}$$

$$\Longrightarrow \quad \frac{\overline{X} - \overline{Y}}{\sigma_d/\sqrt{n}} \overset{H_0:\mu_x=\mu_y}{\sim} \mathcal{N}(0,1). \tag{5.8}$$

where σ_d^2 is essentially the sum of the variances minus the "dependence" between X_i and Y_i. We formalize this dependence, called covariance, starting in Chapter 8.

The paired two-sample t-test can thus be considered a one sample t-test of the differences with mean $\mu_0 = 0$.

Example 5.10. We consider the pododermatitis measurements from July 2016 and June 2017 and test if there is a progression over time. We have the following summaries for the differences (see R-Code 5.5 and Test 3). Mean $\bar{d} = 0.21$; standard deviation $s_d = 1.26$; and sample size $n = 17$.

$H_0 : d = 0$ versus $H_1 : d \neq 0$; or equivalently $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x \neq \mu_y$;

$t_{\text{obs}} = \dfrac{|0.210|}{1.262/\sqrt{17}} = 0.687;$

$t_{\text{crit}} = t_{16;0.05} = 2.12$ $\qquad$ p-value: 0.502.

There is no evidence that there is a progression over time. ♣

5.3.4 Comparing Sample Variances of two Samples

Just as means can be compared, there are also tests to compare variances of two samples, where, instead of taking differences, we take the ratio of the estimated variances. If the two estimates are similar, the ratio should be close to one. The test statistic is accordingly S_x^2/S_y^2

Test 3: Comparing means from two paired samples

Question: Are the means $\bar{x}$ and $\bar{y}$ of two paired samples significantly different?

Assumptions: The samples are paired. The differences are normally distributed with unknown mean δ. The variance is unknown.

Calculation: $t_{\text{obs}} = \dfrac{|\,\bar{d}\,|}{s_d/\sqrt{n}}$, where

- $d_i = x_i - y_i$ is the i-th observed difference,
- $\bar{d}$ and s_d are the mean and the standard deviation of the differences d_i .

Decision: Reject $H_0 : \delta = 0$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n-1,1-\alpha/2}$.

Calculation in R: `t.test(x, y, paired=TRUE, conf.level=1-alpha)` or

`t.test(x-y, conf.level=1-alpha)`

R-Code 5.5 Two-sample t-test with paired samples, `pododermatitis` (see Example 5.10 and Test 3).

```
podoV1V13 <- podo[podo$Visit %in% c(1,13),] # select visits from 2016 and 2017
PDHmean2 <- matrix(podoV1V13$PDHmean[order(podoV1V13$ID)], ncol=2, byrow=TRUE)
t.test( PDHmean2[,2], PDHmean2[,1], paired=TRUE)
##
##  Paired t-test
##
## data:  PDHmean2[, 2] and PDHmean2[, 1]
## t = 0.687, df = 16, p-value = 0.5
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.43870  0.85929
## sample estimates:
## mean difference
##         0.21029
# Same result as with
# t.test( PDHmean2[,2] - PDHmean2[,1])
```

and the distribution thereof is an F-distribution (see also Section 3.2.3 with quantile, density, and distribution functions implemented in R with `[q,d,p]f`). This “classic” F-test is given in Test 4.

In latter chapters we will see more natural settings, where we need to compare variances (not necessary from a priori two different samples).

Test 4: Comparing two variances

Question: Are the variances s_x^2 and s_y^2 of two samples significantly different?

Assumptions: Both samples are normally distributed and independent.

Calculation: $F_{\text{obs}} = \dfrac{s_x^2}{s_y^2}$

Decision: Reject H_0: $\sigma_x^2 = \sigma_y^2$ if $F_{\text{obs}} > F_{\text{crit}}$, where F_{crit} is the $1-\alpha/2$ quantile of an F-distribution with $n_x - 1$ and $n_y - 1$ degrees of freedom or if $F_{\text{obs}} < F_{\text{crit}}$, where F_{crit} is the $\alpha/2$ quantile of an F-distribution with $n_x - 1$ and $n_y - 1$ degrees of freedom.

Calculation in R: `var.test( x, y, conf.level=1-alpha)`

Example 5.11 (continuation of Example 5.2)**.** As shown by R-Code 5.6 the pododermatitis mean scores of the two barns do not have any evidence against the null hypothesis of having equal variances. ♣

R-Code 5.6 Comparison of two variances, *PDH* (see Example 5.11 and Test 4).

```
var.test( PDHmeanB1, PDHmeanB2)
##
##  F test to compare two variances
##
## data:  PDHmeanB1 and PDHmeanB2
## F = 1.04, num df = 19, denom df = 13, p-value = 0.97
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.35027 2.78376
## sample estimates:
## ratio of variances
##             1.0384
```

It is important to note that for the two-sample t-test we should not test first if the variances are equal and then deciding on whether to use the classical or Welch's two-sample t-test. With such a sequential approach, we cannot maintain the nominal significance level (we use the same data for several tests, see also Section 5.5.2). It should be rather the experimental setup that should argue if conceptually the variances should be equivalent (see also Chapter 12).

Remark 5.2. When flipping the samples in Test 4, the value of the observed test statistic and the associated confidence interval of σ_y^2/σ_x^2 changes. However, the p-value of the test remains the same, see the output of `var.test( PDHmeanB2, PDHmeanB1)`. This is because if $W \sim F_{n,m}$ then $1/W \sim F_{m,n}$. ♣

5.4 Duality of Tests and Confidence Intervals

There is a close connection between a significance test for a particular parameter and a confidence interval for the same parameter. Rejecting $H_0 : \theta = \theta_0$ with significance level α is equivalent to θ_0 not being in the $(1-\alpha)$ confidence interval of θ.

To illustrate, we consider the one sample t-test where we compare the mean with a theoretical value μ_0. As shown in Test 1, $H_0 : \mu = \mu_0$ is rejected if

$$t_{\text{obs}} = \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} > t_{\text{crit}} = t_{n-1,1-\alpha/2}. \tag{5.9}$$

Hence, H_0 cannot be rejected if

$$t_{\text{obs}} = \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} \leq t_{\text{crit}} = t_{n-1,1-\alpha/2} \iff |\bar{x}-\mu_0| \leq t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}}, \tag{5.10}$$

which also means that H_0 is not rejected if

$$-t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \bar{x}-\mu_0 \quad \text{and} \quad \bar{x}-\mu_0 \leq t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}}. \tag{5.11}$$

We can again rewrite this as

$$\bar{x}-t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}} \leq \mu_0 \quad \text{and} \quad \mu_0 \leq \bar{x}+t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}}, \tag{5.12}$$

which correspond to the boundaries of the sample $(1-\alpha)$ confidence interval for μ_0. Analogously, this duality can be established for the other tests described in this chapter.

Example 5.12 (reconsider the situation from Example 5.8)**.** The 95%-confidence interval for the mean μ is $[\,3.54, 4.20\,]$. Since the value of $\mu_0 = 3.33$ is not in this interval, the null hypothesis $H_0 : \mu = \mu_0 = 3.33$ is rejected at level $\alpha = 5\%$. ♣

In R, most test functions return the corresponding confidence intervals (named element `$conf.int` of the returned list) together with the value of the statistic (`$statistic`), p-value (`$p.value`) and other information. Some test functions may explicitly require setting additional argument `conf.int=TRUE`.

5.5 Missuse of p-Values and Other Dangers

p-Values and their use have been criticized lately. By not carefully and properly performing statistical tests it is possible to “obtain” p-values that are small (especially smaller than 0.05), to “observe” a significant result. In this section we discuss and illustrate a few possible pitfalls of statistical testing. Note that wrong statistical results are often due to insufficient statistical knowledge and not due to deliberate manipulation of data or suchlike.

5.5.1 Interpretation of p-Values

The definition and interpretation of p-values are not as easy as it seems and quite often lead to confusion or a misinterpretation. Some scientific journals went as far to ban articles with p-values altogether. In the last few years many articles have emerged discussing what p-values are and what they are not, often jointly with the interpretation of confidence intervals. Here, we cite verbatim from the ASA's Statement (Wasserstein and Lazar, 2016):

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance does measure the size of and effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

See also Greenland *et al.* (2016).

5.5.2 Multiple Testing and p-Value Adjustments

In many cases, we want to perform not just one test, but a series of tests for the same data. We must then be aware that the significance level α only holds for a single test. In the case of a single test, the probability of a falsely significant test result equals the significance level, usually $\alpha = 0.05$. The probability that the null hypothesis H_0 is correctly not rejected is then $1 - 0.05 = 0.95$.

Consider the situation in which $m > 1$ tests are performed. The probability that at least one false significant test result is obtained is then equal to one minus the probability that no false significant test results are obtained. It holds that

$$\begin{aligned} \text{P(at least 1 false significant results)} &= 1 - \text{P(no false significant results)} && (5.13)\\ &= 1 - (1-\alpha)^m. && (5.14) \end{aligned}$$

Table 5.2 gives the probabilities of at least one false significant result for $\alpha = 0.05$ and various m. Even for just a few tests, the probability increases drastically.

Table 5.2: Probabilities of at least one false significant test result when performing m tests at level $\alpha = 5\%$ (top row) and at level $\alpha_{\text{new}} = \alpha/m$ (bottom row).

m	1	2	3	4	5	6	8	10	20	100
$1-(1-\alpha)^m$	0.05	0.098	0.143	0.185	0.226	0.265	0.337	0.401	0.642	0.994
$1-(1-\alpha_{\text{new}})^m$	0.05	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049

There are several different methods that allow multiple tests to be performed while maintaining the selected significance level. The simplest and most well-known of them is the Bonferroni correction. Instead of comparing the p-value of every test to α, they are compared to a new significance level, $\alpha_{\text{new}} = \alpha/m$, see second row of Table 5.2.

There are several alternative methods, which, according to the situation, may be more appropriate. We recommend to use at least `method="holm"` (default) in `p.adjust`. For more details see, for example, Farcomeni (2008).

5.5.3 p-Hacking, HARKing and Publication Bias

There are other dangers with p-values. It is often very easy to tweak the data such that we observe a significant p-value (declaring values as outliers, removing certain observations, use secondary outcomes of the experiment). Such tweaking is often called p-hacking: manage the data until we get a significant result.

Hypothesizing after the results are known (*HARKing*) is another inappropriate scientific practice in which a post hoc hypothesis is presented as an a priori hypotheses. In a nutshell, we collect the data of the experiment and adjust the hypothesis after we have analysed the data, e.g., select effects small enough such that significant results have been observed.

Along similar lines, analyzing a dataset with many different methods will likely lead to several significant p-values. In fact, even if in the case of the true underlying null hypothesis, on average $\alpha \cdot 100\%$ of the tests are significant. Due to various inherent decisions often even more. When searching for a good statistical analysis one often has to make many choices and thus inherently selects the best one among many. This danger is often called the 'garden of forking paths'. Conceptually, adjusting the p-value for the many (not-performed) test would mitigate the problem.

If a result is not significant, the study is often not published and is left in a 'file-drawer'. A seemingly significant result might be well due to Type I error but this is not evident as many similar experiments lead to non-significant outcomes that are not published. Hence, the so-called publication bias implies that there are more Type I error results than the nominal α level.

For many scientific domains, it is possible to *preregister* the study, i.e., to declare the study experiment, analysis methods, etc. before the actual data has been collected. In other words, everything is determined except the actual data collection and actual numbers of the statistical analysis. The idea is that the scientific question is worthwhile investigating and reporting independent of the actual outcome. Such an approach reduces HARKing, garden-of-forking-paths issue, publication bias and more.

5.6 Bibliographic Remarks

Virtually all introductory statistics books contain material about statistical testing. The classics Lehmann and Casella (1998); Lehmann and Romano (2005) are quite advanced and mathematical.

5.7 Exercises and Problems

Problem 5.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Derive the power of a one-sample t-test for $H_0 : \mu = \mu_0$ and $H_0 : \mu = \mu_1$ for sample size n.

b) Show that the value of the test statistic of the two sample t-test with equal variances and equal sample sizes simplifies to $\sqrt{n}(\bar{x} - \bar{y}) \big/ \sqrt{s_x^2 + s_y^2}$.

c) Starting from results in (5.6), derive the test statistic of Test 2 for the case general case of sample sizes n_x and n_y.

d) We give some background to Welch's two sample t-test with test statistic $\big(\overline{X} - \overline{Y}\big)/S_W$ with $S_W^2 = S_x^2/n_x + S_y^2/n_y$. The distribution thereof will not be exactly a t-distribution because $\sigma_x \neq \sigma_y$. And thus the denominator is not exactly a scaled chi-squared random variable. However, it can be shown that rS_W^2/σ_W^2 is approximately chi-squared with r degrees of freedom. Determine r such that $\text{Var}(rS_W^2/\sigma_W^2) = 2r$.

Problem 5.2 (t-Test) Use again the sickle-cell disease data introduced in Problem 4.2. For the cases listed below, specify the null and alternative hypothesis. Then use R to perform the tests and give a careful interpretation.

a) $\mu_{\text{HbS}\beta} = 10$ ($\alpha = 5\%$, two-sided)

b) $\mu_{\text{HbS}\beta} = \mu_{\text{HbSS}}$ ($\alpha = 1\%$, two-sided)

c) What changes, if one-sided tests are performed instead?

Problem 5.3 (t-Test) Anorexia is an eating disorder that is characterized by low weight, food restriction, fear of gaining weight and a strong desire to be thin. The dataset `anorexia` in the package `MASS` gives the weight of 29 females before and after a cognitive behavioral treatment (in pounds). Test whether the treatment was effective.

Problem 5.4 (Testing the variance of one sample) In this problem we develop a "one-sample variance" test. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We consider the estimator S^2 for the parameter σ^2. We test $H_0 : \sigma^2 = \sigma_0^2$ against $H_0 : \sigma^2 \geq \sigma_0^2$.

a) What is the distribution of $(n-1)S^2/\sigma^2$?

b) Find an expression for the p-value for the estimate s^2. For simplicity, we assume $s^2 > \sigma_0^2$.

c) We now assume explicit values $n = 17$, $s^2 = 0.41$ and $\sigma_0^2 = 0.25$. What is the p-value?

d) Construct a one-sided sample confidence interval $[0, b_u]$ for the parameter σ^2 in the general setting and with the values from **c**).

Problem 5.5 (p-values under mis-specified assumptions) In this problem investigate the effect of deviations of statistical assumptions on the p-value. For simplicity, we use the one sample t-test.

a) For 10000 times, sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$, $\sigma = 1$ and $n = 10$. For each sample perform a t-test for $H_0 : \mu = 0$. Plot the p-values in a histogram. What do you observe? For $\alpha = 0.05$, what is the observed Type I error?

b) We repeat the experiment with a different distribution. Same questions as in **a**), but for

- $X_1, \dots, X_n$ from a t-distribution with 4 degrees of freedom and $H_0 : \mu = 0$;
- $X_1, \dots, X_n$ from a chi-square distribution with 10 degrees of freedom and $H_0 : \mu = 10$.

c) Will the observed Type I error be closer to the nominal level α when we increase n? Justify.

Problem 5.6 (Misinterpretation of overlapping confidence intervals when comparing groups) Suppose we have measurements $x_1, \dots, x_{n_x}$ and $y_1, \dots, y_{n_y}$ from two groups. There is a common misinterpretation that if the confidence intervals of the corresponding means are overlapping then the corresponding two-sample test will not show a statistically significant difference of the two samples. In this problem we analyze the setting and learn to understand the reason of this “apparent” violation of the duality of between tests and confidence intervals.

For simplicity, assume that we have $n = n_x = n_y$ and $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_x, \sigma^2)$ and $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_y, \sigma^2)$, $\mu_x < \mu_y$.

a) We assume $n = 25$, $\sigma^2 = 1$ and $\alpha = 5\%$. Generate for $\mu_x = 0$ a sample $x_1, \dots, x_n$ and for $\mu_y = 0$ a sample $y_1, \dots, y_n$. Shift the sample $y_1, \dots, y_n$ such that both confidence intervals “touch”. Based on the shifted sample, what is the p-value of the two sample t-test?

Based on trial-and-error, which shift would yield approximately a p-value of $\alpha = 5\%$?

Hint: use `t.test(...)$conf` to access the sample confidence intervals.

b) Calculate the difference between the lower bound of the sample confidence interval for μ_y and upper bound of the sample confidence interval for μ_x. Show that the two intervals “touch“ each other when $(\bar{x} - \bar{y}) \Big/ \big((s_x + s_y)/\sqrt{n}\big) = t_{n-1,1-\alpha/2}$.

c) Which test would be adequate to consider here? What is t_{obs} and t_{crit} in this specific setting?

d) Comparing **b**) and **c**), why do we not have this apparent duality between the confidence intervals and the hypothesis test?

Problem 5.7 (BMJ Endgame) Discuss and justify the statements about ‘Independent samples t test’ given in doi.org/10.1136/bmj.c2673.

Chapter 6

Estimating and Testing Proportions

Learning goals for this chapter:

- ◇ Identify if the situation involves proportion
- ◇ Explain and apply estimation, confidence interval and hypothesis testing for proportions
- ◇ Compare different CI for a single proportion (Wald, Wilson)
- ◇ Explain and apply different methods of comparing proportions (difference between proportions, odds ratio, relative risk)
- ◇ Explain the concept of Person's chi-squared test

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter06.R.

One of the motivational examples in the last chapter was tossing a coin. In this chapter, we generalize this specific idea and discuss the estimation and testing of a single proportion as well as two or several proportions. The following example serves as a motivation.

The end-of-semester exam of the lecture 'Analysis for the natural sciences' at my university consisted of two sightly different versions. The exam type is assigned according to the student listing. It is of utmost importance, that the two versions are identical. For one particular year, among the 589 students that participated, 291 received exam version A, the others version B. There were 80 students failing version A versus 85 failing version B. Is there enough evidence in the data to claim that the exams were not of equal difficulty and hence some students were disadvantaged.

In this chapter we have a closer look at statistical techniques that help us to correctly answer the above and similar questions. More precisely, we will estimate and compare proportions. To simplify the exhibition, we discuss the estimation of one proportion followed by comparing two proportions. The third section discusses statistical test for different cases.

6.1 Estimation

We start with a simple setting where we observe occurrences of a certain event and are interested in the proportion of the events over the total population. More specifically, we consider the number of successes in a sequence of experiments, e.g., whether a certain treatment had an effect or whether a certain test has been passed. We first discuss point estimation followed by the construction a confidence intervals for a proportion.

6.1.1 Point Estimation for a Proportion

For our setting, we use a binomial random variable $X \sim \mathcal{B}in(n, p)$, where n is given or known and p is unknown, the parameter of interest. Intuitively, we find x/n to be an estimate of p and X/n the corresponding estimator.

More formally, with the method of moments we obtain the estimator $\widehat{p}_{\text{MM}} = X/n$, since $np = \text{E}(X)$ and we have only one observation (total number of cases). The estimator is thus identical to the intuitive estimator.

The likelihood estimator is constructed as follows:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{6.1}$$

$$\ell(p) = \log\big(L(p)\big) = \log\binom{n}{x} + x\log(p) + (n-x)\log(1-p) \tag{6.2}$$

$$\frac{d\ell(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p} \qquad \Longrightarrow \qquad \frac{x}{\widehat{p}_{\text{ML}}} = \frac{n-x}{1-\widehat{p}_{\text{ML}}} \qquad \Longrightarrow \qquad x - x\widehat{p}_{\text{ML}} = n\widehat{p}_{\text{ML}} - x\widehat{p}_{\text{ML}}. \tag{6.3}$$

Thus, the maximum likelihood estimator is (again) $\widehat{p}_{\text{ML}} = X/n$.

In our example we have the following estimates: $\widehat{p}_{\text{A}} = 211/291 \approx 72.51\%$ for exam version A and $\widehat{p}_{\text{B}} = 213/298 \approx 71.48\%$ for version B. Once we have an estimate of an proportion, we can now answer questions (for each version separately), such as:

1. How many cases of failure can be expected in a group of 100 students?
2. What is the probability that more than 100 failures occur in a cohort of 300 students?

(where we use results from Sections 2.2.1 and 3.3 only). We revisit questions like “Is the failure rate significantly lower than 30%?” later in the chapter.

The estimator $\widehat{p} = X/n$ does not have a “classical” distribution (it is a “scaled binomial”). Figure 6.1 illustrates the probability mass function based on the estimate of exam version A. The figure visually suggest to use a Gaussian approximation. Formally, we approximate the binomial distribution of X by a Gaussian distribution (see Section 3.3), which is well justified here as $np(1-p) \gg 9$. The approximation for X is then used to state that the estimator $\widehat{p}$ is also approximately Gaussian with adjusted parameters: $\widehat{p} \overset{\text{app}}{\sim} \mathcal{N}(x/n, p(1-p)/n)$. In this chapter, we will often use this approximation and thus we implicitly assume $np(1-p) > 9$.

Figure 6.1 also indicates that shifting the Gaussian density slightly to the right, the approximation would improve. This shift is linked to the *continuity correction* and is performed in practice. In our derivations we often omit the correction for clarity.

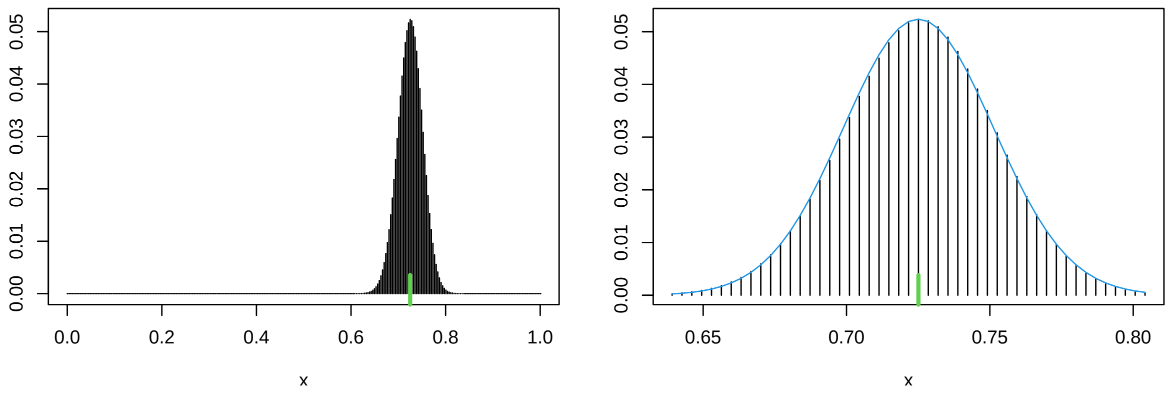


Figure 6.1: Probability mass function of $\widehat{p} = 211/291$. The blue curve in the right panel is the normal approximation. The actual estimate is indicated with a green tick mark.

When dealing with proportions we often speak of odds, or simply of chance, defined by $\omega = p/(1-p)$. The corresponding intuitive estimator is $\widehat{\omega} = \widehat{p}/(1-\widehat{p})$ for an estimator $\widehat{p}$. Similarly, $\widehat{\theta} = \log(\widehat{\omega}) = \log\big((\widehat{p}/(1-\widehat{p})\big)$ is an intuitive estimator of log odds. If $\widehat{p}$ is an estimate, then the quantities are the corresponding estimates. As a side note, these estimators also coincide with the maximum likelihood estimators.

6.1.2 Confidence Intervals for a Proportion

To construct a confidence intervals for the parameter p we use the Gaussian approximation for the binomial distribution and thus we have

$$1-\alpha \approx \mathrm{P}\Big(z_{\alpha/2} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq z_{1-\alpha/2}\Big). \tag{6.4}$$

This can be rewritten as

$$\begin{aligned} 1-\alpha &\approx \mathrm{P}\Big(z_{\alpha/2}\sqrt{np(1-p)} \leq X-np \leq z_{1-\alpha/2}\sqrt{np(1-p)}\Big) && (6.5)\\ &= \mathrm{P}\Big(-\frac{X}{n} + z_{\alpha/2}\frac{1}{n}\sqrt{np(1-p)} \leq -p \leq -\frac{X}{n} + z_{1-\alpha/2}\frac{1}{n}\sqrt{np(1-p)}\Big). && (6.6)\end{aligned}$$

As a further approximation we replace p with $\widehat{p}$ in the argument of the square root to obtain

$$1-\alpha \approx \mathrm{P}\Big(-\frac{X}{n} + z_{\alpha/2}\frac{1}{n}\sqrt{n\widehat{p}(1-\widehat{p})} \leq -p \leq -\frac{X}{n} + z_{1-\alpha/2}\frac{1}{n}\sqrt{n\widehat{p}(1-\widehat{p})}\Big). \tag{6.7}$$

Since $\widehat{p} = x/n$ (as estimate) and $q := z_{1-\alpha/2} = -z_{\alpha/2}$, we have as sample confidence interval

$$b_{l,u} = \widehat{p} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \widehat{p} \pm q \cdot \mathrm{SE}(\widehat{p})\,. \tag{6.8}$$

Remark 6.1. For (regular) models with parameter θ, as $n \to \infty$, likelihood theory states that, the estimator $\widehat{\theta}_{\mathrm{ML}}$ is normally distributed with expected value θ and variance $\mathrm{Var}(\widehat{\theta}_{\mathrm{ML}})$.

Since $\mathrm{Var}(X/n) = p(1-p)/n$, one can assume that $\mathrm{SE}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$. The so-called Wald confidence interval rests upon this assumption (which can be shown more formally) and is identical to (6.8). ♣

If the inequality in (6.4) is solved through a quadratic equation, we obtain the sample Wilson confidence interval

$$b_{l,u} = \frac{1}{1+q^2/n} \cdot \left(\widehat{p} + \frac{q^2}{2n} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} \right), \tag{6.9}$$

where we use the estimates $\widehat{p} = x/n$ (see Problem 6.1.**a**).

CI 3 summarizes both the Wald and Wilson confidence interval. The latter can be constructed in R with *prop.test(x, n, correct=FALSE)$conf.int*. An exact $(1-\alpha)$ confidence interval for a proportion based on quantiles of the binomial distriubtion is computed with *binom.test(x, n)$conf.int*.

CI 3: Confidence intervals for proportions

An approximate $(1-\alpha)$ Wald confidence interval for a proportion is

$$B_{l,u} = \widehat{p} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{6.10}$$

with estimator $\widehat{p} = X/n$ and quantile $q = z_{1-\alpha/2}$.
An approximate $(1-\alpha)$ Wilson confidence interval for a proportion is

$$B_{l,u} = \frac{1}{1+q^2/n} \cdot \left(\widehat{p} + \frac{q^2}{2n} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} \right). \tag{6.11}$$

The Wilson confidence interval is "more complicated" than the Wald confidence interval. Is it also "better" because of one fewer approximation during the derivation?

Ideally the coverage probability of a $(1-\alpha)$ confidence interval should be $1-\alpha$. Because of the approximations this will not be the case and the coverage probability can be used to assess confidence intervals. For a discrete random variable, the coverage is

$$\mathrm{P}(p \in [\,B_l, B_u\,]) = \sum_{x=0}^{n} \mathrm{P}(X = x) \mathbb{I}_{\{p \in [\,b_l, b_u\,]\}}. \tag{6.12}$$

(see Problem 6.1.**b**). R-Code 6.1 calculates the coverage of the 95% confidence intervals for $X \sim \mathcal{B}in(n = 40, p = 0.4)$. For the particular setting, the Wilson confidence interval does not seem to have a better coverage (96% compared to 94%).

R-Code 6.1: Coverage of 95% confidence intervals for $X \sim \mathcal{B}in(n = 40, p = 0.4)$.

```
p <- .4 ;   n <- 40            # defining the binomial random variable
x <- 0:n                       # sequence of all possible values
WaldCI <- function(x, n){      # CI formula based on eq (6.8)
  mid <- x/n
  se <- sqrt(x*(n-x)/n^3)
  cbind( pmax(0, mid - 1.96*se),  pmin(1, mid + 1.96*se))
}
WaldCIs <- WaldCI(x,n)         # calculate CI for our values n and x
Waldind <- (WaldCIs[,1] <= p) & (WaldCIs[,2] >= p) # p in CI?
Waldcoverage <- sum( dbinom(x, n, p)*Waldind)    # eq (6.12)

WilsonCI <- function(x, n){    # CI formula based on eq (6.9)
  mid <- (x + 1.96^2/2)/(n + 1.96^2)
  se <- sqrt(n)/(n+1.96^2)*sqrt(x/n*(1-x/n)+1.96^2/(4*n))
  cbind( pmax(0, mid - 1.96*se),  pmin(1, mid + 1.96*se))
}
WilsonCIs <- WilsonCI(x,n)     # calculate CI for our values n and x
Wilsonind <- (WilsonCIs[,1] <= p) & (WilsonCIs[,2] >= p)  # p in CI?
Wilsoncoverage <- sum( dbinom(x, n, p)*Wilsonind)
print( c(true=0.95, Wald=Waldcoverage, Wilson=Wilsoncoverage))
##    true    Wald  Wilson
## 0.95000 0.94587 0.96552
```

Figure 6.2 illustrates the coverage for the confidence intervals for different value of p. Overall the Wilson confidence interval now dominates the Wald one. The Wilson confidence interval has better nominal coverage at the center. This observation also holds when n is varied, as seen in Figure 6.3 which shows the coverage for different values of n and p. The Wilson confidence interval has a slight tendency of too large coverage (more blueish areas) but is overall much better than the Wald one. Note that the top "row" of the left and right part of the panel corresponds to the left and right part of Figure 6.2.

R-Code 6.2 Tips for construction of Figure 6.2.

```
p <- seq(0, 0.5, .001)    # we exploit the symmetry
  # either a loop over all elements in p or a few `apply()`'s
  # over the functions 'Wilsonind' and 'Wilsoncoverage'
  # `Waldcoverage` and `Wilsoncoverage` are thus vectors!
plot( p, Waldcoverage, type='l', ylim=c(.8,1))
Waldsmooth <- loess( Waldcoverage ~ p, span=.12)      # "guide the eye" curve
lines( Waldsmooth$x, Waldsmooth$fitted, col=3, lw=2) # add to plot.
lines( c(-1, 2), c(.95, .95), col=2, lty=2)          # nominal level
```

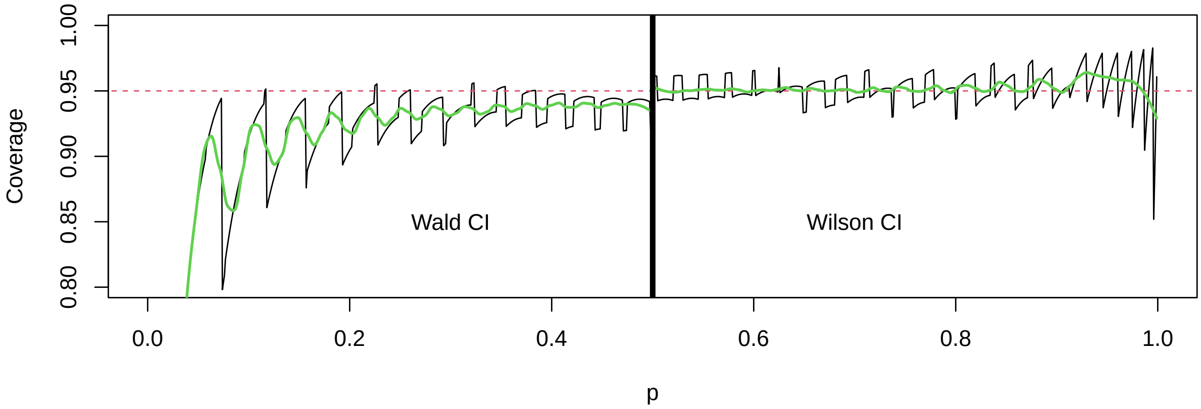


Figure 6.2: Coverage of the 95% confidence intervals for $X \sim \mathcal{B}in(n = 40, p)$ for the Wald CI (left) and Wilson CI (right). The red dashed line is the nominal level $1 - \alpha$ and in green we have a smoothed curve to "guide the eye". (See R-Code 6.2.)

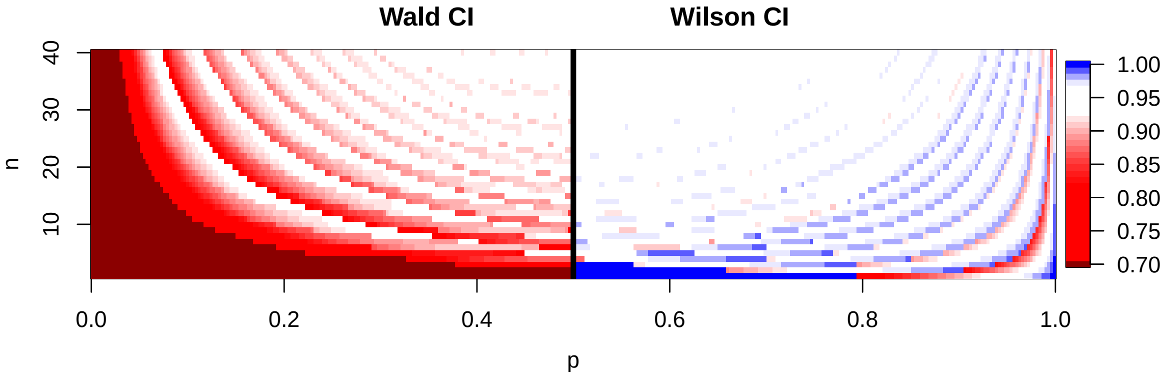


Figure 6.3: Coverage of the 95% confidence intervals for $X \sim \mathcal{B}in(n, p)$ as functions of p and n. The probabilities are symmetric around $p = 1/2$. All values smaller than 0.7 are represented with dark red. Left is for the Wald CI, the right is for the Wilson CI.

The width of a sample confidence interval is $b_u - b_l$. For the Wald confidence interval we obtain

$$2q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 2q\sqrt{\frac{x(n-x)}{n^3}} \tag{6.13}$$

and for the Wilson confidence interval we have

$$\frac{2q}{1+q^2/n}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} = \frac{2q}{1+q^2/n}\sqrt{\frac{x(n-x)}{n^3} + \frac{q^2}{4n^2}}. \tag{6.14}$$

The widths vary with the observed value of X and are shown in Figure 6.4. For $5 < x < 36$, the Wilson confidence interval has a smaller width and a better nominal coverage (over small ranges of p). For small and very large values x, the Wald confidence interval has a too small coverage and thus wider intervals are desired.

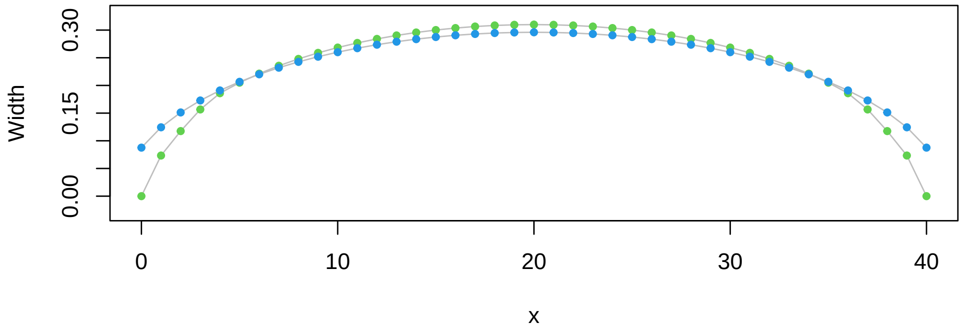


Figure 6.4: Widths of the sample 95% confidence intervals for $X \sim \mathcal{B}in(n = 40, p)$ as a function of the observed value x (the Wald CI is in green, the Wilson CI in blue).

6.2 Comparison of two Proportions

We assume that we have two binomial random variables $X_1 \sim \mathcal{B}in(n_1, p_1)$ and $X_2 \sim \mathcal{B}in(n_2, p_2)$ that define two groups. The resulting observations are often presented in a 2×2 *contingency table* as shown in Table 6.1. Sometimes, n_i are also denoted by r_i (row totals). The goal of this section is to introduce formal approaches for a comparison of two proportions p_1 and p_2. This can be accomplished using (i) a difference $p_1 - p_2$, (ii) a quotient p_1/p_2, or (iii) an odds ratio $p_1/(1-p_1)\Big/(p_2/(1-p_2))$, which we consider in the following three sections. The approaches are illustrated based on the following example.

Table 6.1: Example of a two-dimensional contingency table displaying frequencies. The first index refers to the group category and the second to the result category.

		Result		
		positive	negative	Total
Group	A	h_{11}	h_{12}	n_1
	B	h_{21}	h_{22}	n_2
	Total	c_1	c_2	n

Example 6.1. Pre-eclampsia is a hypertensive disorder occurring during pregnancy (gestational hypertension) with symptoms including: edema, high blood pressure, and proteinuria. In a double-blinded randomized controlled trial (RCT) 2706 pregnant women were treated with either a diuretic or with a placebo (Landesman *et al.*, 1965). Pre-eclampsia was diagnosed in 138 of the 1370 subjects in the treatment group and in 175 of the subjects receiving placebo. The medical question is whether diuretic medications, which reduce water retension, reduce the risk of pre-eclampsia.

The two treatments (control/placebo and diuretic medication) define the risk factor, which is the group category and the diagnosis is the result category in Table 6.1. R-Code 6.3 presents the 2×2 table. ♣

R-Code 6.3 Contingency table for the pre-eclampsia data of Example 6.1.

```
xD <- 138;          xC <- 175          # positive diagnosed counts
nD <- 1370;         nC <- 1336         # totals in both groups
tab <- rbind( Diuretic=c(xD, nD-xD), Control=c(xC, nC-xC))
colnames( tab) <- c( 'pos', 'neg')
tab

##          pos  neg
## Diuretic 138 1232
## Control  175 1161
```

6.2.1 Difference Between Proportions

The risk difference RD describes the (absolute) difference in the probability of experiencing the event in question.

Using the notation introduce above, the difference $h_{11}/(h_{11}+h_{12}) - h_{21}/(h_{21}+h_{22}) = h_{11}/n_1 - h_{21}/n_2$ can be seen as a realization of $X_1/n_1 - X_2/n_2$, which is approximately normally distributed

$$\mathcal{N}\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big) \tag{6.15}$$

(based on the normal approximation of the binomial distribution). Hence, a corresponding confidence interval can be derived.

6.2.2 Relative Risk

The relative risk estimates the size of the effect of a risk factor compared with the size of the effect when the risk factor is not present:

$$\text{RR} = \frac{\text{P(Positive diagnosis with risk factor)}}{\text{P(Positive diagnosis without risk factor)}}. \tag{6.16}$$

The groups with or without the risk factor can also be considered the treatment and control groups.

The relative risk is a positive values. A value of $RR = 1$ means that the risk is the same in both groups and there is no evidence of a association between the diagnosis/disease/event and the risk factor. A $RR \geq 1$ is evidence of a possible positive association between a risk factor and a diagnosis/disease. If the relative risk is less than one, the exposure has a protective effect, as is the case, for example, for vaccinations.

An estimate of the relative risk is (see Table 6.1)

$$\widehat{\text{RR}} = \frac{\widehat{p}_1}{\widehat{p}_2} = \frac{\dfrac{h_{11}}{h_{11}+h_{12}}}{\dfrac{h_{21}}{h_{21}+h_{22}}} = \frac{h_{11}n_2}{h_{21}n_1}. \tag{6.17}$$

To construct confidence intervals, we consider first $\widehat{\theta} = \log(\widehat{\text{RR}})$. The standard error of $\widehat{\theta}$ is determined with the delta method and based on Equation (3.31), applied to a Binomial instead of a Bernoulli random variable:

$$\text{Var}(\,\widehat{\theta}\,) = \text{Var}\big(\,\log(\widehat{\text{RR}})\,\big) = \text{Var}\Big(\log(\frac{\widehat{p}_1}{\widehat{p}_2})\Big) = \text{Var}\big(\log(\widehat{p}_1) - \log(\widehat{p}_2)\big) \tag{6.18}$$

$$= \text{Var}\big(\log(\widehat{p}_1)\big) + \text{Var}((\log(\widehat{p}_2)\big) \approx \frac{1-\widehat{p}_1}{n_1 \cdot \widehat{p}_1} + \frac{1-\widehat{p}_2}{n_2 \cdot \widehat{p}_2} \tag{6.19}$$

$$\approx \frac{1 - \dfrac{h_{11}}{h_{11}+h_{12}}}{(h_{11}+h_{12}) \cdot \dfrac{h_{11}}{h_{11}+h_{12}}} + \frac{1 - \dfrac{h_{22}}{h_{21}+h_{22}}}{(h_{21}+h_{22}) \cdot \dfrac{h_{22}}{h_{21}+h_{22}}} \tag{6.20}$$

$$= \frac{1}{h_{11}} - \frac{1}{h_{11}+h_{12}} + \frac{1}{h_{21}} - \frac{1}{h_{21}+h_{22}}\,. \tag{6.21}$$

A back-transformation

$$\Big[\,\exp\big(\,\widehat{\theta} \pm z_{1-\alpha/2}\,\text{SE}(\widehat{\theta})\,\big)\,\Big] \tag{6.22}$$

implies positive confidence boundaries. Note that with the back-transformation we loose the 'symmetry' of estimate plus/minus standard error.

CI 4: Confidence interval for relative risk (RR)

An approximate $(1-\alpha)$ confidence interval for RR, based on the two-dimensional contingency table (Table 6.1), is

$$\Big[\exp\Big(\,\log(\widehat{\text{RR}}) \pm z_{1-\alpha/2}\,\text{SE}\big(\log(\widehat{\text{RR}})\big)\,\Big)\Big] \tag{6.23}$$

where for the sample confidence intervals we use the estimates

$$\widehat{\text{RR}} = \frac{h_{11}(h_{21}+h_{22})}{(h_{11}+h_{12})h_{21}}, \quad \text{SE}\big(\log(\widehat{\text{RR}})\big) = \sqrt{\frac{1}{h_{11}} - \frac{1}{h_{11}+h_{12}} + \frac{1}{h_{21}} - \frac{1}{h_{21}+h_{22}}}\,.$$

Example 6.2 (continuation of Example 6.1)**.** The relative risk and corresponding confidence interval for the pre-eclampsia data are given in R-Code 6.4. The relative risk is smaller than one (diuretics reduce the risk). An approximate 95% confidence interval does not include the value one. ♣

The relative risk cannot be applied to so-called case-control studies, where we match for each subject in the risk group one or several subjects in the second, control group. This matching implies that the risk in the control group is not representative but influenced by the first group.

R-Code 6.4 Relative Risk with confidence interval.

```
print( RR <- ( tab[1,1]/ sum(tab[1,])) /  ( tab[2,1]/ sum(tab[2,])) )
## [1] 0.769
s <- sqrt( 1/tab[1,1] + 1/tab[2,1] - 1/sum(tab[1,]) - 1/sum(tab[2,]) )
exp( log(RR) + qnorm(c(.025,.975))*s)
## [1] 0.62333 0.94872
```

6.2.3 Odds Ratio

The relative risk is closely related to the *odds ratio*, which is defined as

$$\text{OR} = \frac{\dfrac{\text{P(Positive diagnosis with risk factor)}}{\text{P(Negative Diagnosis with risk factor)}}}{\dfrac{\text{P(Positive diagnosis without risk factor)}}{\text{P(Negative diagnosis without risk factor)}}} = \frac{\dfrac{\text{P}(A)}{1-\text{P}(A)}}{\dfrac{\text{P}(B)}{1-\text{P}(B)}} = \frac{\text{P}(A)(1-\text{P}(B))}{\text{P}(B)(1-\text{P}(A))} \tag{6.24}$$

with A and B the positive diagnosis with and without risk factors. The odds ratio indicates the strength of an association between factors (association measure). The calculation of the odds ratio also makes sense when the number of diseased is determined by study design, as is the case for case-control studies. When a disease is rare (very low probability of disease), the odds ratio and relative risk are approximately equal.

The odds ratio is one of the most used association measures in statistics. Additionally there are several statistical models that are linked to odds ratio.

An estimate of the odds ratio is

$$\widehat{\text{OR}} = \frac{\dfrac{h_{11}}{h_{12}}}{\dfrac{h_{21}}{h_{22}}} = \frac{h_{11}\,h_{22}}{h_{12}\,h_{21}}\,. \tag{6.25}$$

The construction of confidence intervals for the odds ratio is based on Equation (3.30) and Equation (3.31), analogous to that of the relative risk.

CI 5: Confidence interval for the odds ratio (OR)

An approximate $(1-\alpha)$ confidence interval for OR, based on a two-dimensional contingency table (Table 6.1), is

$$\Big[\exp\Big(\log(\widehat{\text{OR}}) \pm z_{1-\alpha/2}\,\text{SE}(\log(\widehat{\text{OR}}))\Big)\Big] \tag{6.26}$$

where for the sample confidence intervals we use the estimates

$$\widehat{\text{OR}} = \frac{h_{11}h_{22}}{h_{12}h_{21}} \quad\text{and}\quad \text{SE}(\log(\widehat{\text{OR}})) = \sqrt{\frac{1}{h_{11}} + \frac{1}{h_{21}} + \frac{1}{h_{12}} + \frac{1}{h_{22}}}\quad .$$

R-Code 6.5 Odds ratio with confidence interval, approximate and exact.

```
print( OR <- tab[1]*tab[4]/(tab[2]*tab[3]))
## [1] 0.74313
s <- sqrt( sum( 1/tab) )
exp( log(OR) + qnorm(c(.025,.975))*s)
## [1] 0.58626 0.94196
# Exact test from Fisher:
fisher.test(tab)
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.016
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.58166 0.94826
## sample estimates:
## odds ratio
##    0.74321
```

Example 6.3 (continuation of Example 6.1)**.** The odds ratio with confidence interval for the pre-eclampsia data is given in R-Code 6.5. The 95% confidence interval is again similar as calculated for the relative risks and does also not include one, strengthening the claim (i.e., significant result).

Notice that the function *fisher.test()* also calculates the odds ratio. As it is based on a likelihood calculation, there are very minor differences between both estimates. ♣

6.3 Statistical Tests

In this section we look at the statistical tests for proportions. We separate the discussion whether the test involves a single, two or several proportions.

6.3.1 Tests Involving a Single Proportion

We start the discussion for the hypothesis test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. This case is straightforward when relying on the duality of test and confidence intervals. That means, we reject the null hypothesis at level α if p_0 is not in the $(1-\alpha)$ sample confidence interval $[\,b_l, b_u\,]$. The confidence interval can be obtained based on a Wald, Wilson or some other approach. If the confidence interval is not exact, the test may not have exact size α.

In R, there is the possibility to use *binom.test(n,p)*, *prop.test(n,p)*, the latter with the argument *correct=TRUE* (default) to include a continuity correction or *correct=FALSE*.

Example 6.4. In one of his famous experiments Gregor Mendel crossed peas based on AA and aa-type homozygotes. Under the Mendelian inheritance assumption, the third generation should consist of AA and Aa genotypes with a ratio of 1:2. For one particular genotype, Mendel reported the counts 8 and 22 (see, e.g., Table 1 of Ellis *et al.*, 2019). We cannot reject the hypothesis $H_0: p = 1/3$ based on the p-value 0.56. As shown in R-Code 6.6, both `binom.test(8, 8+22, p=1/3)` and `prop.test(8, 8+22, p=1/3)` yield the same p-value (up to two digits). The corresponding confidence intervals are also very similar. Whereas when using the argument `correct=FALSE` the outcome changes noticeably. ♣

R-Code 6.6 Testing a proportion.

```
rbind(binom=unlist( binom.test(8, 8+22, p=1/3)[c("p.value", "conf.int")]),
      prop1=unlist( prop.test(8, 8+22, p=1/3)[c("p.value", "conf.int")]),
      prop2=unlist( prop.test(8, 8+22, p=1/3, correct=FALSE)[c("p.value",
                                                               "conf.int")]))

##       p.value conf.int1 conf.int2
## binom 0.56215   0.12279   0.45889
## prop1 0.56128   0.12975   0.46173
## prop2 0.43858   0.14183   0.44448
```

6.3.2 Tests Involving two Proportions

We now consider the 2×2 framework as shown in Table 6.1. More specifically, we assume that each row of the table shows the successes and failures of a binomial random variable. A test for equality of the proportions tests $H_0: p_1 = p_2$, where p_1 and p_2 are the two proportions.

One way to derive such a test is to start with the difference of proportions. Under the null hypothesis, (6.15) simplifies to $\mathcal{N}\big(0, p(1-p)(1/n_1+1/n_2)\big)$ and a test statistic can be derived. In practice one often works with the squared difference of the proportions for which the test statistic takes quite a simple form, as given in Test 5. Under the null hypothesis, the distribution thereof is a chi-squared distribution (square of a normal random variable). The quantile, density, and distribution functions are implemented in R with `[q,d,p]chisq` (see Section 3.2.1). This test is also called Pearson's χ^2 test.

Example 6.5 (continuation of Example 6.1)**.** The R-Code 6.7 shows the results for the pre-eclampsia data, once using a proportion test and once using a chi-squared test (comparing expected and observed frequencies). ♣

Remark 6.2. We have presented the rows of Table 6.1 in terms of two binomials, i.e., with two fixed marginals. In certain situations, such a table can be seen from a hypergeometric distribution point of view (see `help( dhyper)`), where three margins are fixed. For this latter view, `fisher.test` is the test of choice. ♣

Test 5: Test of proportions

Question: Are the proportions in the two groups the same?

Assumptions: Both samples are independent from a binomial distribution with equal probability.

Calculation: Using the same notation as in Table 6.1,

$$\chi^2_{\text{obs}} = \frac{(h_{11}h_{22} - h_{12}h_{21})^2(h_{11} + h_{12} + h_{21} + h_{22})}{(h_{11} + h_{12})(h_{21} + h_{22})(h_{12} + h_{22})(h_{11} + h_{21})} = \frac{(h_{11}h_{22} - h_{12}h_{21})^2 n}{n_1 n_2 c_1 c_2}.$$

Decision: Reject if $\chi^2_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{1,1-\alpha}$.

Calculation in R: `prop.test( tab)` or `chisq.test(tab)` with default continuity correction.

R-Code 6.7 Test of proportions

```
det(tab)^2*sum(tab) / prod(c(rowSums(tab), colSums(tab))) # as in Test 5
## [1] 6.0541
chisq.test( tab, correct=FALSE)  # same value of the test statistic
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 6.05, df = 1, p-value = 0.014
prop.test( tab)   # continuity correction by default
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tab
## X-squared = 5.76, df = 1, p-value = 0.016
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0551074 -0.0054088
## sample estimates:
##  prop 1  prop 2
## 0.10073 0.13099
```

6.3.3 Tests Involving a Several Proportions

We now extend the tests of the last two sections by first considering so-called multinomial setting and then second by extending the 2×2-tables to general two-way tables. Both settings result in a so-called chi-square test.

In the binomial setting we have two proportions p and $1-p$ and a test $H_0: p = p_0$ consists of comparing x with $p_0 n$ or equivalently $n - x$ with $(1-p_0)n$. If there is a large discrepancy between the observed and the theoretical counts, there is evidence against the null hypothesis. Hence it is quite natural to extend the same idea to the setting of K proportions $(p_1, \dots, p_K)$ with observed counts $(x_1, \dots, x_K)$. Note that $\sum_{i=1}^K x_i = n$ and $\sum_{i=1}^K p_i = 1$. Hence we have only $K-1$ free parameters and in the binomial setting, $K = 2$.

There is a natural extension of the test of proportions in which we compare "arbitrary" many proportions. The so-called chi-square test ($\mathcal{X}^2$ test) compares if the observed data follow a particular distribution by comparing the frequencies of binned observations with the expected frequencies.

Under the null hypothesis, the chi-square test is $\mathcal{X}^2$ distributed (Test 6). The test is based on approximations and thus the categories should be aggregated so that all bins contain a reasonable amount of counts, e.g., $e_i \geq 5$ (trivially, $K - k > 1$).

Test 6: Comparison of observations with expected frequencies

Question: Do the observed frequencies o_i of a sample deviate significantly from the expected frequencies e_i of a certain distribution?

Assumptions: Sample with data from any scale.

Calculation: Calculate the observed values o_i and the expected frequencies e_i (with the help of the expected distribution) and then compute

$$\chi^2_{\text{obs}} = \sum_{i=1}^{K} \frac{(o_i - e_i)^2}{e_i}$$

where K is the number of categories.

Decision: Reject H_0: "no deviation between the observed and expected" if $\chi^2_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{K-1-k,1-\alpha}$, where k is the number of parameters estimated from the data to calculate the expected counts.

Calculation in R: `chisq.test( obs, p=expect/sum(obs))` or `chisq.test( obs, p=expect, rescale.p=TRUE)`

Example 6.6. With few observations (10 to 50) it is often pointless to test for normality of the data. Even for larger samples, a Q-Q plot is often more informative. For completeness, we illustrate a simple goodness-of-fit test by comparing the pododermatitits data with expected counts constructed from Gaussian density with matching mean and variance (R-Code 6.8). We pool over both periods and barns ($n = 34$) (there is no significant difference in the means and the variances).

The binning of the data is done through a histogram-type binning (an alternative way would be *table(cut(podo$PDHmean))*). As we have less than five observations in several bins, the function *chisq.test()* issues a warning. This effect could be mitigated if we calculate the p-value using a bootstrap simulation by setting the argument *simulate.p.value=TRUE*. Pooling the bins, say *breaks=c(1.5,2.5,3.5,4,4.5,5)* would be an alternative as well.

The degrees of freedom are $K - 1 - k = 7 - 1 - 2 = 4$, as we estimate the mean and standard deviation to determine the expected counts. ♣

R-Code 6.8 Testing normality, *pododermatitis* (see Example 6.6 and Test 6).

```
observed <- hist( podo$PDHmean, plot=FALSE, breaks=5)
       # without 'breaks' argument there are too many categories
observed[1:2]
## $breaks
## [1] 1 2 3 4 5 6
##
## $counts
## [1]  3  9 27 27  1
m <- mean( podo$PDHmean)
s <- sd( podo$PDHmean)
p <- pnorm( observed$breaks, mean=m, sd=s)
chisq.test( observed$counts, p=diff( p), rescale.p=TRUE)
## Warning in chisq.test(observed$counts, p = diff(p), rescale.p = TRUE):
Chi-squared approximation may be incorrect
##
##  Chi-squared test for given probabilities
##
## data:  observed$counts
## X-squared = 8.71, df = 4, p-value = 0.069
```

General distribution tests (also goodness-of-fit tests) differ from the other tests discussed here, in the sense that they do not test or compare a single parameter or a vector of proportions. Such tests run under the names of Kolmogorov-Smirnov Tests (*ks.test()*), Shapiro–Wilk Normality test (*shapiro.test()*), Anderson–Darling test (*goftest::ad.test()*), etc.

Example 6.7. The dataset *haireye* gives the hair and eye colors of 592 persons collected by students (Snee, 1974). The cross-tabulation data helpful to get information about individual combinations or comparing two combinations. The overall picture is best assessed with

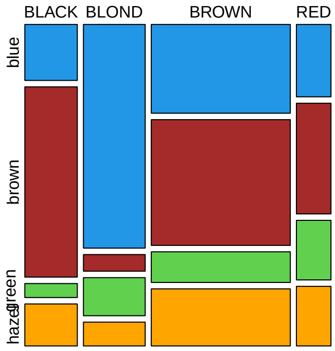


Figure 6.5: Mosaic plot of hair and eye colors. (See R-Code 6.9.)

a mosaic-plot, which indicates that there is a larger proportion of persons with blond hair having blue eyes compared to other hair colors, which is well known. When examining the individual terms $(o_i - e_i)^2/e_i$ of the χ^2 statistic of Test 6, we observe three very large values (`BLONDE-blue`, `BLONDE-brown`, `BLACK-brown`), and thus the very low p-value is no surprise. These residual terms do not indicate if there is an excess or lack of observed pairs. Restricting to persons with brown and red hair only, the eye color seems independent (p-value 0.15, `chisq.test(HAIReye[c("BROWN","RED"),])$p.value`).

R-Code 6.9 Pearson's Chi-squared test for hair and eye colors data. (See Figure 6.5.)

```
HAIReye <- read.csv("data/HAIReye.csv")  # hair color in upper case
mosaicplot(HAIReye, color=c(4,"brown",3,"orange"), main="")
chisq.test(HAIReye)     # Pearson's Chi-squared test for contingency table
##
##  Pearson's Chi-squared test
##
## data:  HAIReye
## X-squared = 138, df = 9, p-value <2e-16
n <- sum(HAIReye)       # 592 persons
rs <- rowSums(HAIReye)  # row totals
cs <- colSums(HAIReye)  # column totals
(outer(rs,cs)/n-HAIReye)^2/(outer(rs,cs)/n) # (o-e)^2/e, sum thereof is 138.3
##          blue      brown   green   hazel
## BLACK  9.4211 19.3459095 3.81688 0.22786
## BLOND 49.6967 34.2341707 0.37540 4.96329
## BROWN  3.8005  1.5214189 0.11909 1.83138
## RED    2.9933  0.0056217 5.21089 0.72633
## Re-run with brown and red colored persons
# (HAIReye <- read.csv("data/HAIReye.csv")[c("BROWN","RED"),])
```

6.4 Bibliographic Remarks

Agresti (2007) or the more technical and advanced version Agresti (2002) are ultimate references for the analysis of categorical data. Brown *et al.* (2002) is a reference for CI for the binomial model.

See also the package `binom` and the references therein for further confidence intervals for proportions.

6.5 Exercises and Problems

Problem 6.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Derive the Wilson confidence interval (6.11).

b) Derive formula (6.12) to calculate the coverage probablity for $X \sim \mathcal{B}in(n, p)$.

c) Derive the test statistic of Test 5 (without continuity correction).

d) Show that the test statistic of Test 5 is a particular case of Test 6 (without continuity correction).

e) Derive standard error of the odds ratio $\text{SE}\big(\log(\widehat{\text{OR}})\big)$.

Problem 6.2 (Binomial distribution) Suppose that among $n = 95$ Swiss males, eight are red-green colour blind. We are interested in estimating the proportion p of people suffering from such disease among the male population.

a) Is a binomial model suitable for this problem?

b) Calculate the maximum likelihood estimate (ML) $\hat{p}_{\text{ML}}$ and the ML of the odds $\hat{\omega}$.

c) Using the central limit theorem (CLT), it can be shown that $\widehat{p}$ follows approximately $\mathcal{N}\big(p, \frac{1}{n}\, p(1-p)\big)$. Compare the binomial distribution to the normal approximation for different n and p. To do so, plot the exact cumulative distribution function (CDF) and compare it with the CDF obtained from the CLT. For which values of n and p is the approximation reasonable? Is the approximation reasonable for the red-green colour blindness data?

d) Use the R functions `binom.test()` and `prop.test()` to compute two-sided 95%-confidence intervals for the exact and for the approximate proportion. Compare the results.

e) What is the meaning of the p-value?

f) Compute the Wilson 95%-confidence interval and compare it to the confidence intervals from (d).

	Treatment A	Treatment B
Cleared	9	5
Not cleared	18	22

Problem 6.3 (A simple clinical trial) A clinical trial is performed to compare two treatments, A and B, that are intended to treat a skin disease named psoriasis. The outcome shown in the following table is whether the patient's skin cleared within 16 weeks of the start of treatment.

Use $\alpha = 0.05$ throughout this problem.

a) Compute for each of the two treatments a Wald type and a Wilson confidence interval for the proportion of patients whose skin cleared.

b) Test whether the risk difference is significantly different to zero (i.e., $RD = 0$). Use both an exact and an approximated approach.

c) Compute CIs for both, relative risk (RR) and odds ratio (OR).

d) How would the point estimate of the odds ratio change if we considered the proportions of patients whose skin did *not* clear?

Problem 6.4 (BMJ Endgame) Discuss and justify the statements about 'Relative risks versus odds ratios' given in doi.org/10.1136/bmj.g1407.

Chapter 7

Rank-Based Methods

Learning goals for this chapter:

- ◇ Explain robust estimators and their properties
- ◇ Explain the motivation and the concept of ranks
- ◇ Describe, apply and interpret rank based tests as counter pieces to the classical t-test
- ◇ Describe the general approach to compare the means of several samples, apply and interpret the rank based tests
- ◇ Explain the idea of a permutation test

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter07.R.

As seen in previous chapters “classic” estimates of the expectation and the variance are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{7.1}$$

If, hypothetically, we set one (arbitrary) value x_i to an infinitely large value (i.e., we create an extreme outlier), these estimates “explode”. A single value may exert enough influence on the estimate such that the estimate is not representative of the bulk of the data anymore. In a similar fashion, outliers may not only influence estimates drastically but also the value of test statistics and thus render the result of the test questionable.

Until now we have often assumed that we have a realization of a Gaussian random sample. We have argued that the t-test family is exact for Gaussian data but remains usable for moderate deviations thereof since we use the central limit theorem in the test statistic. If we have very small sample sizes or if the deviation is substantial, the result of the test may be again questionable.

In this chapter, we discuss basic approaches to estimation and testing for cases that include the presence of outliers and deviations from Gaussianity.

7.1 Robust Point Estimates

A *robust estimators* is an estimator of a parameter that is not sensitive to one or possibly several outliers. Sensitive here should be understood in the sense that if we replace one or possibly several values of the sample with arbitrary values, the corresponding estimate does not or only marginally change.

As argued above, the mean is therefore not a robust estimate of location. However, the trimmed mean (i.e., the biggest and smallest values are trimmed away and not considered) is a robust estimate of location. The sample median (the middle value if the sample size is odd or the center of the two middle-most values otherwise, see (1.2)) is another robust estimate of location.

Robust estimates of the spread are (i) the sample interquartile range (IQR), calculated as the difference between the third and first quartiles, and (ii) the sample median absolute deviation (MAD), calculated as

$$\text{MAD} = c \cdot \text{med}\big(|x_i - \text{med}\{x_1, \dots, x_n\}|\big), \tag{7.2}$$

where most software programs (including R) use $c = 1.4826$. The choice of c is such, that for Gaussian random variables we have an unbiased estimator, i.e., $\text{E}(\text{MAD}) = \sigma$ for MAD seen as an estimator. Since for Gaussian random variables $\text{IQR}= 2\Phi^{-1}(3/4)\sigma$, IQR/1.349 is an estimator of σ; for IQR seen as an estimator.

Example 7.1. Let the values $1.1, 3.2, 2.2, 1.8, 1.9, 2.1, 2.7$ be given. Suppose that we have erroneously entered the final number as 27. R-Code 7.1 compares several statistics (for location and scale) and illustrates the effect of this single outlier on the estimates. ♣

R-Code 7.1 Classic and robust estimates from Example 7.1.

```
sam <- c(1.1, 3.2, 2.2, 1.8, 1.9, 2.1, 27)
print( c(mean(sam), mean(sam, trim=.2), median(sam)))
## [1] 5.6143 2.2400 2.1000
print( c(sd(sam), IQR(sam)/1.349, mad(sam)))
## [1] 9.45083 0.63010 0.44478
```

Remark 7.1. An intuitive approach to quantify the "robustness" of an estimator is the *breakdown point* which quantifies the proportion of the sample that can be set to arbitrary large values before the estimate takes an arbitrary value, i.e., before it breaks down. The mean has a breakdown point of 0, (1 out of n values is sufficient to break the estimate), an α-trimmed mean of α ($\alpha \in [0, 1/2]$), the median of 1/2, the latter is the maximum possible value.

The IQR has a breakdown point of 1/4 and the MAD of 1/2. See also Problem 7.1.a. ♣

Robust estimators have two main drawbacks compared to classical estimators. The first disadvantage is that robust estimators do not possess simple distribution functions and for this reason the corresponding exact confidence intervals are not easy to calculate. More specifically, for a robust estimator $\widehat{\theta}$ of the parameter θ we rarely have the exact quantiles q_l and q_u (which depend on θ), to construct a confidence interval starting from

$$1 - \alpha = P\big(q_l(\theta) \leq \widehat{\theta} \leq q_u(\theta)\big) \tag{7.3}$$

If we could assume that the distribution of robust estimators are somewhat Gaussian (for large samples) we could calculate approximate confidence intervals based on

$$\widehat{\text{robust estimator}} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\text{Var}}(\widehat{\text{robust estimator}})}{n}}, \tag{7.4}$$

which is of course equivalent to $\widehat{\theta} \pm z_{\alpha/2}\,\text{SE}(\widehat{\theta})/n$, for a robust estimator $\widehat{\theta}$. Note that we have deliberately put a *hat* on the variance term in (7.4) as the variance often needs to be estimated as well (which is reflected in a precise definition of the standard error). For example, the R expression `median( x)+c(-2,2)*mad( x)/sqrt(length( x))` yields an approximate sample 95% confidence interval for the median.

The second disadvantage of robust estimators is their lower efficiency, i.e., these estimators have larger variances compared to classical estimators. Formally, the *efficiency* is the ratio of the variance of one estimator to the variance of the second estimator.

In some cases the exact variance of robust estimators can be determined, often approximations or asymptotic results exist. For example, for a continuous random variable with distribution function $F(x)$ and density function $f(x)$, asymptotically, the median is also normally distributed around the true median $\eta = Q(1/2) = F^{-1}(1/2)$ with variance $1/(4nf(\eta)^2)$. The following example illustrates this result and R-Code 7.2 compares the finite sample efficiency of two estimators of location based on repeated sampling.

Example 7.2. Let $X_1, \dots, X_{10} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. We simulate realizations of this random sample and calculate the corresponding sample mean and sample median. We repeat $R = 1000$ times. Figure 7.1 shows the histogram of these means and medians including a (smoothed) density of the sample. The histogram and the density of the sample medians are wider and thus the mean is more efficient.

For this particular example, the sample efficiency is roughly 72% for $n = 10$. As the density is symmetric, $\eta = \mu = 0$ and thus the asymptotic efficiency is

$$\frac{\sigma^2/n}{1/\big(4nf(0)^2\big)} = \frac{\sigma^2}{n} \cdot 4n\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^2 = \frac{2}{\pi} \approx 64\%. \tag{7.5}$$

Of course, if we change the distribution of $X_1, \dots, X_{10}$, the efficiency changes. For example let us consider the case of a t-distribution with 4 degrees of freedom, a density with heavier tails than the normal. Now the sample efficiency for sample size $n = 10$ is 1.26, which means that the median is better compared to the mean. ♣

Robust estimation approaches have the advantage of not having to identify outliers and eliminating these for the estimation process. The decision as to whether a realization of a

R-Code 7.2 Distribution of sample mean and median, see Example 7.2. (See Figure 7.1.)

```
set.seed( 14)          # to reproduce the numbers!
n <- 10                # sample size
R <- 1000              # How often we repeat the sampling
samples <- matrix( rnorm(n*R), nrow=R, ncol=n)  # each row one sample
means <- apply( samples, 1, mean)                # R means
medians <- apply( samples, 1, median)            # R medians
print( c( var( means), var( medians), var( means)/var( medians)))

## [1] 0.10991 0.15306 0.71809

hist( medians, border=7, col=7, prob=TRUE, main='', ylim=c(0, 1.2),
     xlab='estimates')
hist( means, add=TRUE, prob=TRUE)
lines( density( medians), col=2)
lines( density( means), col=1)
# with a t_4 distribution the situation is different:
samples <- matrix( rt(n*R, df=4), nrow=R, ncol=n) # row a t-sample
means <- apply( samples, 1, mean)
medians <- apply( samples, 1, median)
print( c( var( means), var( medians),  var( means)/var( medians)))

## [1] 0.19268 0.15315 1.25815
```

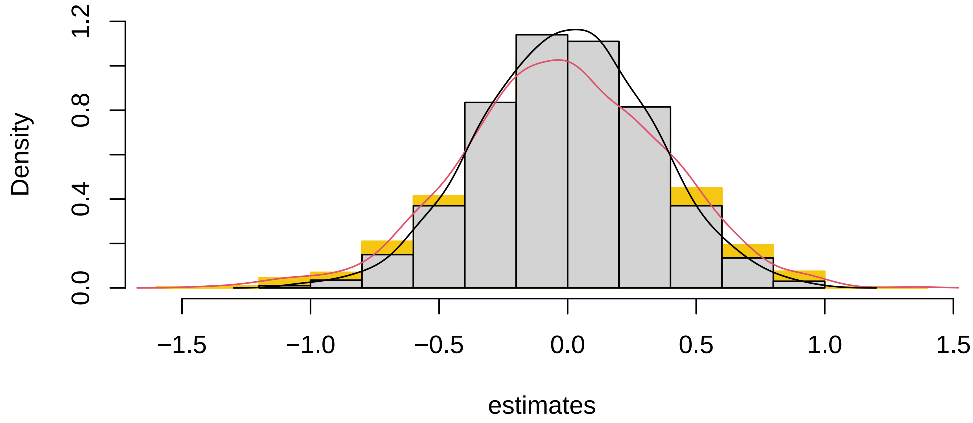


Figure 7.1: Comparing finite sample efficiency of the mean and median for a Gaussian sample of size $n = 10$. Medians in yellow with red smoothed density of the sample, means in black. (See R-Code 7.2.)

random sample contains outliers is not always easy and some care is needed. For example it is not possible to declare a value as an outlier if it lays outside the whiskers of a boxplot. For all distributions with values from $\mathbb{R}$, observations will lay outside the whiskers when n is sufficiently large (see Problem 7.1.**b**). Obvious outliers are easy to identify and eliminate, but in less clear cases robust estimation methods are preferred.

Outliers can be very difficult to recognize in multivariate random samples, because they are not readily apparent with respect to the marginal distributions. Robust methods for random vectors exist, but are often computationally intense and not as intuitive as for scalar values.

It has to be added that independent of the estimation procedures, if an EDA finds outliers, these should be noted and scrutinized.

7.2 Rank-Based Tests as Alternatives to t-Tests

We now turn to tests. To start, recall that the tests to compare means in Chapter 5 assume normally distributed data resulting in test statistics that are t-distributed. Slight deviations from a normal distribution has typically negligible consequences as the central limit theorem reassures that the mean is approximately normally distributed.

However, if outliers are present or the data are skewed or the data are measured on the ordinal scale, the t-test family is not exact and, depending on the degree of departure from the assumption, may not be useful. In such settings, the use of so-called 'rank-based' tests is recommended. Compared to classical tests which typically assume a parametrized distribution (e.g., μ, σ^2 in $\mathcal{N}(\mu, \sigma^2)$ or p in $\mathcal{Bin}(n, p)$), rank based tests do not prescribe a detailed or specific distribution and are thus also called *non-parametric tests*.

To illustrate the nature of a non-parametric test, consider paired data $X_1, \dots, X_n \overset{\text{iid}}{\sim} F_X$, $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} F_Y$, with continuous distributions F_X and F_Y. We consider the null hypothesis H_0 that $\text{P}(X_i < Y_i) = \text{P}(Y_i < X_i) = 1/2$ (because of the continuous distribution $\text{P}(X_i = Y_i) = 0$). In other words, the null hypothesis is that the medians of both distributions are the same. However, the test does not make any assumption on the distributions F_X and F_Y.

The test compares the cases when $x_i < y_i$ with $y_i < x_i$ and we consider the sign of the difference $x_i - y_i$. If there are too many or two few negative signs ($x_i < y_i$) or equivalently too many or two few positive signs ($x_i > y_i$), the data is not supporting the null hypothesis. More formally, under H_0, the sign of each pair $x_i - y_i$ is distributed according a Bernoulli random variable with probability $p = 1/2$.

The resulting statistical test is called the *sign test* and the procedure is illustrated in Test 7. In practice, there might be the case that $y_i = x_i$ and all such cases will be dropped from the sample and the sample size n will be reduced accordingly to $n_\star$. One-sided versions of Test 7 are straightforward to construct, by not taking the minimum in Step (3) and comparing to the α-quantile or $1 - \alpha$-quantile of a $\mathcal{Bin}(n_\star, 1/2)$ distribution, respectively.

The sign test is based on very weak assumptions and therefore has little power. We introduce now the concept of "ranks", as an extension of "signs", and resulting rank tests. The *rank* of a value in a set of values is the position (order) of that value in the ordered sequence (from smallest to largest). In particular, the smallest value has rank 1 and the largest rank n. In the case of ties, the arithmetic mean of the ranks is used.

Example 7.3. The values 1.1, −0.6, 0.3, 0.1, 0.6, 2.1 have ranks 5, 1, 3, 2, 4 and 6. However, the ranks of the absolute values are 5, (3+4)/2, 2, 1, (3+4)/2 and 6. ♣

Rank-based tests consider the ranks of the observations or ranks of the differences, not the observations or the differences between the observations and thus mitigate their effect on the test

Test 7: Comparing the locations of two paired samples with the sign test

Question: Are the medians of two paired samples significantly different?

Assumptions: The samples are paired and from continuous distributions.

Calculation: (1) Calculate the differences $d_i = x_i - y_i$. Ignore all differences $d_i = 0$ and consider only the $n_\star$ differences $d_i \neq 0$.

(2) Categorize each difference d_i by its sign ($+$ or $-$) and count the number of positive signs: $S^+ = \sum_{i=1}^{n_\star} \mathbb{I}_{\{d_i>0\}}$. $S_{\text{obs}} = \min(S^+, n_\star - S^+)$.

Decision: Reject H_0 : "the medians are the same", if $S_{\text{obs}} < S_{\text{crit}}$, where S_{crit} is the $\alpha/2$-quantile of a $\mathcal{B}in(n_\star, 1/2)$ distribution.

Calculation in R:

```
binom.test( sum( d>0), sum( d!=0), conf.level=1-alpha)
```

statistic. For example, the largest value has always the same rank and therefore has the same influence on the test statistic independent of its value.

Compared to classical t-tests as discussed in Chapter 5 and further tests that we will discuss in Chapter 11, rank tests should be used if

- the underlying distributions are skewed.
- the underlying distributions have heavy or light tails compared to the Gaussian distribution.
- if the data has potentially "outliers".

For smaller sample sizes, it is difficult to check model assumptions and it is recommended to use rank tests in such situations. Rank tests have fewer assumptions on the underlying distributions and have a fairly similar power (see Problem 7.6).

We now introduce two classical rank tests which are the Wilcoxon signed rank test and the Wilcoxon–Mann–Whitney U test (also called the Mann–Whitney test), i.e., rank-based versions of Test 2 and Test 3 respectively.

7.2.1 Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is used to test an effect in paired samples, i.e., two matched samples or two repeated measurements on the same subject). As for the sign test or the two-sample paired t-test, we start by calculating the differences of the paired observations, followed by calculating the ranks of the negative differences and of the positive differences. Note that we omit pairs having zero differences and denote with $n_\star$ the possibly adjusted sample size. Under the null-hypotheses, the paired samples are from the same distribution and thus the ranks of the positive and negative differences should be comparable, not be too small or too large. If this is not the case, the data indicates evidence against the hypothesis. This approach is formally presented in Test 8.

Test 8: Comparing the distribution of two paired samples

Question: Are the distributions of two paired samples significantly different?

Assumptions: Both samples are from continuous distributions of the same shape, the samples are paired.

Calculation: (1) Calculate the differences $d_i = x_i - y_i$. Ignore all differences $d_i = 0$ and consider only the remaining $n_\star$ differences $d_i \neq 0$.

(2) Order the $n_\star$ differences d_i by their absolute differences $|d_i|$.

(3) Calculate the sum of the ranks of the positive differences:
$W^+ = \sum_{i=1}^{n_\star} \mathbb{I}_{\{d_i>0\}}$.
$W^- = \frac{n_\star(n_\star+1)}{2} - W^+$ (sum of the ranks of the negative differences).

$$W_{\text{obs}} = \min(W^+, W^-) \qquad\qquad (W^+ + W^- = \frac{n_\star(n_\star+1)}{2})$$

Decision: Reject H_0 : "the distributions are the same" if $W_{\text{obs}} < W_{\text{crit}}(n_\star; \alpha/2)$, where W_{crit} is the critical value.

Calculation in R: `wilcox.test(x-y, conf.level=1-alpha)` or

`wilcox.test(x, y, paired=TRUE, conf.level=1-alpha)`

The quantile, density and distribution functions of the Wilcoxon signed rank test statistic are implemented in R with `[q,d,p]signrank()`. For example, the critical value $W_{\text{crit}}(n; \alpha/2)$ mentioned in Test 8 is `qsignrank( .025, n)` for $\alpha = 5\%$ and the corresponding p-value is `2*psignrank( Wobs, n)` with `n` $= n_\star$.

It is possible to approximate the distribution of test statistic with a normal distribution. The observed test statistic W_{obs} is z-transformed as follows:

$$z_{\text{obs}} = \frac{\left|\, W_{\text{obs}} - \dfrac{n_\star(n_\star+1)}{4} \,\right|}{\sqrt{\dfrac{n_\star(n_\star+1)(2n_\star+1)}{24}}}, \tag{7.6}$$

and z_{obs} is then compared with the corresponding quantile of the standard normal distribution (see Problem 7.1.c). This approximation may be used when the samples are sufficiently large, which is, as a rule of thumb, $n_\star \geq 20$.

In case of ties, R may not be capable to calculate exact p-values and thus will issue a warning. The warning can be avoided by not requiring exact p-values through setting the argument `exact=FALSE`. When setting the argument `conf.int=TRUE` in `wilcox.test()`, a nonparametric confidence interval is constructed. It is possible to specify the confidence level with `conf.level`, default value is 95%. The numerical values of the confidence interval are accessed with the element `$conf.int`, also illustrated in the next example.

Example 7.4 (continuation of Example 5.2)**.** We consider again the `podo` data as introduced in Example 5.2. R-Code 7.3 performs Wilcoxon signed rank test. As we do not have any outliers the p-value is similar to the one obtained with a paired t-test in Example 5.10. There are no ties and thus the argument `exact=FALSE` is not necessary.

The advantage of robust methods becomes clear when the first value from the second visit 3.75 is changed to 37.5, as shown towards the end of the same R-Code. While the p-value of the signed rank test does virtually not change, the one from the paired two-sample t-test changes from 0.5 (see R-Code 5.5) to 0.31. More importantly, the confidence intervals are now drastically different as the outlier inflated the estimated standard deviation of the t-test. In other situations, it is quite likely that with or without a particular "outlier" the p-value falls below the magical threshold α (recall the discussion of Section 5.5.3).

Of course, a corrupt value, as introduced in this example, would be detected with a proper EDA of the data (scales are within zero and ten). ♣

R-Code 7.3: Rank tests and comparison of a paired tests with a corrupted observation.

```
# Possibly relaod the 'podo.csv' and construct the variables as in Example 4.1
wilcox.test(  PDHmean2[,2], PDHmean2[,1], paired=TRUE)
##
##  Wilcoxon signed rank exact test
##
## data:  PDHmean2[, 2] and PDHmean2[, 1]
## V = 88, p-value = 0.61
## alternative hypothesis: true location shift is not equal to 0
# wilcox.test( PDHmean2[,2]-PDHmean2[,1], exact=FALSE)  # is equivalent
PDHmean2[1, 2] <- PDHmean2[1, 2]*10  # corrupted value, decimal point wrong
rbind(t.test=unlist( t.test( PDHmean2[,2], PDHmean2[,1], paired=TRUE)[3:4]),
      wilcox=unlist( wilcox.test( PDHmean2[,2], PDHmean2[,1], paired=TRUE,
            conf.int=TRUE)[c( "p.value", "conf.int")]))
##         p.value conf.int1 conf.int2
## t.test 0.31465   -2.2879    6.6791
## wilcox 0.61122   -0.5500    1.1375
```

For better understanding of the difference between a sign test and the Wilcoxon signed rank test consider an arbitrary distribution $F(x)$. The assumptions of the Wilcoxon signed rank test are $X_1, \dots, X_n \overset{\text{iid}}{\sim} F(x)$ and $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} F(x-\delta)$, where δ represents the shift. Hence, under the null hypothesis, $\delta = 0$ which further implies that for all $i = 1, \dots, n$, (1) $\text{P}(X_i > Y_i) = \text{P}(X_i < Y_i) = 1/2$ and (2) the distribution of the difference $X_i - Y_i$ is symmetric. The second point is not required by the sign test. Hence, the Wilcoxon signed rank test requires more assumptions and has thus generally a higher power.

For a symmetric distribution, we can thus use `wilcox.test()` with argument `mu=mu0` to test $H_0 : \mu = \mu_0$, where μ is the median (and by symmetry also the mean). This setting is the rank test equivalent of Test 1.

7.2.2 Wilcoxon–Mann–Whitney Test

The Wilcoxon–Mann–Whitney test is probably the most well-known rank test and represents the two-sample version of the Wilcoxon signed rank test. To motivate this test, assume that we have two samples from one common underlying distribution. In this case the observations and ranks of both samples mingle nicely. If both samples have the same or quite similar sample sizes, the two sums of the ranks are comparable. Alternatively, assume that the first sample has a much smaller median (or mean) and the ranks of the first sample would be smaller than those of the sample with the larger median (or mean).

More specifically, the Wilcoxon–Mann–Whitney test calculates the rank sums of the two samples and corrects them based on the corresponding sample size. The smaller of the resulting values is then compared with an appropriate quantile, see Test 9. Note that we do not require a symmetric distribution, but rely on the less stringent assumption that the distribution of both samples have the same shape, possibly shifted. Under the null hypothesis, the shift is zero and the Wilcoxon–Mann–Whitney test can be interpreted as comparing the medians between the two distributions.

Test 9: Comparing the locations of two independent samples

Question: Are the medians of two independent samples significantly different?

Assumptions: Both samples are from continuous distributions of the same shape, the samples are independent and the data are at least ordinally scaled.

Calculation: Let $n_x \leq n_y$, otherwise switch the samples. Assemble the $(n_x + n_y)$ sample values into a *single* set ordered by rank and calculate the sum R_x and R_y of the sample ranks. Let

$$U_x = R_x - \frac{n_x(n_x + 1)}{2} \qquad U_y = R_y - \frac{n_y(n_y + 1)}{2}$$

$$U_{\text{obs}} = \min(U_x, U_y) \qquad \text{Note: } U_x + U_y = n_x n_y.$$

Decision: Reject H_0: “medians are the same” if $U_{\text{obs}} < U_{\text{crit}}(n_x, n_y; \alpha/2)$, where U_{crit} is the critical value.

Calculation in R: `wilcox.test(x, y, conf.level=1-alpha)`

The quantile, density and distribution functions of the Wilcoxon–Mann–Whitney test statistic are implemented in R with `[q,d,p]wilcox()`. For example, the critical value $U_{\text{crit}}(n_x, n_y; \alpha/2)$ used in Test 9 is `qwilcox( .025, nx, ny)` for $\alpha = 5\%$ and corresponding p-value `2*pwilcox( Uobs, nx, ny)`.

It is possible to approximate the distribution of the test statistic by a Gaussian one, provided we have sufficiently large samples, e.g., $n_x, n_y \geq 10$. The value of the test statistic U_obs is z-transformed to

$$z_\text{obs} = \frac{\left| U_\text{obs} - \dfrac{n_x n_y}{2} \right|}{\sqrt{\dfrac{n_x n_y (n_x + n_y + 1)}{12}}}, \tag{7.7}$$

where $n_x n_y / 2$ is the mean of all ranks and the denominator is the standard deviation of the sum of the ranks under the null (see Problem 7.1.**d**). This value is then compared with the respective quantile of the standard normal distribution. With additional continuity corrections, the approximation may be improved.

To construct confidence intervals, the argument *`conf.int=TRUE`* must be used in the function *`wilcox.test()`* and unless $\alpha = 5\%$ a specification of *`conf.level`* is required. The numerical values of the confidence interval are accessed with the list element *`$conf.int`*.

Example 7.5 (continuation of Example 5.2 and 7.4)**.** We compare if there is a difference between the pododermatitis scores between both barns. Histogram of the densities support the assumption that both samples come from one underlying distribution. There is no evidence of difference (same conclusion as in Example 5.9).

Several scores appear multiple times and thus the ranks contain ties. To avoid a warning message we set *`exact=FALSE`*. ♣

R-Code 7.4: Two-sample rank test.

```
#  hist(PDHmeanB1, density=15)  # quick check on shape of distribution
#  hist(PDHmeanB2, add=T, border=2, col=2, density=15, angle=-45)
wilcox.test( PDHmeanB1, PDHmeanB2, exact=FALSE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  PDHmeanB1 and PDHmeanB2
## W = 153, p-value = 0.66
## alternative hypothesis: true location shift is not equal to 0
```

Remark 7.2.

1. When constructing a boxplot in R, the argument *`notch=TRUE`* draws a "notch" a the median. The idea is, that if the notches of two boxplots do not overlap then the medians are approximately significantly different at the fixed 95% confidence level. The notches have width of $3.15\text{IQR}/\sqrt{n}$ and are based on asymptotic normality of the median and roughly equal sample sizes for the two medians being compared (McGill *et al.*, 1978).

2. The interpretation of the confidence interval resulting from argument *`conf.int=TRUE`* requires some care, see *`?wilcox.test`*. With this approach, it is possible to set the confidence level via the argument *`conf.level`*. ♣

7.3 Comparing More Than two Groups

Until now we have considered at most two groups. Imagine the situations where n subjects are measured over several times points or where several treatments are applied to different subjects. In such situations we should not perform blindly individual tests comparing all combinations of time points or all combinations of treatments. Based on Section 5.5.2, if we were to administer placebo to six different groups, detecting at least one significant result at level $\alpha = 5\%$ is more likely than not (`1-0.95^choose(6,2)`).

Instead of correcting for multiple testing the following more powerful approach is used. In a first step we test if there is any effect at all. If there is, in a second step, we investigate, which of the groups shows an effect.

In more details, for the first step, we compare the locations of the different groups. In case of an effect, these exhibit a lot of variability, compared to the variability with the individual groups. If there is no effect, the variability between the locations of the groups is small compared to the variability within the groups. Of course some care is needed when comparing the variablities and we likely have to take into account the number of groups and group sizes. In general, such an approach is called *ANOVA* which stands for analysis of variance. We will revisit ANOVA in the Gaussian framework in Chapter 11.

Similar as in the setting of two groups, we also have to differentiate betweeen matched or repeated measures and independent groups. We will discuss two classical rank tests now. The latter two complete the layout shown in Figure 7.2, which summarizes the different non-parametric tests to compare locations (median with a theoretical value or medians of two or several groups). Note that this layout is by no means complete.

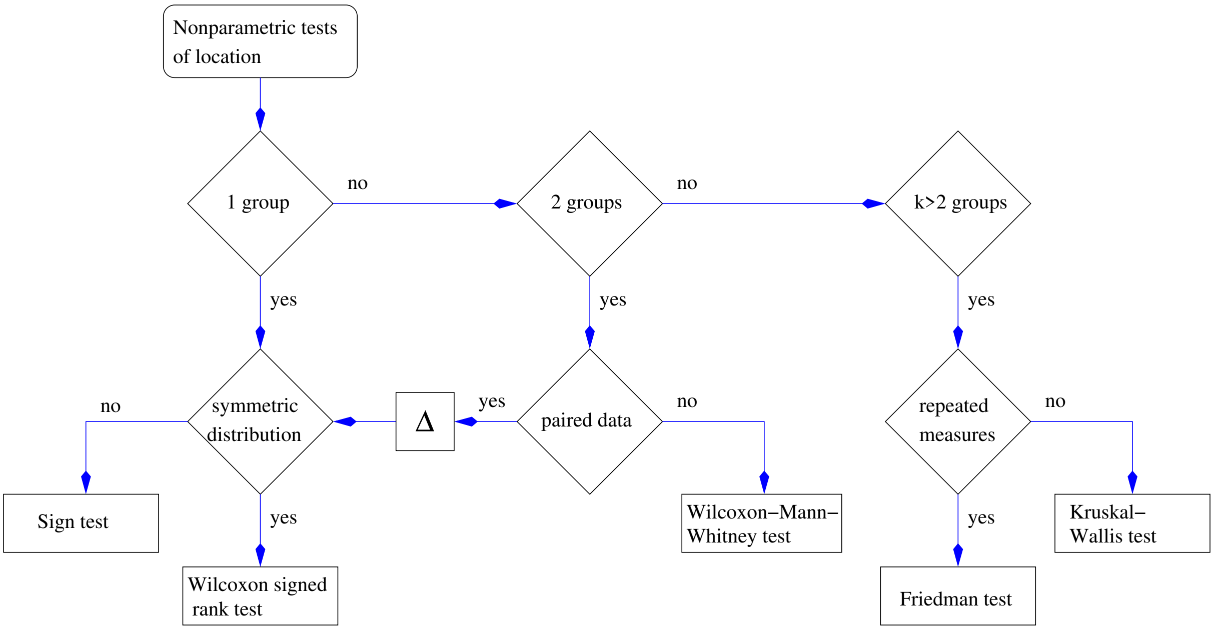


Figure 7.2: Different non-parametric statistical tests of location. In the case of paired data, the Δ operation takes the difference of the two samples and reduces the problem to a single sample setting.

7.3.1 Friedman Test

Suppose we have I subjects that are tested on J different treatments. The resulting data matrix represents for each subject the J measures. While the subjects are assumed independent, the measures within each subject are not necessarily.

Such a simulation setup is also called one-way repeated measures analysis of variance, one-way because each subject receives at any point one single treatment and all other factors are kept the same. We compare the variability between the groups with the variability within the groups based on ranks.

The data is then often presented in a matrix form with I rows and J columns, resulting in a total of $n = IJ$ observations. In a first step we calculate the ranks within each subject. If all the treatments are equal then there is no preference group for small or large ranks. That means, the sum of the ranks within each group R_j are similar. As we have more than two groups we look at the variability of the column ranks. There are different ways to present the test statistics and the following is typically used to calculate the observed value

$$Fr_{\text{obs}} = \Big(\frac{12}{IJ(J+1)}\sum_{j=1}^{J} R_j^2\Big) - 3I(J+1), \tag{7.8}$$

where R_j is the sum of the ranks of the jth column. Under the null hypothesis, the value Fr_{obs} is small.

At first sight it is not evident that the test statistic represents a variability measure. See Problem 7.1.e for a different and more intuitive form of the statistic. If there are ties then the test statistic needs to be adjusted because ties affect the variance in the groups. The form of the adjustment is not intuitive and consists essentially of a larger denominator compared to $IJ(J+1)$. For very small values of J and I, critical values are tabulated in selected books (e.g., Siegel and Castellan Jr, 1988). For $J > 5$ (or for $I > 5, 8, 13$ when $J = 5, 4, 3$, respectively) an approximation based on a chi-square distribution with $J - 1$ degrees of freedom is used. Test 10 summarizes the Friedman test and Example 7.6 uses the `podo` dataset as an illustration.

Test 10: Comparing the locations of repeated measures

Question: Are the medians of J matched samples significantly different?

Assumptions: All distributions are continuous of the same shape, possibly shifted, the samples are independent.

Calculation: For each subject the ranks are calculated. For each sample, sum the ranks to get R_j. Calculate Fr_{obs} according to (7.8).

Decision: Reject H_0: "the medians are the same" if $Fr_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{J-1,1-\alpha}$.

Calculation in R: `friedman.test(y)`

In case H_0 is rejected, we have evidence that at least one of the groups has a different location compared to at least one other. The test does not say which one differs or how many are different. To check which of the group pairs differ, individual *post-hoc tests* can be performed. Typically, for the pair (r, s) the following approximation is used

$$|R_r - R_s| \geq z_{\alpha/(J(J-1))}\sqrt{\frac{IJ(J+1)}{6}} \tag{7.9}$$

where R_s is as above. Note that we divide the level α of the quantile by $J(J-1) = \binom{J}{2}$ which corresponds to the classical Bonferroni correction, Section 5.5.2.

Example 7.6 (continuation of Example 5.2)**.** We consider all four visits of the dataset. Note that one rabbit was not measured in the third visit (ID *15* in visit called *5*) and thus the remaining three measures have to be eliminated. The data can be arranged in a matrix form, here of dimension 16 × 4. R-Code shows that there is no significant difference between the visits, even if we would further stratify according to the barn. ♣

R-Code 7.5 Friedman test, *pododermatitis* (see Example 7.6 and Test 10)

```
table(podo$ID)
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
##  4  4  4  4  4  4  4  4  4  4  4  4  4  4  3  4  4
podoCompl <- podo[podo$ID !=15,]  # different ways to eliminate
PDHmean4 <- matrix(podoCompl$PDHmean[order(podoCompl$ID)], ncol=4, byrow=TRUE)
colMeans(PDHmean4)                # means of the four visits
## [1] 3.7500 3.4484 3.8172 3.8531
friedman.test( PDHmean4)          # no post-hoc tests necessary.
##
##  Friedman rank sum test
##
## data:  PDHmean4
## Friedman chi-squared = 2.62, df = 3, p-value = 0.45
## further stratification does not change the result
# podoCompl <- podoCompl[podoCompl$Barn==2,] # Barn 2 only, 6 animals
```

Often the Friedman test is also called a Friedman two-way ANOVA by ranks. Two-way in the sense that there is a symmetry with respect to groups and subjects.

7.3.2 Kruskal–Wallis Test

The Kruskal–Wallis test compares I different groups, with possibly different group sizes n_j and thus represents the extension of the Wilcoxon–Mann–Withney test to arbitrary number of groups. We start by calculating the ranks for all observations. The test statistic calculates the average of the ranks in each group and compares it to the overall average. Similar as in a variance estimate, we square the difference and additionally weight the difference by the number of observations in the group. Formally, we have

$$KW_{\text{obs}} = \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j \big(\overline{R}_j - \overline{R}\big)^2, \tag{7.10}$$

where n is the total number of observations, $\overline{R}_j$ the average of the ranks in group j and $\overline{R}$ the overall mean of the ranks, i.e., $(n+1)/2$. The additional denominator is such that for $J > 3$ and $I > 5$, the Kruskal–Wallis test statistics is approximately distributed as a chi-square distribution with $J - 1$ degrees of freedom.

Similar comments as for the Friedman test hold: for very small datasets critical values are tabulated; in case of ties the test statistic needs to be adjusted; if the null hypothesis is rejected post-hoc tests can be performed.

Test 11: Comparing the locations of several independent samples

Question: Are the medians of several dependent samples significantly different?

Assumptions: All distributions are continuous of the same shape (possibly shifted), the samples are independent.

Calculation: Calculate the rank of the observations among the entire dataset. For $\overline{R}_j$ the average of the ranks in group j and $\overline{R} = (n+1)/2$, calculate KW_{obs} according to (7.10).

Decision: Reject H_0: “the medians are the same” if $KW_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{J-1,1-\alpha}$.

Calculation in R: `Kruskal.test(x, g)`

7.3.3 Family-Wise Error Rate

In this short paragraph we give an alternative view on correcting for multiple testing. A Bonferoni correction as used in the post-hoc tests (see Section 5.5.2) guarantees the specified Type I error but may not be ideal as discussed now. Suppose that we perform several test, with resulting counts denoted as shown in Table 7.1. The *family-wise error rate* (FWER) is the probability of making at least one Type I error in the family of tests FWER $= \mathrm{P}(V \geq 1)$. Hence, when performing the tests such that FWER $\leq \alpha$ we keep the probability of making one or more Type I errors in the family at or below level α.

The Bonferroni correction, where we replace the level α of the test with $\alpha_{\text{new}} = \alpha/m$, maintains the FWER. Another popular correction method is the Holm–Bonferroni procedure for which we order the p-values (lowest to largest) and the associated hypothesis. We successively test if the k^{th} p-value is smaller than $\alpha/(m-k+1)$. If so we reject the k^{th} hypothesis and move

Table 7.1: Typical notation for multiple testing.

	H_0 is true	H_1 is true	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

on to $k + 1$, if not we stop the procedure. The Holm–Bonferroni procedure also maintains the FWER but has (uniformly) higher power than the simple Bonferroni correction.

Remark 7.3. In bioinformatics and related fields m may be thousands and much larger. In such settings, a FWER procedure may be too stringent as many effects may be missed due to the very low thresholds. An alternative correction is to control the *false discovery rate* (FDR) at level q with $\text{FDR} = \text{E}(V/R)$. FDR procedures have greater power at the cost of increased rates of type I errors.

If all of the null hypotheses are true ($m_0 = m$), then controlling the FDR at level q also controls the FWER

$$\text{FWER} = \text{P}(V \geq 1) = \text{E}(V/R) = \text{FDR} \leq q, \tag{7.11}$$

because the event $\{V \geq 1\}$ is identical to $\{V/R = 1\}$ (for $V = 0$ we make appropriate definitions of the events). However, if $m_0 < m$ it holds $\text{FWER} \geq \text{FDR}$. ♣

7.4 Permutation Tests

We conclude the chapter with additional tests, relying on few statistical assumptions and rather being a toolbox to construct arbitrary tests. The general idea of permutation tests is to "reassign" the group or treatment to the observations and recalculate the test statistic. Under the null hypothesis, the observed test statistic will be similar compared to the ones obtained by reassigning the groups. The resulting p-value is the proportion of cases that the permutation yielded a more extreme observation than the data.

Permutation tests are straightforward to implement manually and thus are often used in settings where the distribution of the test statistic is complex or even unknown. Two classical examples are given in Tests 12 and 13. Permutation tests assume that under the null hypothesis, the underlying distributions being compared are the same and exchangeable. For large samples, the test may be computationally demanding.

Example 7.7 (continuation of Examples 5.2, 7.4 and 7.5)**.** R-Code 7.6 illustrates an alternative to the Wilcoxon signed rank test and the Wilcoxon–Mann–Whitney test and compares the difference in locations between the two visits and the two barns respectively. For the paired setting, see also Problem 7.4. Without too much surprise, the resulting p-values are in the same range as seen in Examples 7.4 and 7.5. Note that the function `oneway_test()` requires a formula input. ♣

Test 12: Comparing the locations of two paired samples

Question: Are the locations of two paired samples significantly different?

Assumptions: The null hypothesis is formulated, such that under H_0 the groups are exchangeable.

Calculation: (1) Calculate the mean t_{obs} of the differences $d_1, \dots, d_n$.

(2) Multiply each difference d_i with -1 with probability $p = 1/2$.

(3) Calculate the mean of differences.

(4) Repeat this procedure R times (R large).

Decision: Compare the selected significance level with the p-value:

$$\frac{1}{R}(\text{number of permuted differences more extreme than } t_{\text{obs}})$$

Calculation in R: `require(exactRankTests); perm.test(x, y, paired=TRUE)`

Test 13: Comparing the locations of two independent samples

Question: Are the locations of two independent samples significantly different?

Assumptions: The null hypothesis is formulated, such that under H_0 the groups are exchangeable.

Calculation: (1) Calculate the difference t_{obs} in the means of the two groups to be compared (m observations in group 1, n observations in group 2).

(2) Form a random permutation of the values of both groups by randomly allocating the observed values to the two groups (m observations in group 1, n observations in group 2).

(3) Calculate the difference in means of the two new groups.

(4) Repeat this procedure R times (R large).

Decision: Compare the selected significance level with the p-value:

$$\frac{1}{R}(\text{number of permuted differences more extreme than } t_{\text{obs}})$$

Calculation in R: `require(coin); oneway_test( formula, data)`

R-Code 7.6 Permutation test comparing means of two samples.

```
require(exactRankTests)
perm.test( PDHmean2[,1],  PDHmean2[,2], paired=TRUE)
##
##  1-sample Permutation Test (scores mapped into 1:m using rounded
##  scores)
##
## data:  PDHmean2[, 1] and PDHmean2[, 2]
## T = 28, p-value = 0.45
## alternative hypothesis: true mu is not equal to 0
require(coin)
oneway_test( PDHmean ~ as.factor(Barn), data=podo)
##
##  Asymptotic Two-Sample Fisher-Pitman Permutation Test
##
## data:  PDHmean by as.factor(Barn) (1, 2)
## Z = 1.24, p-value = 0.21
## alternative hypothesis: true mu is not equal to 0
```

Note that the package *exactRankTests* will no longer be further developed. The package *coin* is the successor of the latter and includes extensions to other rank-based tests.

Remark 7.4. The two tests presented in this section are often refered to as randomization tests. There is a subtle difference between permutation tests and randomization tests. For simplicity we use the term permutation test only and refer to Edgington and Onghena (2007) for an indepth discussion. ♣

7.5 Bibliographic Remarks

There are many ("elderly") books about robust statistics and rank test. Siegel and Castellan Jr (1988) was a very accessible classic for many decades. The book by Hollander and Wolfe (1999) treats the topic of rank based methods in much more details and depth.

The classical theory of robust statistics is summarized by Huber (1981); Hampel *et al.* (1986).

7.6 Exercises and Problems

Problem 7.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Justify (intuitively) the breakdown points stated in Remark 7.1.

b) Determine the proportion of observations which would be marked on average as outliers when the data is from a (i) Gaussian distribution, (ii) from a t-distribution with ν degrees of freedom, (iii) from an exponential distribution with rate λ.

c) Show that under the null hypothesis, $\mathrm{E}(W_{\text{obs}}) = n_\star(n_\star + 1)/4$ and $\mathrm{Var}(W_{\text{obs}}) = n_\star(n_\star + 1)(2n_\star + 1)/24$.

Hint: Write $W_{\text{obs}} = \sum_{k=1}^{n_\star} kI_k$ where I_k is a binomial random variable with $p = 1/2$ under H_0. Hence, $\mathrm{E}(I_k) = 1/2$ and $\mathrm{Var}(I_k) = 1/4$.

d) Show that under the null hypothesis, $\mathrm{E}(U_{\text{obs}}) = n_x n_y/2$ and $\mathrm{Var}(U_{\text{obs}}) = n_x n_y(n_x + n_y + 1)/12$.

Hints: Use the result of Problem 2.5.**c**. If we draw without replacement n_y elements from a set of $N = n_x + n_y$ iid random variables, each having mean μ and variance σ^2, then the mean and variance of the sample is μ and $\frac{\sigma^2}{n_y}\Bigl(1 - \frac{n_y - 1}{n_x + n_y - 1}\Bigr)$. That means that the variance needs to be adjusted because of the finite population size $N = n_x + n_y$, see also Problem 4.

e) Show that the Friedman test statistic Fr given in equation (7.8) can be written in the "variance form"

$$Fr = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \bigl(R_j - \overline{R}\bigr)^2,$$

where $\overline{R}$ is the average of R_j.

Hint: start showing $\overline{R} = I(J + 1)/2$

Problem 7.2 (Robust estimates) For the `mtcars` dataset (available by default through the package `datasets`), summarize location and scale for the milage, number of cylinders and the weight with standard and robust estimates. Compare the results.

Problem 7.3 (Weight changes over the years) We consider `palmerpenguins` and assess if there is a weight change of the penguins over the different years (due to exceptionally harsh environmental conditions in the corresponding years).

a) Is there a significant weight change for Adelie between 2008 and 2009?

b) Is there evidence of change for any of the species? Why would a two-by-two comparison be suboptimal?

Problem 7.4 (Permutation test for paired samples) Test if there is a significant change in pododermatitis between the first and last visit using a manual implementation of a permutation test. Compare the result to Example 7.4.

Problem 7.5 (Rank and permutation tests) Download the `water_transfer.csv` data from the course web page and read it into R. The dataset describes tritiated water diffusion across human chorioamnion and is taken from Hollander and Wolfe (1999, Table 4.1, page 110). The `pd` values for age `"At term"` and `"12-26 Weeks"` are denoted with y_A and y_B, respectively. We will statistically test if the water diffusion is different at the two time points. That means we test whether there is a shift in the distribution of the second group compared to the first.

a) Use a Wilcoxon–Mann–Whitney test to test for a shift in the groups. Interpret the results.

b) Now, use a permutation test as implemented by the function `wilcox_test()` from R package `coin` to test for a potential shift. Compare to **a**).

c) Under the null hypothesis, we are allowed to permute the observations (all y-values) while keeping the group assignments fix. Keeping this in mind, we will now manually construct a permutation test to detect a potential shift. Write an R function `perm_test()` that implements a two-sample permutation test and returns the p-value. Your function should execute the following steps.

- Compute the test statistic $t_{\text{obs}} = \widetilde{y}_A - \widetilde{y}_B$, where $\widetilde{\cdot}$ denotes the empirical median.
- Then repeat many times (e.g., $R = 1000$)
  - Randomly assign all the values of `pd` to two groups x_A and x_B of the same size as y_A and y_B.
  - Store the test statistic $t_{\text{sim}} = \widetilde{x}_A - \widetilde{x}_B$.
- Return the two-sided p-value, i.e., the number of permuted test statistics t_{sim} which are smaller or equal than $-|t_{\text{obs}}|$ or larger or equal than $|t_{\text{obs}}|$ divided by the total number of permutations (in our case $R = 1000$).

Problem 7.6 (Comparison of power) In this problem we compare the power of the one sample t-test to the Wilcoxon signed rank test with a simulation study.

a) We assume that $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$, where $n = 15$ and $\sigma^2 = 1$. For $\mu_1 = 0$ to 1.2 in steps of 0.05, simulate $R = 1000$ times a sample and based on these replicates, estimate the power for a one sample t-test and a Wilcoxon signed rank test (at level $\alpha = 0.1$).

b) Redo for $Y_i = \sqrt{(m-2)/m}V_i + \mu_1$, where $V_1, \dots, V_n \overset{\text{iid}}{\sim} t_m$, with $m = 4$. Interpret.

c) Redo for $Y_i = (X_i - m)/\sqrt{2m} + \mu_1$, where $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{X}^2_m$, with $m = 10$. Why is the Wilcoxon signed rank test having an apparent lower power?

Problem 7.7 (Power of the sign test) In this problem we will analyze the power of the sign test in two situations and compare it to the Wilcoxon signed rank test. To emphasize the effects, we assume a large sample size $n = 100$.

a) We assume that $d_1, \dots, d_{100}$ are the differences of two paired samples. Simulate the d_i according to a normal random variables (with variance one) and different means μ_1 (e.g., a fine sequence from 0 to 1 `seq(0, to=1, by=.1)`). For the sign test and the Wilcoxon signed rank test, calculate the proportion of rejected cases out of 500 of the test H_0 : “median is zero” and visualize the resulting empirical power curves. What are the powers for $\mu_1 = .2$? Interpret the power for $\mu_1 = 0$.

b) We now look at the Type I error if we have a skewed distribution and we use the chi-squared distribution with varying degrees of freedom. For very large values thereof, the distribution is almost symmetric, for very small values we have a pronounced asymmetry. Note that the median of a chi-square distribution with ν degrees of freedom is approximately $\nu(1-2/(9\nu))^3$.

We assume that we have a sample $d_1, \dots, d_{100}$ from a shifted chi-square distribution with ν degrees of freedom and median one, e.g., from `rchisq(100, df=nu)-nu*(1-2/(9*nu))^3`. As a function of ν, $\nu = 2, 14, 26, \dots, 50$, (`seq(2, to=50, by=12)`) calculate the proportion of rejected cases out of 500 tests for the test H_0 : "median is zero". Calculate the resulting empirical power curve and interpret.

c) Similar to **a**) we calculate the power for a shifted chi-squared distribution with $\nu = 2$. To simplify, simulate first `rchisq(100, df=2)-2*(1-2/(9*2))^3` and then shift by the same amount as done in **a**). Calculate the resulting empirical power curve and interpret. Why should the resulting power curves of the sign test not be compared with the power curve of **a**)?

Problem 7.8 (Anorexia treatments) Anorexia is an eating disorder that is characterized by low weight, food restriction, fear of gaining weight and a strong desire to be thin. The dataset `anorexia` in the package `MASS` gives the weight in pounds of 72 females before and after a treatment, consisting of control, cognitive behavioral treatment and family treatment. Visualize the data wisely and test for effectiveness of the treatments. If there are several possibilities, discuss the appropriateness of each.

Problem 7.9 (Water drinking preference) Pachel and Neilson (2010) conducted a pilot study to assess if house cats prefer water if provided still or flowing. Their experiment consisted of providing nine cats either form of water over 4 days. The data provided in mL consumption over the 22h hour period is provided in the file `cat.csv`. Some measurement had to be excluded due to detectable water spillage. Assess with a suitable statistical approach if the cats prefer flowing water over still water.

Problem 7.10 (BMJ Endgame) Discuss and justify the statements about 'Parametric v non-parametric statistical tests' given in doi.org/10.1136/bmj.e1753.

Chapter 8

Multivariate Normal Distribution

Learning goals for this chapter:

- ⋄ Describe a random vector, cdf, pdf of a random vector and its properties
- ⋄ Give the definition and intuition of E, Var and Cov for a random vector
- ⋄ Know basic properties of E, Var and Cov for a random vector
- ⋄ Recognize the density of Gaussian random vector and know properties of Gaussian random vector
- ⋄ Explain and work with conditional and marginal distributions.
- ⋄ Estimate the mean and the variance of the multivariate distribution.
- ⋄ Explain the relationship between the eigenvalues and eigenvector of the covariance matrix and the shape of the density function.

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter08.R.

In Chapter 2 we have introduced univariate random variables and in Chapter 3 random samples. We now extend the framework to random vectors (i.e., multivariate random variables) where the individual random variables are not necessarily independent (see Definition 3.1). Within the scope of this book, we can only cover a tiny part of a beautiful theory. We are pragmatic and discuss what will be needed in the sequel. Hence, we discuss a discrete case as a motivating example only and then focus on continuous random vectors, especially Gaussian random vectors. In this chapter we cover the theoretical details, the next one focuses on estimation.

8.1 Random Vectors

A random vector is a (column) vector $\mathbf{X} = (X_1, \dots, X_p)^\top$ with p random variables as components. The following definition is the generalization of the univariate cumulative distribution function (cdf) to the multivariate setting (compare with Definition 2.1).

Definition 8.1. The multivariate (or multidimensional) distribution function of a random vector $\mathbf{X}$ is defined as

$$F_{\mathbf{X}}(\boldsymbol{x}) = \mathrm{P}(\mathbf{X} \leq \boldsymbol{x}) = \mathrm{P}(X_1 \leq x_1, \ldots, X_p \leq x_p), \tag{8.1}$$

where the list in the right-hand-side is to be understood as intersections ($\cap$) of the p events. $\diamondsuit$

The multivariate distribution function generally contains more information than the set of marginal distribution functions $\mathrm{P}(X_i \leq x_i)$, because (8.1) only simplifies to $F_{\mathbf{X}}(\boldsymbol{x}) = \prod\limits_{i=1}^{p} \mathrm{P}(X_i \leq x_i)$ under independence of all random variables X_i (compare to Equation (3.3)).

Example 8.1. Suppose we toss two fair four-sided dice with face values 0, 1, 2 and 3. The first player marks the absolute difference between both and the second the maximum value. To cast the situation in a probabilistic framework, we introduce the random variables T_1 and T_2 for the result of the tetrahedra and assume that each value appears with probability $1/4$. (As a side note, tetrahedron die have often marked corners instead of faces. For the later, appearance means the face that is turned down.) Table 8.1 gives the frequency table for $X = |T_1 - T_2|$ and $Y = \max(T_1, T_2)$. The entries are to be interpreted as the *joint pdf* $f_{X,Y}(x, y) = \mathrm{P}(X = x, Y = y)$. From the joint pdf, we can derive various probabilities, for example $\mathrm{P}(X \geq Y) = 9/16$, $\mathrm{P}(X \geq 2 \mid Y \neq 3) = 1/8$. It is also possible to obtain the marginal pdf of, say, X: $f_X(x) = \sum_{y=0}^{3} f_{X,Y}(x, y)$, also summarized by Table 8.1. More specifically, from the joint we can find the marignal quantities, but not the other way around. $\clubsuit$

Table 8.1: Probability table for X and Y of Example 8.1. Last column and last row give the marginal probabilities of X and Y, respectively.

$X \backslash Y$	0	1	2	3	
0	1/16	1/16	1/16	1/16	1/4
1	0	1/8	1/8	1/8	3/8
2	0	0	1/8	1/8	1/4
3	0	0	0	1/8	1/8
	1/16	3/16	5/16	7/16	1

8.1.1 Density Function of Continuous Random Vectors

A random vector $\mathbf{X}$ is a continuous random vector if each component X_i is a continuous random variable. The probability density function for a continuous random vector is defined in a similar manner as for univariate random variables.

Definition 8.2. The probability density function (or density function, pdf) $f_{\mathbf{X}}(\boldsymbol{x})$ of a p-dimensional continuous random vector $\mathbf{X}$ is defined by

$$\mathrm{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \qquad \text{for all } A \subset \mathbb{R}^p. \tag{8.2}$$

$\diamondsuit$

For convenience, we summarize here a few properties of random vectors with two continuous components, i.e., for a bivariate random vector $(X, Y)^\top$. The univariate counterparts are stated in Properties 2.1 and 2.3. The properties are illustrated with a subsequent extensive example.

Property 8.1. *Let $(X, Y)^\top$ be a bivariate continuous random vector with joint density function $f_{X,Y}(x, y)$ and joint distribution function $F_{X,Y}(x, y)$.*

1. *The distribution function is monotonically increasing:*
   for $x_1 \leq x_2$ and $y_1 \leq y_2$, $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$.
2. *The distribution function is normalized:*
   $\lim\limits_{x,y\nearrow\infty} F_{X,Y}(x, y) = F_{X,Y}(\infty, \infty) = 1.$
   (We use the slight abuse of notation by writing ∞ in arguments without a limit.)
   $F_{X,Y}(-\infty, -\infty) = F_{X,Y}(x, -\infty) = F_{X,Y}(-\infty, y) = 0.$
3. *$F_{X,Y}(x, y)$ and $f_{X,Y}(x, y)$ are continuous (almost) everywhere.*
4. $f_{X,Y}(x, y) = \dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$
5. $$\begin{aligned}\mathrm{P}(a < X \leq b, c < Y \leq d) &= \int_a^b \int_c^d f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x \\ &= F_{X,Y}(b, d) - F_{X,Y}(b, c) - F_{X,Y}(a, d) + F_{X,Y}(a, c).\end{aligned}$$
6. *Marginal distributions:*
   $F_X(x) = \mathrm{P}(X \leq x, Y \text{ arbitrary}) = F_{X,Y}(x, \infty)$ *and* $F_Y(y) = F_{X,Y}(\infty, y)$;
7. *Marginal densities:*
   $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\,\mathrm{d}y$ *and* $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\,\mathrm{d}x.$

The last two points of Property 8.1 refer to marginalization, i.e., reduce a higher-dimensional random vector to a lower dimensional one. Intuitively, we “neglect” components of the random vector in allowing them to take any value.

Example 8.2. The joint distribution of $(X, Y)^\top$ is given by $F_{X,Y}(x, y) = y - \big(1 - \exp(-xy)\big)/x$, for $x \geq 0$ and $0 \leq y \leq 1$. Top left panel of Figure 8.1 illustrates the joint cdf and the mononicity of the joint cdf is nicely visible. The top right panel illustrates that the joint cdf is normalized: for all values $x = 0$ or $y = 0$ the joint cdf is also zero and it reaches one for $x \to \infty$ and $y = 1$. The joint cdf is defined for the entire plane (x, y), if any of the two variables is negative, the joint cdf is zero. For values $y > 1$ we have $F_{X,Y}(x, y) = F_{X,Y}(x, 1)$. That means, that the joint cdf outside the domain $[0, \infty[\times [0, 1]$ is determined by its properties (here, for rotational simplicity, we only state the functions in that specific domain).

The joint density is given by $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = \frac{\partial}{\partial x} 1 - \exp(-xy) = \exp(-xy)y$ and shown in the top right panel of Figure 8.1. The joint density has a particular simple form but it cannot be factored into the marginal cdf and marginal densities:

$$F_Y(y) = F_{X,Y}(\infty, y) = y, \qquad F_X(x) = F_{X,Y}(x, 1) = F_{X,Y}(x, \infty) = 1 - (1 - \mathrm{e}^{-x})/x, \tag{8.3}$$

$$f_Y(y) = \int_0^\infty \mathrm{e}^{-xy}\, y\,\mathrm{d}x = 1, \qquad f_X(x) = \int_0^1 \mathrm{e}^{-xy}\, y\,\mathrm{d}y = \frac{1 - \mathrm{e}^{-x}(1 + x)}{x^2}. \tag{8.4}$$

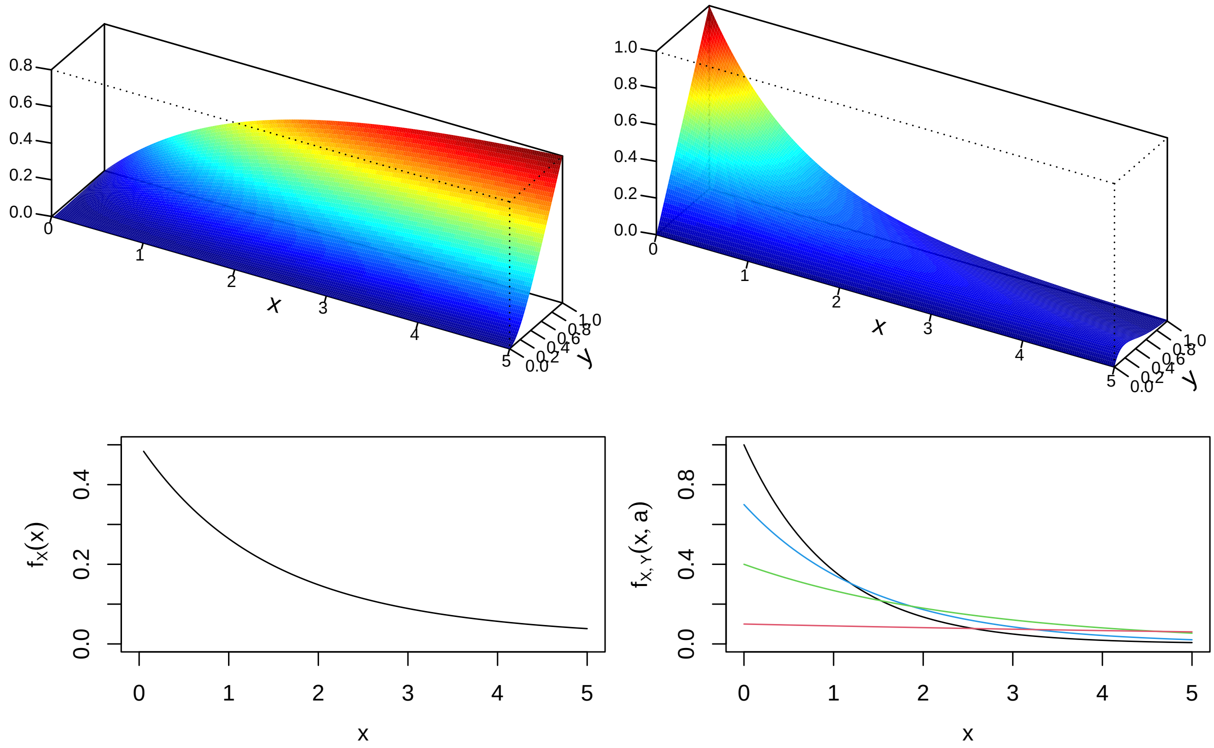


Figure 8.1: Top row: joint cdf and joint density function for case discussed in Example 8.2. Bottom row: marginal density of X and joint density evaluated at $a = 1, .7, .4, .1$ (in black, blue, green and red).

(Of course, we could have differentiated the expressions in the first line to get the second line or integrated the expressions in the second line to get the ones in the first line, due to Property 2.3.) The marginal distribution of Y is $\mathcal{U}(0, 1)$! Bottom left panel of Figure 8.1 gives the marginal density of X. The product of the two marginal densities is not equal to the joint one. Hence, X and Y are not independent.

For coordinate aligned rectangular domains, we use the joint cdf to calculate probabilities, for example

$$\begin{aligned}
\mathrm{P}(1 < X \leq 2, Y \leq 1/2) &= F_{X,Y}(2, 1/2) - F_{X,Y}(2, 0) - F_{X,Y}(1, 1/2) + F_{X,Y}(1, 0) \\
&= 1/2 - (1 - \mathrm{e}^{-2\cdot 1/2})/2 - 0 - 1/2 - (1 - \mathrm{e}^{-1\cdot 1/2})/1 - 0 = \frac{(\sqrt{\mathrm{e}} - 1)^2}{2\,\mathrm{e}} \approx 0.077.
\end{aligned} \tag{8.5}$$

and via integration for other cases

$$\mathrm{P}(XY \leq 1) = F_{X,Y}(1, 1) + \int_0^1 \int_1^{1/y} \mathrm{e}^{-xy}\, y \,\mathrm{d}x \,\mathrm{d}y = \mathrm{e}^{-1} + \int_0^1 \mathrm{e}^{-y} - \mathrm{e}^{-1} \,\mathrm{d}y = \frac{\mathrm{e} - 1}{\mathrm{e}} \approx 0.63. \tag{8.6}$$

♣

8.1.2 Mean and Variance of Continuous Random Vectors

We now characterize the first two moments of random vectors.

Definition 8.3. The expected value of a random vector $\mathbf{X}$ is defined as

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix}. \tag{8.7}$$

◇

Hence the expectation of a random vector is simply the vector of the individual expectations. Of course, to calculate these, we only need the marginal univariate densities $f_{X_i}(x)$ and thus the expectation does not change whether (8.1) can be factored or not. The expectation of products of random variables is defined as

$$\mathrm{E}(X_1 X_2) = \int\int x_1 x_2 f(x_1, x_2)\,\mathsf{d}x_1\,\mathsf{d}x_2 \tag{8.8}$$

(for continuous random variables). The variance of a random vector requires a bit more thought and we first need the following.

Definition 8.4. The *covariance* between two arbitrary random variables X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\big((X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2))\big) = \mathrm{E}(X_1 X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2). \tag{8.9}$$

◇

In case of two independent random variables, the their joint density can be factored and Equation (8.8) shows that $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\,\mathrm{E}(X_2)$ and thus their covariance is zero. The inverse, however, is in general not true.

Using the linearity property of the expectation operator, it is possible to show the following handy properties.

Property 8.2. *We have for arbitrary random variables X_1, X_2 and X_3:*

1. $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$,
2. $\mathrm{Cov}(X_1, X_1) = \mathrm{Var}(X_1)$,
3. $\mathrm{Cov}(a + bX_1, c + dX_2) = bd\,\mathrm{Cov}(X_1, X_2)$, *for arbitrary values a, b, c and d,*
4. $\mathrm{Cov}(X_1, X_2 + X_3) = \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3)$.

It is important to note that the covariance describes the *linear* relationship between the random variables.

Definition 8.5. The variance of a p-variate random vector $\mathbf{X} = (X_1, \dots, X_p)^\top$ is defined as

$$\mathrm{Var}(\mathbf{X}) = \mathrm{E}\big((\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^\top\big) \tag{8.10}$$

$$= \mathrm{Var}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \mathrm{Var}(X_1) & \dots & \mathrm{Cov}(X_i, X_j) \\ & \ddots & \\ \mathrm{Cov}(X_j, X_i) & \dots & \mathrm{Var}(X_p) \end{pmatrix}, \tag{8.11}$$

called the *covariance matrix* or *variance–covariance* matrix. ◇

The covariance matrix is a symmetric matrix and – except for degenerate cases – a positive definite matrix. We will not consider degenerate cases and thus we can assume that the inverse of the matrix $\operatorname{Var}(\mathbf{X})$ exists and is called the precision.

Similar to Properties 2.5, we have the following properties for random vectors.

4 min **Property 8.3.** *For an arbitrary p-variate random vector $\mathbf{X}$, (fixed) vector $\boldsymbol{a} \in \mathbb{R}^q$ and matrix $\mathbf{B} \in \mathbb{R}^{q\times p}$ it holds:*

1. $\operatorname{Var}(\mathbf{X}) = \operatorname{E}(\mathbf{X}\mathbf{X}^\top) - \operatorname{E}(\mathbf{X})\operatorname{E}(\mathbf{X})^\top$,
2. $\operatorname{E}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \boldsymbol{a} + \mathbf{B}\operatorname{E}(\mathbf{X})$,
3. $\operatorname{Var}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\operatorname{Var}(\mathbf{X})\mathbf{B}^\top$.

The covariance itself cannot be interpreted as it can take arbitrary values. The *correlation* between two random variables X_1 and X_2 is defined as

5 min
$$\operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}} \tag{8.12}$$

and corresponds to the *normalized* covariance. It holds that $-1 \leq \operatorname{Corr}(X_1, X_2) \leq 1$, with equality only in the *degenerate* case $X_2 = a + bX_1$ for some a and $b \neq 0$.

8.2 Conditional Densities

In this short section we revisit conditional probabilities and conditional densities with particular focus on bivariate continuous random variables.

We recall the formula for conditional probablity $\operatorname{P}(A \mid B) = \operatorname{P}(A \cap B)/\operatorname{P}(B)$, which can be translated directly to the conditional cdf

$$F_{X|Y}(x \mid y) = \frac{F_{X,Y}(x, y)}{F_Y(y)} \tag{8.13}$$

for $A = \{X \leq x\}$ and $B = \{Y \leq y\}$. The events A and B can be much more general, for example $A = \{X = 2\}$ and $B = \{Y = 2\}$. In a setting like Example 8.1, it is still possible to calculate $\operatorname{P}(X = 2 \mid Y = 2) = \operatorname{P}(X = 2, Y = 2)/\operatorname{P}(Y = 2) = (1/8)/(5/16) = 2/5$. However, in a setting like Example 8.2, an expression like $\operatorname{P}(Y = a)$ is zero and we are - seemingly - stuck.

Definition 8.6. Let $(X, Y)^\top$ be a bivarite continuous random vector with joint density function $f_{X,Y}(x, y)$. The conditional density of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \tag{8.14}$$

whenever $f_Y(y) > 0$ and zero otherwise. ♢

This definition has nevertheless some intuition. Consider a neighborhood of x and y

$$\begin{aligned} \operatorname{P}(x \leq X \leq x + \delta \mid y \leq Y \leq y + \epsilon) &= \frac{\int_x^{x+\delta}\int_y^{y+\epsilon} f_{X,Y}(x, y)\,\mathsf{d}y\,\mathsf{d}x}{\int_y^{y+\epsilon} f_Y(y)\,\mathsf{d}y} & (8.15)\\ &\approx \frac{\delta\epsilon f_{X,Y}(x, y)}{\epsilon f_Y(y)} = \delta f_{X|Y}(x \mid y). & (8.16) \end{aligned}$$

That means that the conditional probability of $X \mid Y = y$ in a neighborhood of (x, y) is proportional to the conditional density.

Visually, the conditional density is like slicing the joint at particular values (of x or y) and renormalizing the resulting curve to get a proper density.

Conditional densities are proper densities and thus it is possible to calculate the conditional expectation $\operatorname{E}(X \mid Y)$, conditional variance $\operatorname{Var}(X \mid Y)$ and so forth.

Example 8.3. In the setting of Example 8.2 the marginal density of Y is constant and thus $f_{X|Y}(x \mid y) = f_{X,Y}(x, y)$. The curves in the lower right panel of Figure 8.2 are actual densities: the conditional density $f_{X|Y}(x \mid y)$ for $y = 1, 0.7, 0.4$ and 0.1.

The conditional expectation of $X \mid Y = y$ is $\operatorname{E}(X \mid Y) = \int_0^\infty x\,\mathrm{e}^{-xy}\,y\,\mathsf{d}x = 1/y$. ♣

8.3 Multivariate Normal Distribution

We now consider a special multivariate distribution: the multivariate normal distribution, by first considering the bivariate case.

8.3.1 Bivariate Normal Distribution

We introduce the bivariate normal distribution by specifying its density, whose expression is quite intimidating and is for reference here.

Definition 8.7. The random variable pair (X, Y) has a bivariate normal distribution if

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x, y)\,\mathsf{d}x\,\mathsf{d}y \tag{8.17}$$

with density

$$\begin{aligned} f(x, y) &= f_{X,Y}(x, y) \qquad (8.18) \\ &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right), \end{aligned}$$

for all x and y and where $\mu_x \in \mathbb{R}$, $\mu_y \in \mathbb{R}$, $\sigma_x > 0$, $\sigma_y > 0$ and $-1 < \rho < 1$. ♢

The role of some of the parameters μ_x, μ_y, σ_x, σ_y and ρ might be guessed. We will discuss their precise meaning after the following example. Note that the joint cdf does not have a closed form and essentially all probability calculations involve some sort of numerical integration scheme of the density.

Example 8.4. Figure 8.2 (based on R-Code 8.1) shows the density of a bivariate normal distribution with $\mu_x = \mu_y = 0$, $\sigma_x = 1$, $\sigma_y = \sqrt{5}$, and $\rho = 2/\sqrt{5} \approx 0.9$. Because of the quadratic form in (8.18), the contour lines (isolines) are ellipses.

The joint cdf is harder to interpret, it is almost impossible to infer the shape an orientation of an isoline of the density. Since (x, y) can take any values in the plane, the joint cdf is strictly positive and the value zero is only reached at $\lim_{x,y \searrow -\infty} F(x, y) = 0$.

Several R packages implement the bivariate/multivariate normal distribution. We recommend the package `mvtnorm`. ♣

R-Code 8.1: Density of a bivariate normal random vector. (See Figure 8.2.)

```
require( mvtnorm)    # providing dmvnorm, pmvnorm
require( fields)     # providing tim.colors() and image.plot()
Sigma <- array( c(1,2,2,5), c(2,2))
x <- y <- seq( -3, to=3, length=100)
grid <- expand.grid( x=x, y=y)
densgrid <- dmvnorm( grid, mean=c(0, 0), sigma=Sigma) # zero mean
jdensity <- array( densgrid, c(100, 100))
image.plot(x, y, jdensity, col=tim.colors())          # left panel
faccol <- fields::tim.colors()[cut(jdensity[-1,-1],64)]
persp(x, y, jdensity, col=faccol, border = NA,  zlab="", # right panel
      tick='detailed', theta=120, phi=30, r=100, zlim=c(0,0.16))
## To calculate the cdf, we need a lower and upper bound. Passing directly
cdfgrid <- apply(grid, 1, function(x) {        #  the grid is not possible
        pmvnorm( upper=x, mean=c(0, 0), sigma=Sigma) } )
jcdf <- array( cdfgrid, c(100, 100))
image.plot(x, y, jcdf, zlim=c(0,1), col=tim.colors())     # left panel
faccol <- fields::tim.colors()[cut(jcdf[-1,-1],64)]
persp(x, y, jcdf, col=faccol, border = NA, zlab="",       # right panel
      tick='detailed', theta=12, phi=50, r=100, zlim=c(0,1))
```

The bivariate normal distribution has many nice properties. A first set is as follows.

Property 8.4. *For the bivariate normal random vector as specified by* (8.18)*, we have: (i) The marginal distributions are* $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ *and* $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ *and (ii)*

$$\mathrm{E}\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \mu_x \\ \mu_y, \end{pmatrix} \qquad \mathrm{Var}\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{8.19}$$

Thus,

$$\mathrm{Cov}(X, Y) = \rho\sigma_x\sigma_y, \qquad \mathrm{Corr}(X, Y) = \rho. \tag{8.20}$$

(iii) If $\rho = 0$*,* X *and* Y *are independent and vice versa.*

Note, however, that the equivalence of independence and uncorrelatedness is specific to jointly normal variables and cannot be assumed for random variables that are not jointly normal.

Example 8.5. Figure 8.3 (based on R-Code 8.2) shows realizations from a bivariate normal distribution with zero mean and unit marginal variance for various values of correlation ρ. Even for large samples as shown here ($n = 200$), correlations between -0.25 and 0.25 are barely perceptible. ♣

```
## Loading required package:  mvtnorm
##
## Attaching package:  'mvtnorm'
## The following objects are masked from 'package:spam':
##
##     rmvnorm, rmvt
```

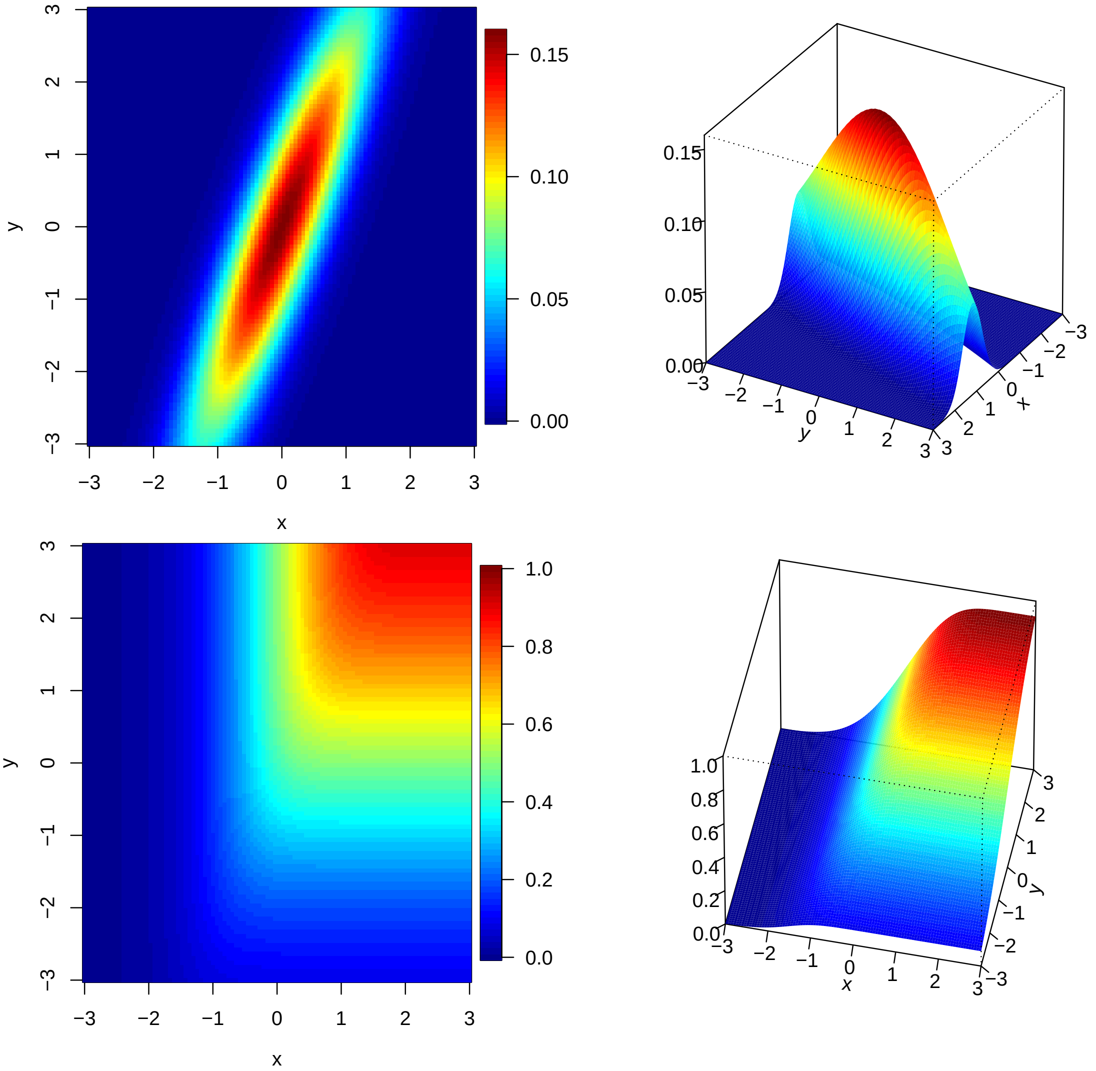


Figure 8.2: Density (top row) and distribution function (bottom row) of a bivariate normal random vector. (See R-Code 8.1.)

8.3.2 Multivariate Normal Distribution

For the general multivarate case we have to use vector notation. Surprisingly, we gain clarity even compared to the bivariate case. We again introduce the distribution through the density

Definition 8.8. The random vector $\mathbf{X} = (X_1, \dots, X_p)^\top$ is multivariate normally distributed if

$$F_{\mathbf{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_{\mathbf{X}}(x_1, \dots, x_p)\, \mathsf{d}x_1 \dots \mathsf{d}x_p \qquad (8.21)$$

R-Code 8.2 Realizations from a bivariate normal distribution for various values of ρ, termed `binorm` (See Figure 8.3.)

```
set.seed(12)
rho <- c(-.25, 0, .1, .25, .75, .9)
for (i in 1:6) {
  Sigma <- array( c(1, rho[i], rho[i], 1), c(2,2))
  sample <- rmvnorm( 200, sigma=Sigma)
  plot( sample, pch=20, xlab='', ylab='', xaxs='i', yaxs='i',
       xlim=c(-4, 4),ylim=c(-4, 4), cex=.4)
  legend( "topleft", legend=bquote(rho==.(rho[i])), bty='n')
}
```

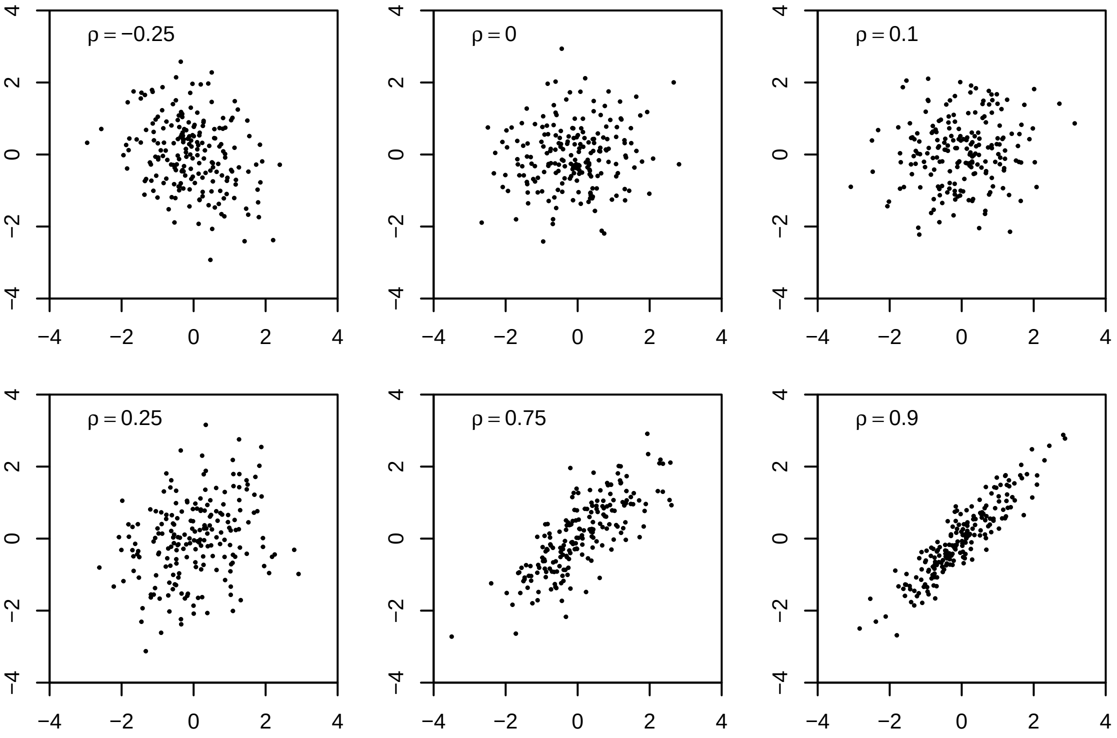


Figure 8.3: Realizations from a bivariate normal distribution with mean zero, unit marginal variance and different correlations ($n = 200$). (See R-Code 8.2.)

with density

$$f_{\mathbf{X}}(x_1, \dots, x_p) = f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big) \tag{8.22}$$

for all $\boldsymbol{x} \in \mathbb{R}^p$ (with $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive-definite $\boldsymbol{\Sigma}$). We denote this distribution with $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. ♢

The following two properties give insight into the meaning of the parameters and give the distribution of linear combinations of Gaussian random variables.

Property 8.5. *For the multivariate normal distribution* $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *we have:*

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}\,, \qquad\qquad \mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}\,. \tag{8.23}$$

Property 8.6. *Let* $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and let* $\boldsymbol{a} \in \mathbb{R}^q$, $\mathbf{B} \in \mathbb{R}^{q\times p}$, $q \leq p$, $\mathrm{rank}(\mathbf{B}) = q$, *then*

$$\boldsymbol{a} + \mathbf{B}\mathbf{X} \sim \mathcal{N}_q\big(\boldsymbol{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top\big). \tag{8.24}$$

This last property has profound consequences. First, it generalizes Property 2.7 and, second, it asserts that the one-dimensional marginal distributions are again Gaussian with $X_i \sim \mathcal{N}\big((\boldsymbol{\mu})_i, (\boldsymbol{\Sigma})_{ii}\big)$, $i = 1, \dots, p$ (by (8.24) with $\boldsymbol{a} = \mathbf{0}$ and $\mathbf{B} = (\dots, 0, 1, 0, \dots)$, a vector with zeros and a one in the ith position).

Similarly, any subset and any (non-degenerate) linear combination of random variables of $\mathbf{X}$ is again Gaussian with appropriate subset selection of the mean and covariance matrix.

5 min We now discuss how to draw realizations from an arbitrary Gaussian random vector, much in the spirit of Property 2.7. We suppose that we can efficiently draw from a standard normal distribution. Recall that $Z \sim \mathcal{N}(0, 1)$, then $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$. We will rely on Equation (8.24) but need to decompose the target covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{L} \in \mathbb{R}^{p\times p}$ such that $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}$. That means, $\mathbf{L}$ is like a "matrix square root" of $\boldsymbol{\Sigma}$.

To draw a realization $\boldsymbol{x}$ from a p-variate random vector $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, one starts with drawing p values from $Z_1, \dots, Z_p \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, and sets $\boldsymbol{z} = (z_1, \dots, z_p)^\top$. The vector is then (linearly) transformed with $\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$. Since $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$, where $\mathbf{I} \in \mathbb{R}^{p\times p}$ be the identity matrix, a square matrix which has only ones on the main diagonal and only zeros elsewhere, Property 8.6 asserts that $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^\top)$.

In practice, the Cholesky decomposition of $\boldsymbol{\Sigma}$ is often used. This factorization decomposes a symmetric positive-definite matrix into the product of a lower triangular matrix $\mathbf{L}$ and its transpose. It holds that $\det(\boldsymbol{\Sigma}) = \det(\mathbf{L})^2 = \prod_{i=1}^p (\mathbf{L})_{ii}^2$, i.e., we get the normalizing constant in (8.22) "for free".

8.3.3 Conditional Distributions

We now extend the concept of bivariate conditional densities to arbitrary dimensions for Gaussian random vectors. To do so, we separate the random vector $\mathbf{X}$ in two parts of size q and $p - q$, taking the role of X and Y in Equation (8.13), (8.14) and so forth. For simplicity we (possibly) reorder the elements and write

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{X}_1 \in \mathbb{R}^q, \quad \mathbf{X}_2 \in \mathbb{R}^{p-q}. \tag{8.25}$$

We divide the vector $\boldsymbol{\mu}$ in similarly sized components $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the matrix $\boldsymbol{\Sigma}$ in 2×2 blocks with according sizes:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right). \tag{8.26}$$

Both (multivariate) marginal distributions $\mathbf{X}_1$ and $\mathbf{X}_2$ are again normally distributed with $\mathbf{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{X}_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ (this can be seen again by Property 8.6).

Property 8.7. *If one conditions a multivariate normally distributed random vector* (8.26) *on a sub-vector, the result is itself multivariate normally distributed with*

1 min

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \boldsymbol{x}_1 \sim \mathcal{N}_{p-q}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right). \tag{8.27}$$

$\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent if* $\boldsymbol{\Sigma}_{21} = \mathbf{0}$ *and vice versa.*

Equation (8.27) is probably one of the most important formulas one encounters in statistics albeit not always in this explicit form. The result is again one of the many features of Gaussian random vectors: the distribution is closed (meaning again Gaussian) with respect to linear combinations and conditioning.

We now have a close look at Equation (8.27) and give a detailed explaination thereof. The expected value of the conditional distribution (conditional expectation) depends linearly on the value of $\boldsymbol{x}_1$, but the variance is independent of the value of $\boldsymbol{x}_1$. The conditional expectation represents an update of $\mathbf{X}_2$ through $\mathbf{X}_1 = \boldsymbol{x}_1$: the difference $\boldsymbol{x}_1 - \boldsymbol{\mu}_1$ is normalized by its variance $\boldsymbol{\Sigma}_{11}$ and scaled by the covariance $\boldsymbol{\Sigma}_{21}$.

The interpretation of the conditional distribution (8.27) is a bit easier to grasp if we use $p = 2$ with X and Y, in which $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \rho\sigma_y\sigma_x(\sigma_x^2)^{-1} = \rho\sigma_y/\sigma_x$ and $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \rho^2\sigma_y^2$, yielding

$$Y \mid X = x \sim \mathcal{N}\left(\mu_y + \rho\sigma_y\sigma_x^{-1}(x - \mu_x), \sigma_y^2 - \rho^2\sigma_y^2\right). \tag{8.28}$$

This equation is illustrated in Figure (8.4). The bivariate density is shown with blue ellipses. The vertical green density is the marginal density of Y and represent the "uncertainty" without any further information. The inclined ellipse indicates dependence between both variables. Hence, if we know $X = x$ (e.g., one of the ticks on the x-axis), we change our knowledge about the second variable. We can adjust the mean and reduce the variance. The red vertical densities are two

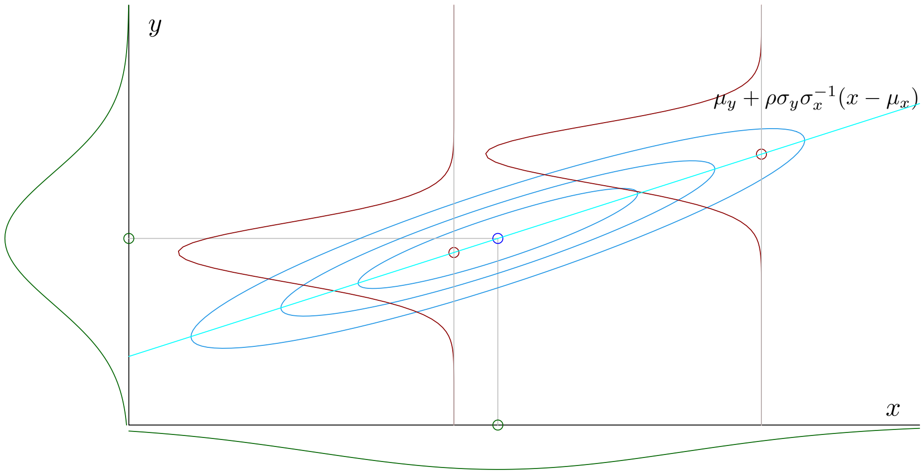


Figure 8.4: Graphical illustration of the conditional distribution of a bivariate normal random vector. Blue: bivariate density with isolines indicating quartiles, green: marginal densities, red: conditional densities. The respective means are indicated with circles. The height of the univariate densities are exaggerated by a factor of five.

examples of conditional densities. The conditional means are on the line $\mu_y + \rho\sigma_y\sigma_x^{-1}(x - \mu_x)$. The unconditional mean μ_y is corrected based the difference $x - \mu_x$, the further x from the mean of X the larger the correction. Of course, this difference needs to be normalized by the standard deviation of X, leading then to $\sigma_x^{-1}(x - \mu_x)$. The tighter the ellipses the stronger we need to correct (further multiplication with ρ) and finally, we have to scale (back) according to the variable Y (final multiplication with σ_y).

The conditional variance remains the same for all possible $X = x$ (the red vertical densities have the same variance) and can be written as $\sigma_y^2(1 - \rho^2)$ and is smaller the larger the absolute value of ρ is. As ρ is the correlation, a large correlation means a stronger linear relationship, the contour lines of the ellipses are tighter and thus the more information we have for Y. A negative ρ simply tilts the ellipses preserving the shape. Hence, similar information for the variance and a mere sign change for the mean adjustment.

This interpretation indicates that the conditional distribution and more specifically the conditional expectation plays a large role in *prediction* where one tries to predict (in the litteral sense of the word) a random variable based on an observation (see Problem 8.4).

Remark 8.1. With the background of this chapter, we are able to close a gap from Chapter 3. A slightly more detailed explanation of Property 3.4.2 is as follows. We represent the vector $(X_1, \dots, X_n)^\top$ by two components, $(X_1 - \overline{X}, \dots, X_n - \overline{X})^\top$ and $\overline{X}\mathbf{1}$, with $\mathbf{1}$ the p-vector containing only ones. Both vectors are orthogonal, thus independent. If two random variables, say Y and Z are independent, then $g(Y)$ and $h(Z)$ as well (for reasonable choices of g and h). Here, g and h are multivariate and have the particular form of $g(y_1, \dots, y_n) = \sum_{i=1}^n y_i^2/(n-1)$ $h(z_1, \dots, z_n) = z_1$. ♣

8.4 Bibliographic Remarks

Many introductory textbooks contain chapters about random vectors and multivariate Gaussian distribution. Classics are Rice (2006) and Mardia *et al.* (1979).

The online book “Matrix Cookbook” is a very good synopsis of the important formulas related to vectors and matrices (Petersen and Pedersen, 2008).

8.5 Exercises and Problems

Problem 8.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Proof Property 8.3. Specifically, let $\mathbf{X}$ be a p-variate random vector and let $\mathbf{B} \in \mathbb{R}^{q\times p}$ and $\boldsymbol{a} \in \mathbb{R}^q$ be a non-stochastic matrix and vector, respectively. Show the following.

(a) $\operatorname{E}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \boldsymbol{a} + \mathbf{B}\operatorname{E}(\mathbf{X})$,

(b) $\operatorname{Var}(\mathbf{X}) = \operatorname{E}(\mathbf{X}\mathbf{X}^\top) - \operatorname{E}(\mathbf{X})\operatorname{E}(\mathbf{X})^\top$,

(c) $\text{Var}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\,\text{Var}(\mathbf{X})\mathbf{B}^\top$.

Problem 8.2 (Bivariate normal) In this exercise we derive the bivariate density in an alternative fashion.

a) Let $X, Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ Show that the joint density is $f_{X,Y}(x,y) = 1/(2\pi)\exp(-(x^2+y^2)/2)$.

b) Define $Z = \rho X + \sqrt{1-\rho^2}Y$ Derive the distribution of Z and the pair (X, Z). Give an intuitive explainaition of the dependency between X and Z in terms of ρ.

c) Show that $Z \mid X = x \sim \mathcal{N}(\rho x, 1-\rho^2)$

Problem 8.3 (Covariance and correlation) Let $f_{X,Y}(x,y) = c \cdot \exp(-x)(x+y)$ for $0 \leq x \leq \infty$, $0 \leq y \leq 1$ and zero otherwise.

a) Show that $c = 2/3$.

b) Derive the marginal densities $f_X(x)$ and $f_Y(y)$ and their first two moments.

c) Calculate $\text{Cov}(X,Y)$ and $\text{Corr}(X,Y)$.

d) Derive the conditional densities $f_{Y|X}(y \mid x)$, $f_{X|Y}(x \mid y)$ and the conditional expectations $\text{E}(Y \mid X = x)$, $\text{E}(X \mid Y = y)$. Give an interpretation thereof.

Problem 8.4 (Prediction) According to https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit.assetdetail.7586022.html Swiss values 164.65 0.15

a) What height do we predict for her dauther after grown up?

b) Could the prediction be further refined?

Problem 8.5 (Sum of dependent random variables) We saw that if X, Y are independent Gaussian variables, then their (weighted) sum is again Gaussian. This problem illustrates that the assumption of independence is necessary.

Let X be a standard normal random variable. We define Y as

$$Y = \begin{cases} X & \text{if } |X| \geq c, \\ -X & \text{otherwise,} \end{cases}$$

for a fixed positive constant c.

a) Argue heuristically that also $Y \sim \mathcal{N}(0,1)$.

b) Argue heuristically and with simulations that (i) X and Y are dependent and (ii) the random variable $X + Y$ is not normally distributed.

Chapter 9

Estimation of Correlation and Simple Regression

Learning goals for this chapter:

- Explain the concept of correlation
- Explain different correlation coefficient estimates
- Explain the least squares criterion
- Explain the statistical model of the linear regression
- Apply linear model in R, check the assumptions, interpret the output

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter09.R.

We now start to introduce more complex and realistic statistical models (compared to (4.1)). In most of this chapter (and most of the remaining ones) we consider "linear models." We start by estimating the correlation and then linking the simple regression model to the bivariate Gaussian distribution, followed by extending the model to several so-called predictors in the next chapter. Chapter 11 presents the least squares approach from a variance decomposition point of view. A detailed discussion of linear models would fill entire books, hence we consider only the most important elements.

The (simple) linear regression is commonly considered the archetypical task of statistics and is often introduced as early as middle school, either in a black-box form or based on intuition. We approach the problem more formally and we will in this chapter (i) quantify the (linear) relationship between variables, (ii) explain one variable through another variable with the help of a "model".

9.1 Estimation of the Correlation

The goal of this section is to quantify the linear relationship between two random variables X and Y with the help of n pairs of values $(x_1, y_n), \dots, (x_n, y_n)$, i.e., realizations of the two random variables (X, Y). More formally, we will derive an estimator and estimate the covariance and correlation of a pair of random variables.

Based on Equation (8.12), an intuitive estimator of the correlation between the random variables X and Y is

$$r = \widehat{\operatorname{Corr}(X,Y)} = \frac{\widehat{\operatorname{Cov}(X,Y)}}{\sqrt{\widehat{\operatorname{Var}(X)}\widehat{\operatorname{Var}(Y)}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}}, \tag{9.1}$$

where we used the same denomiator for the covariance and the variance estimate (e.g., $n-1$). The estimate r is called the *Pearson correlation coefficient*. Just like the correlation, the Pearson correlation coefficient also lies in the interval $[-1, 1]$.

We will introduce a handy notation that is often used in the following:

$$s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \qquad s_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad s_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2. \tag{9.2}$$

Hence, we can express (9.1) as $r = s_{xy}/\sqrt{s_{xx}s_{yy}}$. Similarly, unbiased variance estimates are $s_{xx}/(n-1)$ and $s_{yy}/(n-1)$ and an estimate for the covariance is $s_{xy}/(n-1)$ (again an unbiased).

Example 9.1. In R, we use `cov()` and `cor()` to obtain an estimate of the covariance and of Pearson correlation coefficient. For the `penguin` data (see Example 1.9) we have between body mass and flipper length a covariance of 9824.42 and a correlation of 0.87. The latter was obtained with the command `with(penguins, cor(body_mass_g, flipper_length_mm, use="complete.obs"))`, where the argument `use` specifies the handling of missing values. ♣

A natural question we now tackle is if an observed correlation is statistically significant, i.e., testing the hypothesis $H_0 : \rho = 0$. As seen in Chapter 5 we need the distribution of the corresponding estimator. Let us consider bivariate normally distributed random variables as discussed in the last chapter. Naturally, r as given in (9.1) is an estimate of the correlation parameter ρ explicated in the density (8.18). Let R be the corresponding estimator of ρ based on (9.1), i.e., replacing (x_i, y_i) by (X_i, Y_i). With quite tedious work, it can be shown that the random variable

$$T = R\frac{\sqrt{n-2}}{\sqrt{1-R^2}} \tag{9.3}$$

is, under $H_0 : \rho = 0$, t-distributed with $n-2$ degrees of freedom. Hence, the test consists of using the test statistic (9.3) with the data and comparing the value with the appropriate quantile of the t-distribution. The corresponding test is described under Test 14.

The test is quite particular, as it only depends on the sample correlation and sample size. Astonishingly, for correlation estimates to be significant large sample sizes are required. For

Test 14: Test of correlation

Question: Is the correlation between two paired samples significant?

Assumptions: The pairs of values stem from a bivariate normal distribution.

Calculation: $t_{\text{obs}} = |r|\dfrac{\sqrt{n-2}}{\sqrt{1-r^2}}$ where r is the Pearson correlation coefficient (9.1).

Decision: Reject H_0: $\rho = 0$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n-2,1-\alpha/2}$.

Calculation in R: `cor.test( x, y, conf.level=1-alpha)`

example, $r = 0.25$ is not significant unless $n > 62$. Naturally, if we do not have Gaussian data, Test 14 is not exact and the resulting p-value an approximation only.

In order to construct confidence intervals for correlation estimates, we typically need the so-called *Fisher transformation*

$$W(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r) \tag{9.4}$$

and the fact that, for bivarate normally distributed random variables, the distribution of $W(R)$ is approximately $\mathcal{N}\big(W(\rho), 1/(n-3)\big)$. Then a straight-forward confidence interval can be constructed:

$$\frac{W(R)-W(\rho)}{\sqrt{1/(n-3)}} \overset{\text{app}}{\sim} \mathcal{N}(0,1), \qquad \text{approx. CI for } W(\rho) : \Big[W(R) \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\Big]. \tag{9.5}$$

A confidence interval for ρ requires a back-transformation and is shown in CI 6.

CI 6: Confidence interval for the Pearson correlation coefficient

A sample approximate $(1-\alpha)$ confidence interval for r is

$$\Big[\tanh\Big(\operatorname{arctanh}(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big), \tanh\Big(\operatorname{arctanh}(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big)\Big]$$

where tanh and arctanh are the hyperbolic and inverse hyperbolic tangent functions (see (9.4)).

Example 9.2. We estimate the correlation of the scatter plots from Figure 8.3 and calculate the corresponding 95%-confidence intervals thereof in R-Code 9.1. For these specific samples, the confidence intervals obtained from simulating with $\rho = 0$ and 0.1 cover the value zero. Hence, ρ is not statistically significant (different from zero).

The width of the six intervals are slightly different and the correlation estimate is not precisely in the center, both due to the back-transformation. ♣

R-Code 9.1 Pearson correlation coefficient and confidence intervals of the scatter plots from Figure 8.3.

```
require( mvtnorm)
set.seed(12)
rho <- c(-.25, 0, .1, .25, .75, .9)
n <- 200
out <- matrix(0, 3,6, dimnames=list(c("rhohat","b_low","b_up"), paste(rho)))
for (i in 1:6) {
  Sigma <- array( c(1, rho[i], rho[i], 1), c(2,2))
  sample <- rmvnorm( n, sigma=Sigma)
  out[1,i] <- cor( sample)[2]
  out[2:3,i] <- tanh( atanh( out[1,i]) + qnorm( c(0.025,0.975))/sqrt(n-3))
}
print( out, digits=2)
##         -0.25       0    0.1 0.25 0.75  0.9
## rhohat -0.181 -0.1445  0.035 0.24 0.68 0.89
## b_low  -0.312 -0.2777 -0.104 0.11 0.60 0.86
## b_up   -0.044 -0.0059  0.173 0.37 0.75 0.92
```

There are alternatives to Pearson correlation coefficient, that is, there are alternative estimators of the correlation. The two most common ones are called Spearman's ϱ or Kendall's τ and are based on ranks and thus also called rank correlation coefficients. In brief, Spearman's ϱ is calculated similarly to (9.1), where the values are replaced by their ranks. Kendall's τ compares the number of concordant (if $x_i < x_j$ then $y_i < y_j$) and discordant (if $x_i < x_j$ then $y_i > y_j$) pairs. As expected, Pearson's correlation coefficient is not robust, while Spearman's ϱ or Kendall's τ are "robust".

Example 9.3. Correlation is a measure of linear dependency and it is impossible to deduce general relationship in a scatterplot based on a single value. Figure 9.1 (based on R-Code 9.2) gives the four scatterplots of the so-called *anscombe* data, all having an identical Pearson correlation coefficient of 0.82 but vastly different shape. As given in R-Code 9.2 the robustness of Spearman's ϱ or Kendall's τ are evident. If there are no outliers and the the data does exhibit a linear relationship, there is, of course, quite some agreement between the estimates.

Note that none of the estimators "detects" the squared relationship in the second sample. The functional form can be "approximated" by a linear one and hence the large correlation estimates. ♣

R-Code 9.2 *anscombe* data: visualization and correlation estimates. (See Figure 9.1.)

```
library( faraway)      # dataset 'anscombe' is provided by this package
data( anscombe)
#  head( anscombe)     # dataset with eleven observations and 4x2 variables
with( anscombe, { plot(x1, y1); plot(x2, y2); plot(x3, y3);  plot(x4, y4) })
sel <- c(0:3*9+5)      # extract diagonal entries of sub-block
print(rbind( pearson=cor(anscombe)[sel],
         spearman=cor(anscombe, method='spearman')[sel],
         kendall=cor(anscombe, method='kendall')[sel]), digits=2)
```

```
##          [,1] [,2] [,3] [,4]
## pearson  0.82 0.82 0.82 0.82
## spearman 0.82 0.69 0.99 0.50
## kendall  0.64 0.56 0.96 0.43
```

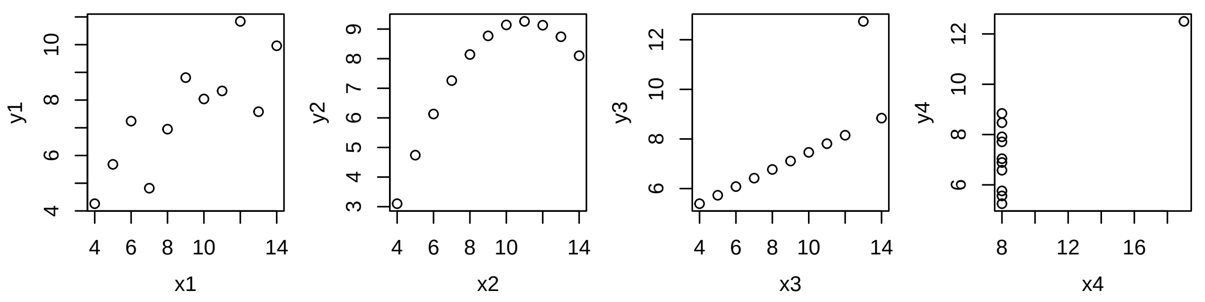


Figure 9.1: *anscombe* data, the four cases all have the same Pearson correlation coefficient of 0.82, yet the scatterplot shows a completely different relationship. (See R-Code 9.2.)

9.2 Estimation of a p-Variate Mean and Covariance

We now extend the estimation of the covariance and correlation of pairs or variables to random vectors of arbitrary dimension and thus we will rely again on vector notation. The estimators for parameters of random vectors are constructed in a manner similar to that for the univariate case. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be a realization of the random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ iid for some p-variate distribution with $n > p$. We have the following estimators for the mean and the variance

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top \tag{9.6}$$

and corresponding estimates

4 min

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top. \tag{9.7}$$

The estimators and estimates are intuitive generalizations of the univariate forms (see Problem 8.3.a). In fact, it is possible to show that the two estimators given in (9.6) are unbiased estimators of $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$.

Remark 9.1. If $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\widehat{\boldsymbol{\mu}}$ is the maximum likelihood and the method of moment estimator. For the variance, the maximum likelihood estimator is $(n-1)/n\widehat{\boldsymbol{\Sigma}}$. ♣

A normally distributed random variable is determined by two parameters, namely the mean μ and the variance σ^2. A multivariate normally distributed random vector is determined by p and $p(p+1)/2$ parameters for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. The remaining $p(p-1)/2$ parameters in $\boldsymbol{\Sigma}$ are determined by symmetry. However, the $p(p+1)/2$ values cannot be arbitrarily chosen since $\boldsymbol{\Sigma}$ must be positive-definite (in the univariate case the variance must be strictly positive as well). As long as $n > p$, the estimator in (9.7) satisfies this condition. If $p \leq n$, additional assumptions about the structure of the matrix $\boldsymbol{\Sigma}$ are needed.

Example 9.4. Similar as in R-Code 8.2, we generate bivariate realizations with different sample sizes ($n = 10, 50, 100, 500$). We estimate the mean vector and covariance matrix according to (9.7); from these we can calculate the corresponding isolines of the bivariate normal density (where with plug-in estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). Figure 9.2 (based on R-Code 9.3) shows the estimated 95% and 50% confidence regions (isolines). As n increases, the estimation improves, i.e., the estimated ellipses are closer to the ellipses based on the true (unknown) parameters.

Note that several packages provide a function called *`ellipse`* with varying arguments. We use the call *`ellipse::ellipse()`* to ensure using the one provided by the package *`ellipse`*. ♣

R-Code 9.3: Bivariate normally distributed random numbers for various sample sizes with contour lines of the density and estimated moments. (See Figure 9.2.)

```
set.seed( 14)
require( ellipse)                     # to draw ellipses
n <- c( 10, 50, 100, 500)             # different sample sizes
mu <- c(2, 1)                         # theoretical mean
Sigma <- matrix( c(4, 2, 2, 2), 2)    #  and covariance matrix
for (i in 1:4) {
  plot(ellipse::ellipse( Sigma, centre=mu, level=.95), col='gray',
       xaxs='i', yaxs='i', xlim=c(-4, 8), ylim=c(-4, 6), type='l')
  lines( ellipse::ellipse( Sigma, centre=mu, level=.5), col='gray')
  sample <- rmvnorm( n[i], mean=mu, sigma=Sigma)          # draw realization
  points( sample, pch=20, cex=.4)                         # add realization
  muhat <- colMeans( sample)        # apply( sample, 2, mean) # is identical
  Sigmahat <- cov( sample)          # var( sample)            # is identical
  lines( ellipse::ellipse( Sigmahat, centre=muhat, level=.95), col=2, lwd=2)
  lines( ellipse::ellipse( Sigmahat, centre=muhat, level=.5), col=4, lwd=2)
  points( rbind( muhat), col=3, cex=2)
  text( -2, 4, paste('n =',n[i]))
}
```

```
muhat      # Estimates for n=500
## [1] 1.93252 0.97785
Sigmahat
##         [,1]   [,2]
## [1,] 4.7308 2.2453
## [2,] 2.2453 1.9619
c(cov2cor( Sigma)[2], cov2cor( Sigmahat)[2])  # correlation= sqrt(2)/2
## [1] 0.70711 0.73700
```

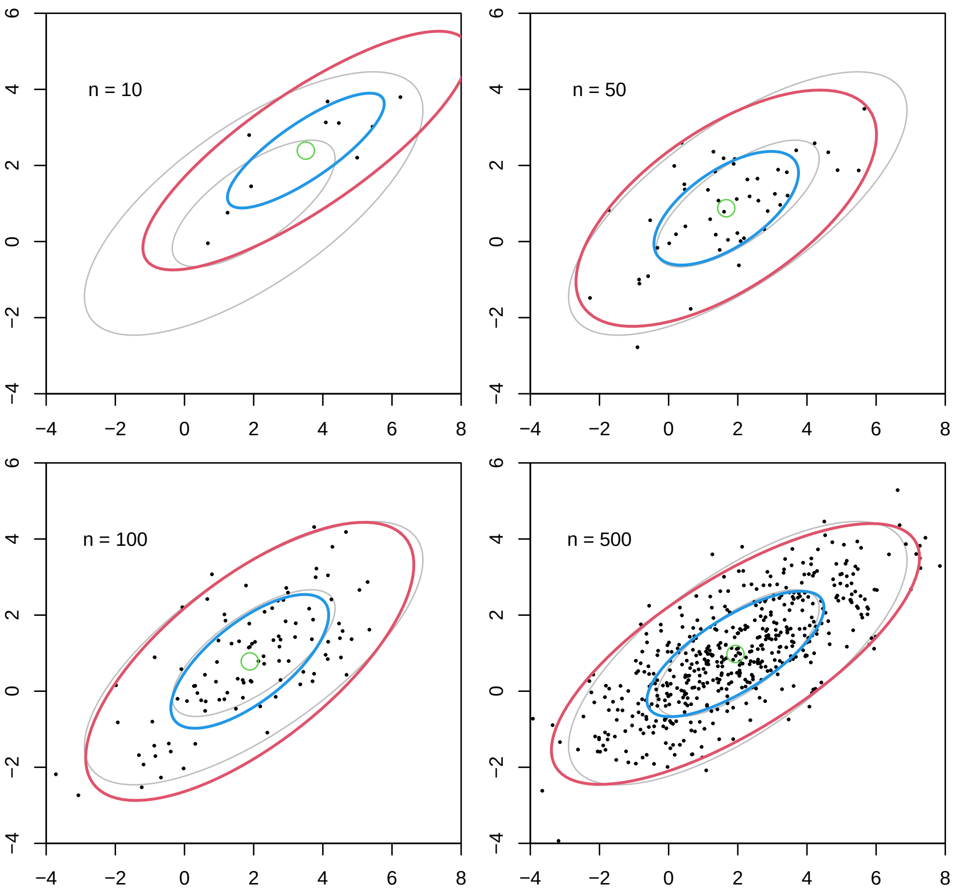


Figure 9.2: Bivariate normally distributed random numbers. The contour lines of the true density are in gray. The isolines correspond to 95% and 50% quantiles and the sample mean in green. (See R-Code 9.3.)

In general, interval estimation within random vectors is much more complex compared to random variables. In fact, due to the multiple dimensions, a better term for 'interval estimation' would be 'area or (hyper-)volume estimation'. In practice, one often marginalizes and constructs intervals for individual elements of the parameter vector (which we will also practice in Chapter 10, for example).

In the trivial case of $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known, we have $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ (can also be shown with Property 8.24) and the uncertainty in the estimate can be expressed with ellipses or hyper-ellipsoids. In case of unknown $\mathbf{\Sigma}$, plug-in estimates can be used, at the price of accuracy. This setting is essentially the multivariate version of the middle panel of Figure 4.3. The discussion of uncertainty in $\widehat{\mathbf{\Sigma}}$ is beyond the scope of this book.

To estimate the correlation matrix we define the diagonal matrix $\mathbf{D}$ with elements $d_{ii} = \big((\widehat{\mathbf{\Sigma}})_{ii}\big)^{-1/2}$. The estimated correlation matrix is given by

$$\mathbf{R} = \mathbf{D}\widehat{\mathbf{\Sigma}}\mathbf{D} \tag{9.8}$$

(see Problem 9.1.**a**). In R, the function *cov2cor()* can be used.

Example 9.5. We retake the setup of Example 9.4 for $n = 50$ and draw 100 realizations. Figure 9.3 visualizes the density based on the estimated parameters (as in Figure 9.2 top left). The variablity seems dramatic. However, if we would estimate the parameters marginally, we would have the same estimates and the same uncertainties. The only additional element here is the estimate of the covariance, i.e., the off-diagonal elements of $\widehat{\mathbf{\Sigma}}$. ♣

R-Code 9.4: Visualizing uncertainty when estimating parameters of bivariate normally distributed random numbers. (See Figure 9.3.)

```r
set.seed( 14)
n <- 50                # fixed sample size
R <- 100               # draw R realizations
plot( 0, type='n', xaxs='i', yaxs='i', xlim=c(-4, 8), ylim=c(-4, 6))
for (i in 1:R) {
  sample <- rmvnorm( n, mean=mu, sigma=Sigma)
  muhat <- colMeans( sample)        # estimated mean
  Sigmahat <- cov( sample)          # estimated variance
  lines( ellipse::ellipse( Sigmahat, centre=muhat, level=.95), col=2)
  lines( ellipse::ellipse( Sigmahat, centre=muhat, level=.5), col=4)
  points( rbind( muhat), col=3, pch=20) # add mean in green
}
```

Assessing multivariate normality from a sample is not straightforward. A simple approach is to reduce the multivariate dataset to univariate ones, which we assess, typically, with QQ-plots. A more elaborate approach to assess if a p dimensional dataset indicates any suspicion against multivariate normality is:

1. assess QQ-plots of all marginal variables;
2. assess pairs plots for elliptic scatter plots of all bivariate pairs;
3. assess QQ-plots of linear combinations, e.g. sums of two, three, ..., all variables;
4. assess $d_1, \dots, d_n$ with $d_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \widehat{\mathbf{\Sigma}}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ with a QQ-plot against a χ^2_p distribution.

If any of the above fails, we have evidence against multivariate normality of the entire dataset; subsets thereof may still be of course. Note that only for the fourth point we do need estimates of the mean and covariance because the marginal QQ-plots are scale invariant.

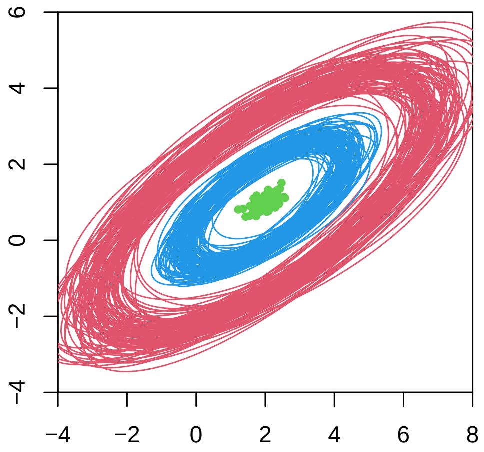


Figure 9.3: Visualizing uncertainty when estimating parameters of bivariate normally distributed random numbers ($n = 50$ and 100 realizations). The red and blue isolines correspond to the 95% and 50% quantiles of the bivariate normal density with plugin estimates. The sample means are in green. (See R-Code 9.4.)

9.3 Simple Linear Regression

The correlation is a symmetric measure, $\text{Corr}(X, Y) = \text{Corr}(Y, X)$, thus there is no preference between either variable. We now extend the idea of quantifying the linear relationship to an asymmetric setting where we have one variable as fixed (or given) and observe the second one as a function of the first. Of course, in practice even the given variable has to be measured. Here, we refer to situations such as "given the height of the mother, the height of the child is", "given the dose, the survival rate is", etc. In a linear regression model we determine a linear relationship between two variables.

More formally, in *simple linear regression* a *dependent variable* is explained linearly through a single *independent variable*. The statistical model writes as

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i && (9.9)\\ &= \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n, && (9.10) \end{aligned}$$

with

- Y_i: dependent variable, measured values, or observations;
- x_i: independent variable, predictor, assumed known or observed and not stochastic;
- β_0, β_1: parameters (unknown);
- ε_i: error, error term, noise (unknown), with symmetric distribution around zero.

It is often also assumed that $\text{Var}(\varepsilon_i) = \sigma^2$ and that the errors are independent of each other. We make another simplification and further assume $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with unknown σ^2. Thus, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \dots, n$, and Y_i and Y_j are independent for $i \neq j$. Because of the varying mean, $Y_1, \dots, Y_n$ are not identically distributed.

Fitting a linear model is based on pairs of data $(x_1, y_1), \dots, (x_n, y_n)$ and the model (9.10) with the goal to find estimates for β_0 and β_1, essentially an estimation problem. In R, this task is straightforward as shown in a motivating example below which has been dissected and will illustrate the theoretical concepts in the remaining part of the chapter.

Example 9.6 (*hardness* data)**.** One of the steps of manufacturing metal springs is a quenching bath that cools metal back to room temperature to prevent that a slow cooling process dramatically changes the metal's microstructure. The temperature of the bath has an influence on the hardness of the springs. Data is taken from Abraham and Ledolter (2006). The Rockwell scale measures the hardness of technical materials and is denoted by HR (Hardness Rockwell). Figure 9.4 shows the Rockwell hardness of coil springs as a function of the temperature (in degrees Celcius) of the quenching bath, as well as the line of best fit. R-Code 9.5 shows how a simple linear regression is performed with the function `lm()` and the 'formula' statement `Hard~Temp`. ♣

R-Code 9.5 *hardness* data from Example 9.6. (See Figure 9.4.)

```
Temp <- c(30, 30, 30, 30, 40, 40, 40, 50, 50, 50, 60, 60, 60, 60)
#  Temp <- rep( 10*3:6, c(4, 3, 3, 4))  # alternative way based on `rep()`
Hard <- c(55.8, 59.1, 54.8, 54.6, 43.1, 42.2, 45.2,
          31.6, 30.9, 30.8, 17.5, 20.5, 17.2, 16.9)
plot( Temp, Hard, xlab="Temperature [C]", ylab="Hardness [HR]")
lm1 <- lm( Hard~Temp)     #  fitting of the linear model
abline( lm1)              #  add fit of the linear model to linear model
```

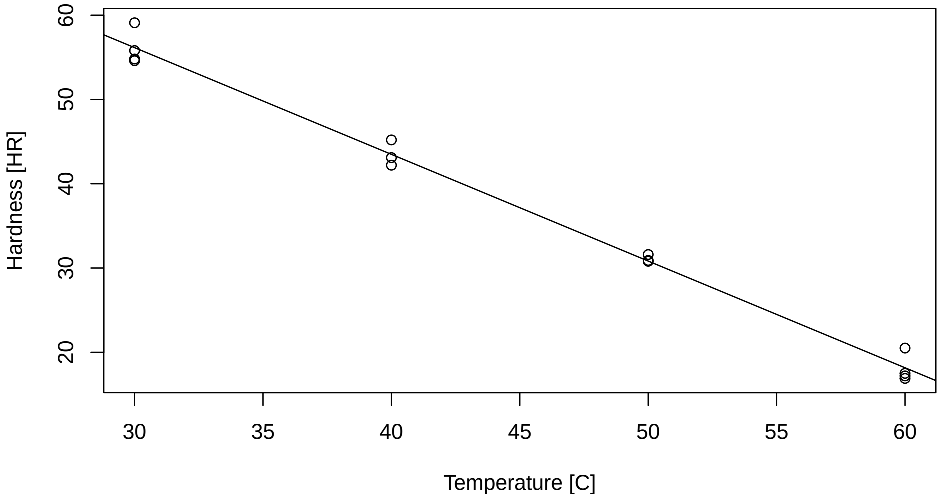


Figure 9.4: *hardness* data: hardness as a function of temperature. The black line is the fitted regression line. (See R-Code 9.5.)

Classically, we estimate $\widehat{\beta}_0$ and $\widehat{\beta}_1$ with a least squares approach (see Section 4.2.1), which minimizes the sum of squared residuals. That means that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are determined such that

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \tag{9.11}$$

is minimized. This concept is also called *ordinary least squares* (OLS) method to emphasize that we have iid errors and no weighting is taken into account. The solution of the minimization is given by

$$\widehat{\beta}_1 = r\sqrt{\frac{s_{yy}}{s_{xx}}} = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\,, \tag{9.12}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}\,. \tag{9.13}$$

and are termed the *estimated regression coefficients*. The first equality in (9.12) emphasizes the difference to a correlation estimate. Here, we correct Pearson's correlation estimate with $\sqrt{s_{yy}/s_{xx}}$. It can be shown that the associated estimators are unbiased (see Problem 9.1.**b**). The *predicted values* are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i\,, \tag{9.14}$$

which lie on the *estimated regression line*

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x\,. \tag{9.15}$$

The *residuals* (observed minus predicted values) are

$$r_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\,. \tag{9.16}$$

Finally, an estimate of the variance σ^2 of the errors ε_i is given by

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2, \tag{9.17}$$

where we have used a slightly different denominator than in Example 4.6.2 to obtain an unbiased estimate. More precisely, instead of estimating one mean parameter we have to estimate two regression coefficients and thus instead of $n-1$ we use $n-2$. For simplicity, we do not introduce another symbol for the estimate here. The variance estimate $\widehat{\sigma}^2$ is often termed *mean squared error* and its root, $\widehat{\sigma}$, *residual standard error*.

Example 9.7 (continuation of Example 9.6)**.** R-Code 9.6 illustrates how to access the estimates, the fitted values and the residuals from the output of the linear model fit of Example 9.6. More specifically, the function `lm()` generates an object of class `"lm"` and the functions `coef()`, `fitted()` and `residuals()` extract the corresponding elements from the object.

To access the residual standard error we need to further "digest" the object via `summary()`. More specifically, an estimate of σ^2 can be obtained via `residuals(lm1)` and Equation (9.17) (e.g., `sum(residuals(lm1)^2)/(14-2)`) or directly via `summary(lm1)$sigma^2`.

There are more methods defined for an object of class `"lm"`, for example `abline()` as used in R-Code 9.5; more will be discussed subsequently. ♣

R-Code 9.6 *hardness* data from Example 9.7.

```
coef( lm1)         # extract coefficients, here hat{beta_0}, here hat{beta_1}
## (Intercept)        Temp
##     94.1341     -1.2662
rbind( observation=Hard, fitted=fitted( lm1), residuals=residuals( lm1))[,1:6]
##                    1       2       3       4        5       6
## observation 55.80000 59.1000 54.8000 54.6000 43.10000 42.2000
## fitted      56.14945 56.1495 56.1495 56.1495 43.48791 43.4879
## residuals   -0.34945  2.9505 -1.3495 -1.5495 -0.38791 -1.2879
head( Hard - (fitted( lm1) + residuals( lm1)))
## 1 2 3 4 5 6
## 0 0 0 0 0 0
summary(lm1)$sigma   # equivalent to:  sqrt( sum(residuals( lm1)^2)/(14-2))
## [1] 1.495
```

9.4 Inference in Simple Linear Regression

In this section we discuss the quality of the fitted regression line. In our setting, ‘quality’ is interpreted in the sense of whether the model makes (statistical) sense by verifying if the estimated slope is significantly different to zero and if the model explains variablity in the data. We will revisit these questions in later chapters again where we further discuss formal justifications. In this section we give the conceptual ideas with some justifications in Problem 9.1.

When fitting a linear model with R, much of the discussion is given by the summary output of linear model fit. R-Code 9.7 gives the summary output which is very classical and further details are elaborated in Figure 9.5.

Example 9.8 (continuation of Examples 9.6 and 9.7)**.** R-Code 9.7 outputs the summary of a linear fit by *summary(lm1)*. ♣

The output is essentially build of four blocks. The first restates the call of the model fit, the second states information about the residuals, the third summarizes the regression coefficients and the fourth digests the overall quality of the regression.

The summary of the residuals serves to check if the residuals are symmetric (median should be close to zero, up to the sign similar quartiles) or if there are outliers (small minimum or maximum value). In the next chapter, we will revisit graphical assessments of the residuals.

The third block of the summary output contains for each estimated coefficient the following four numbers: the estimate itself, its standard error, the ratio of the latter two and the associated p-value of the Test 15. More specifically, for the slope parameters of a simple linear regression these are $\widehat{\beta}_1$, $\mathrm{SE}(\widehat{\beta}_1)$ $\widehat{\beta}_1/\,\mathrm{SE}(\widehat{\beta}_1)$ and a p-value.

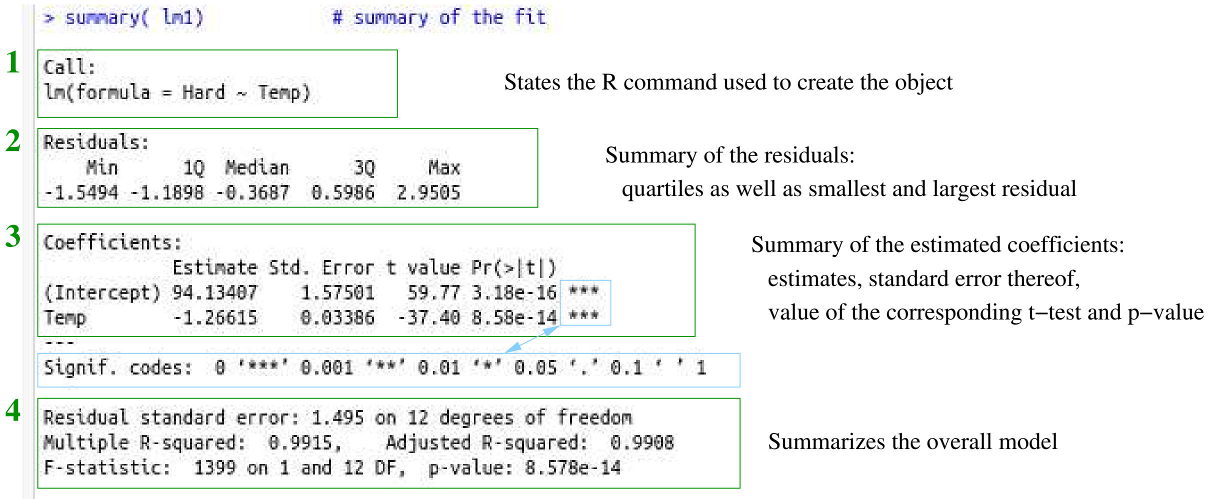
To obtain the aforementioned p-value consider $\widehat{\beta}_1$ as an estimator. When properly scaled it can be shown that the estimator of $\widehat{\beta}_1/\,\mathrm{SE}(\widehat{\beta}_1)$ has a t-distribution and thus we can apply a

R-Code 9.7 *hardness* data from Example 9.6.

```
summary( lm1)             # summary of the fit
##
## Call:
## lm(formula = Hard ~ Temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.550 -1.190 -0.369  0.599  2.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  94.1341     1.5750    59.8  3.2e-16 ***
## Temp         -1.2662     0.0339   -37.4  8.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.5 on 12 degrees of freedom
## Multiple R-squared:  0.991,Adjusted R-squared:  0.991
## F-statistic: 1.4e+03 on 1 and 12 DF,  p-value: 8.58e-14
```

statistical test for $H_0 : \beta_1 = 0$, detailed in Test 15. That means for the slope parameter, the last number is $\text{P}\big(V \geq |\widehat{\beta}_1|/\,\text{SE}(\widehat{\beta}_1)\big)$ with V a t-distributed random variable with $n-2$ degrees of freedom. This latter test is conceptually very similar to Test 1 and essentially identical to Test 14, see Problem 9.1.1.

The final block summarizes the overall fit of the model. In the case of a "good" model fit, the



```
> summary( lm1)             # summary of the fit

Call:
lm(formula = Hard ~ Temp)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5494 -1.1898 -0.3687  0.5986  2.9505

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 94.13407    1.57501   59.77 3.18e-16 ***
Temp        -1.26615    0.03386  -37.40 8.58e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.495 on 12 degrees of freedom
Multiple R-squared:  0.9915,    Adjusted R-squared:  0.9908
F-statistic:  1399 on 1 and 12 DF,  p-value: 8.578e-14
```

States the R command used to create the object

Summary of the residuals:
quartiles as well as smallest and largest residual

Summary of the estimated coefficients:
estimates, standard error thereof,
value of the corresponding t–test and p–value

Summarizes the overall model

Figure 9.5: R summary output of an object of class *lm* (here for the *hardness* data from Example 9.6).

Test 15: Test of a linear relationship in regression

Question: Is there a linear relationship between the dependent and independent variables?

Assumptions: Based on the sample $x_1, \dots, x_n$, the second sample is a realization of a normally distributed random variable with expected value $\beta_0 + \beta_1 x_i$ and variance σ^2.

Calculation: $t_{\text{obs}} = \dfrac{\widehat{\beta}_1}{\text{SE}(\widehat{\beta}_1)}$

Decision: Reject H_0: $\beta_1 = 0$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n-2,1-\alpha/2}$.

Calculation in R: `summary( lm( y ~ x ))`

resulting absolute residuals are small. More precisely, the residual sums of squares are small, at least compared to the total variability in the data s_{yy}. In later chapters, we will formalize that a linear regression decomposes the total variablity in the data in two components, namely the variablity explained by the model and the variablity of the residual. We often write this as

$$\text{SS}_T = \text{SS}_M + \text{SS}_E, \tag{9.18}$$

where the subscript indicates ‘total’, ’model’ and ’error’. This statement is of general nature and holds not only for linear regression models but for virtually all statistical models. Here, the somewhat surprising fact is that we have the simple expressions $\text{SS}_T = \sum_i (y_i - \bar{y})^2$, $\text{SS}_M = \sum_i (\widehat{y}_i - \bar{y})^2$ and $\text{SS}_E = \sum_i (y_i - \widehat{y}_i)^2$ (see also Problem 9.1.**d**). The specific values of the final block for the simple regression are calculated as follows

$$\text{Multiple } R^2\text{:}\quad R^2 = 1 - \frac{\text{SS}_E}{\text{SS}_T} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \tag{9.19}$$

$$\text{Adjusted } R^2\text{:}\quad R^2_{\text{adj}} = R^2 - \frac{1 - R^2}{n-2}, \tag{9.20}$$

$$\text{Observed value of } F\text{-test:}\quad \frac{\text{SS}_M}{\text{SS}_E/(n-2)}, \tag{9.21}$$

The multiple R^2 is sometimes called the *coefficient of determination* or simply *R-squared*. The adjusted R^2 is a weighted version of R^2 and we revisit it later for more details. For a simple linear regression the F-test is an alternative form to test $H_0 : \beta_1 = 0$ (see Problem 9.1.**e**) and a deeper meaning will be revealed with more complex models. The denominator of (9.21) is $\widehat{\sigma}^2$. Note that for the simple regression the degrees of freedom of the F-test are 1 and $n - 2$.

Hence, if the regression model explains a lot of variability in the data, SS_M is large compared to SS_E, that means SS_E is small compared to SS_T, which implies R^2 is close to one and the observed value of the F-test (F-statistic) is large.

In simple linear regression, the central task is often to determine whether there exists a linear relationship between the dependent and independent variables. This can be tested with the hypothesis $H_0 : \beta_1 = 0$ (Test 15). We do not formally derive the test statistic here. The idea is to replace in Equation (9.12) the observations y_i with random variables Y_i with distribution specified by (9.10) and derive the distribution of a test statistic.

For prediction of the dependent variable at x_0, a specific, given value of the independent variable, we plug-in x_0 in Equation (9.15), i.e., we determine the corresponding value of the regression line. The function `predict()` can be used for prediction in R.

Prediction at a (potentially) new value x_0 can also be written as

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_0 = \overline{y} + s_{xy}(s_{xx})^{-1}(x_0 - \overline{x})\,. \tag{9.22}$$

Thus, the last expression is equivalent to Equation (8.27) but with estimates instead of (unknown) parameters. Or in other words, simple linear regression is equivalent to conditional expectation of a bivariate normal distribution with plug-in estimates.

The uncertainty of the prediction depends on the uncertainty of the estimated parameter. Specifically:

$$\text{Var}(\widehat{\mu}_0) = \text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0), \tag{9.23}$$

is an expression in terms of the independent variables, the dependent variables and (implicitly) the variance of the error term. Because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not necessarily independent, it is quite tedious to derive the explicit expression. However, with matrix notation, the above variance is straightforward to calculate, as will be illustrated in Chapter 10.

To construct confidence intervals for a prediction, we must discern whether the prediction is for the mean response $\widehat{\mu}_0$ or for an unobserved (e.g., future) observation $\widehat{y}_0$ at x_0. The former prediction interval depends on the variability of the estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. For the latter the prediction interval depends on the uncertainty of $\widehat{\mu}_0$ and additionally on the variability of the error ε, i.e., $\widehat{\sigma}^2$. Hence, the latter is always wider than the former. In R, these two types are denoted — somewhat intuitively — with `interval="confidence"` and `interval="prediction"`, (see Example 9.9 and R-Code 9.8). The confidence interval summary CI 7 gives the precise formulas, we refrain from the derivation here.

Example 9.9 (continuation of Examples 9.6 to 9.8)**.** We close this example by constructing pointwise confidence intervals for the mean response and for an unobserved observation. Figure 9.6 (based on R-Code 9.8) illustrates the confidence intervals for a very fine grid of temperatures. For each temperature, we calculate the intervals (hence "pointwise") but as classically done they are visualized by lines. The width of both increases as we move further from $\bar{x}$. ♣

Linear regression framework is based on several statistical assumptions, including on the errors ε_i. After fitting the model, we must check whether these assumptions are realistic or if the residuals indicate evidence against these (see Figure 1.1). For this task, we return in Chapter 10 to the analysis of the residuals and their properties.

R-Code 9.8 *hardness* data: predictions and pointwise confidence intervals. (See Figure 9.6.)

```
new <- data.frame( Temp = seq(25, 65, by=.5))
pred.w.clim <- predict( lm1, new, interval="confidence") # for hat(mu)
pred.w.plim <- predict( lm1, new, interval="prediction") # for hat(y), wider!
plot( Temp, Hard, xlab="Temperature [C]", ylab="Hardness [HR]",
     xlim=c(28,62), ylim=c(10,65))       # Plotting observations
matlines( new$Temp, cbind(pred.w.clim, pred.w.plim[,-1]),
     col=c(1,2,2,3,3), lty=c(1,1,1,2,2)) # Prediction intervals are wider!
```

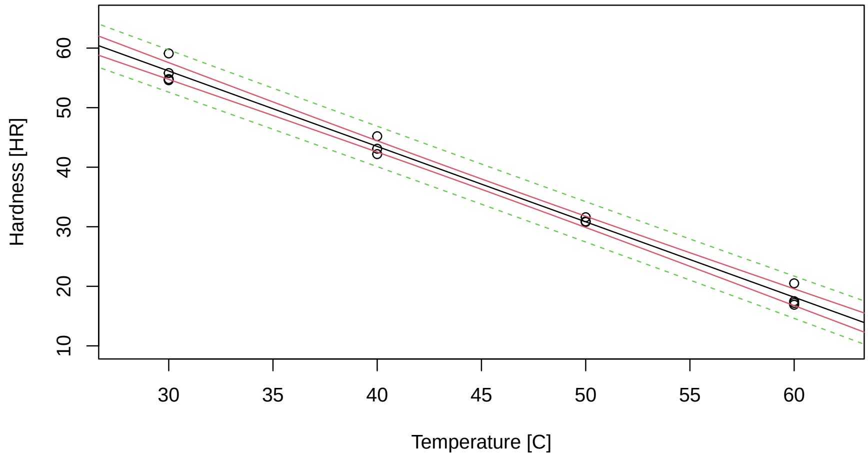


Figure 9.6: *hardness* data: hardness as a function of temperature. Black: fitted regression line, red: confidence intervals for the mean $\widehat{\mu}_i$ (pointwise), green: prediction intervals for (future) predictions $\widehat{y}_i$ (pointwise). (See R-Code 9.8.)

CI 7: Confidence intervals for mean response and for prediction

A $(1-\alpha)$ confidence interval for the mean response $\widehat{\mu}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is

$$\Big[\ \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{n-2,1-\alpha/2}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{x_0 - \bar{x}}{\sum_i (x_i - \bar{x})^2}}\ \Big].$$

A $(1-\alpha)$ prediction interval for an unobserved observation of $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is

$$\Big[\ \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{n-2,1-\alpha/2}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{x_0 - \bar{x}}{\sum_i (x_i - \bar{x})^2}}\ \Big].$$

In both intervals we use estimators, i.e., all the estimates y_i are to be replaced with Y_i to obtain estimators.

9.5 Bibliographic Remarks

We recommend the book from Fahrmeir *et al.* (2009) (German) or Fahrmeir *et al.* (2013) (English), which is both detailed and accessible. Many other books contain a single chapter on simple linear regression.

There exists a fancier version of `anscombe` data, called the 'DataSaurus Dozen', found on https://blog.revolutionanalytics.com/2017/05/the-datasaurus-dozen.html. Even the transitional frames in the animation https://blog.revolutionanalytics.com/downloads/DataSaurus%20Dozen.gif maintain the same summary statistics to two decimal places.

9.6 Exercises and Problems

Problem 9.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Show that the off-diagonal entries of the matrix $\mathbf{R}$ of (9.8) are equivalent to (9.8) and that the diagonal elements are one.

b) Show that the ordinary least squares regression coefficient estimators are unbiased.

c) Show that the rightmost expression of (9.12) can be written as $\sum_{i=1}^n (x_i - \overline{x})y_i / s_{xx}$ and deduce $\operatorname{Var}(\widehat{\beta}_1) = \sigma^2 / s_{xx}$

d) Show the decomposition (9.18) by showing that $\mathrm{SS}_M = r^2 s_{yy}$ and $\mathrm{SS}_E = (1 - r^2)s_{yy}$.

e) Show that the value of the F-test in the lm summary is equivalent to t^2_{obs} of Test 15 and to t^2_{obs} of Test 14.

Problem 9.2 (Correlation) For the `swiss` dataset, calculate the correlation between the variables `Catholic` and `Fertility` as well as `Catholic` and `Education`. What do you conclude?

Hint: for the interpretation, you might use the parallel coordinate plot, as shown in Figure 1.10 in Chapter 1.

Problem 9.3 (Bivariate normal distribution) Consider the random sample $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{\text{iid}}{\sim} \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

We define estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{X}_i - \widehat{\boldsymbol{\mu}})(\mathbf{X}_i - \widehat{\boldsymbol{\mu}})^\top,$$

a) Explain in words that these estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ "generalize" the univariate estimators for μ and σ.

b) Simulate $n = 500$ iid realizations from $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the function `rmvnorm()` from package `mvtnorm`. Draw a scatter plot of the results and interpret the figure.

c) Add contour lines of the density of $\mathbf{X}$ to the plot. Calculate an eigendecomposition of $\boldsymbol{\Sigma}$ and place the two eigenvectors in the center of the ellipses.

d) Estimate $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the correlation between X_1 and X_2 from the 500 simulated values using `mean()`, `cov()` and `cor()`, respectively.

e) Redo the simulation with several different covariance matrices, i.e., choose different values as entries for the covariance matrices. What is the influence of the diagonal elements and the off-diagonal elements of the covariance matrix on the shape of the scatter plot?

Problem 9.4 (Variance under simple random sampling) In this problem we proof the hint of Problem 7.1.d. Let $Y_1, \dots, Y_m$ be iid with $\mathrm{E}(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$. We take a random sample $Y_1', \dots, Y_n'$ of size n without replacement and denote the associated mean with $\overline{Y'}$.

a) Show that $\mathrm{E}\big(\overline{Y'}\big) = \mu$.

b) Show that $\mathrm{Cov}(Y_i', Y_j') = \dfrac{-\sigma^2}{m-1}$, if $i \neq j$ and $\mathrm{Cov}(Y_i', Y_j') = \sigma^2$, if $i = j$.

c) Show that $\mathrm{Var}\big(\overline{Y'}\big) = \dfrac{\sigma^2}{n}\Big(1 - \dfrac{n-1}{m-1}\Big)$. Discuss this result.

Problem 9.5 (Linear regression) The dataset `twins` from the package `faraway` contains the IQ of mono-zygotic twins where one of the siblings has been raised by the biological parents and the second one by foster parents. The dataset has been collected by Cyril Burt. Explain the IQ of the foster sibling as a function of the IQ from his sibling. Give an interpretation. Discuss the shortcomings and issues of the model.

Problem 9.6 (Linear regression) In a simple linear regression, the data are assumed to follow $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$. We simulate $n = 15$ data points from that model with $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 2$ and the follwoing values for x_i.
Hint: to start, copy–paste the following lines into your R-Script.

```
set.seed(5)          ## for reproducable simulations
beta0.true <- 1      ## true parameters, intercept
beta1.true <- 2      ##    and slope
## observed x values:
x <- c(2.9, 6.7, 8.0, 3.1, 2.0, 4.1, 2.2, 8.9, 8.1, 7.9, 5.7, 1.6, 6.6, 3.0, 6.3)
## simulation of y values:
y <- beta0.true + x * beta1.true + rnorm(15, sd = 2)
data <- data.frame(x = x, y = y)
```

a) Plot the simulated data in a scatter plot. Calculate the Pearson correlation coefficient and the Spearman's rank correlation coefficient. Why do they agree well?

b) Estimate the linear regression coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using the formulas from the script. Add the estimated regression line to the plot from (a).

c) Calculate the fitted values $\widehat{Y}_i$ for the data in `x` and add them to the plot from (a).

d) Calculate the residuals $(y_i - \widehat{y}_i)$ for all n points and the residual sum of squares SS $= \sum_i (y_i - \widehat{y}_i)^2$. Visualize the residuals by adding lines to the plot with `segments()`. Are the residuals normally distributed? Do the residuals increase or decrease with the fitted values?

e) Calculate standard errors for β_0 and β_1. For $\widehat{\sigma}_\varepsilon = \sqrt{\mathrm{SS}/(n-2)}$, they are given by

$$\widehat{\sigma}_{\beta_0} = \widehat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}}\,, \qquad \widehat{\sigma}_{\beta_1} = \widehat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum_i (x_i - \overline{x})^2}}\,.$$

f) Give an empirical 95% confidence interval for β_0 and β_1. (The degree of freedom is the number of observations minus the number of parameters in the model.)

g) Calculate the values of the t statistic for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and the corresponding two-sided p-values.

h) Verify your result with the R function `lm()` and the corresponding S3 methods `summary()`, `fitted()`, `residuals()` and `plot()` (i.e., apply these functions to the returned object of `lm()`).

i) Use `predict()` to add a "confidence" and a "prediction" interval to the plot from (a). What is the difference?
Hint: The meanings of "confidence" and "predict" here are based on the R function. Use the help of those functions to understand their behaviour.

j) Fit a linear model without intercept (i.e., force β_0 to be zero). Add the corresponding regression line to the plot from (a). Discuss if the model fits "better" the data.

k) How do outliers influence the model fit? Add outliers:

```
data_outlier <- rbind(data, data.frame(x = c(10, 11, 12), y = c(9, 7, 8)))
```

Fit a linear model and discuss it (including diagnostic plots).

l) What is the difference between a model with formula `y ~ x` and `x ~ y`? Explain it from a stochastic and fitting perspective.

Problem 9.7 (BMJ Endgame) Discuss and justify the statements about 'Simple linear regression' given in doi.org/10.1136/bmj.f2340.

Chapter 10

Multiple Regression

Learning goals for this chapter:

- ⋄ Statistical model of multiple regression
- ⋄ Multiple regression in R, including:
  - Multicollinearity
  - Influential points
  - Interactions between variables
  - Categorical variables (factors)
  - Model validation and information criterion (basic theory and R)
- ⋄ Be aware of nonlinear regression - examples

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter10.R.

In many situations a dependent variable is associated with more than one independent variable. The simple linear regression model can be extended by the addition of further independent variables. We first introduce the model and estimators. Subsequently, we become acquainted with the most important steps in model validation. Two typical examples of multiple regression are given. At the end, several typical examples of extensions of linear regression are illustrated.

10.1 Model and Estimators

A natural extension of the simple linear regression model to p independent variables is as follows

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i, && (10.1)\\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, && (10.2)\\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad\qquad i = 1, \dots, n, \quad n > p, && (10.3)\end{aligned}$$

with

- Y_i: dependent variable, measured value, observation, response;
- $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$: (known) independent/explanatory variables, predictors;
- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$: (unknown) parameter vector, regression coefficients;
- ε_i: (unknown) error term, noise, with symmetric distribution around zero, $\mathrm{E}(\varepsilon_i) = 0$.

It is often also assumed that $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and/or that the errors are independent of each other. In matrix notation, Equation (10.3) is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{10.4}$$

with $\mathbf{X}$ an $n \times (p+1)$ matrix with rows $\boldsymbol{x}_i^\top$. The model is linear in $\boldsymbol{\beta}$ and often referred to as *multiple linear regression* model.

To derive estimators and obtain simple, closed form distributions for these, we assume that $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with unknown σ^2, also when simply referring to the model (10.3). With a Gaussian error term we have (in matrix notation)

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}), \tag{10.5}$$

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{10.6}$$

The mean of the response varies, implying that $Y_1, \dots, Y_n$ are only independent and not iid.

There are some constraints on the independent variables. We assume that the rank of $\mathbf{X}$ equals $p+1$ ($\mathrm{rank}(\mathbf{X}) = p+1$, column rank). This assumption guarantees that the inverse of $\mathbf{X}^\top\mathbf{X}$ exists. In practical terms, this implies that we do not include twice the same predictor, or that a predictor has additional information on top of the already included predictors.

The parameter vector $\boldsymbol{\beta}$ is estimated with the method of ordinary least squares (see Section 4.2.1). This means, that the estimate $\widehat{\boldsymbol{\beta}}$ is such that the sum of the squared errors (residuals) is minimal and is thus derived as follows:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{10.7}$$

$$\Longrightarrow \quad \frac{d}{d\boldsymbol{\beta}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{10.8}$$

$$= \frac{d}{d\boldsymbol{\beta}}(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}) = -2\mathbf{X}^\top\boldsymbol{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \tag{10.9}$$

$$\Longrightarrow \quad \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\boldsymbol{y} \tag{10.10}$$

$$\Longrightarrow \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \tag{10.11}$$

Equation (10.10) is also called the *normal equation* and Equation (10.11) indicates why we need to assume full column rank of the matrix $\mathbf{X}$.

We now derive the distributions of the estimator and other related and important vectors. The derivation of the results are based directly on Property 8.6. Starting from the distributional assumption of the errors (10.5), jointly with Equations (10.6) and (10.11), it can be shown that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}, \qquad \widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big), \tag{10.12}$$

$$\widehat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} = \mathbf{H}\boldsymbol{y}, \qquad \widehat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}), \tag{10.13}$$

$$\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y} \qquad \mathbf{R} \sim \mathcal{N}_n\big(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})\big), \tag{10.14}$$

where we term the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ as the *hat matrix*. In the left column we find the estimate, predicted value and the residuals, in the right column the according functions of the random sample and its distribution. Notice the subtle difference in the covariance matrix of the distributions $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$: the hat matrix $\mathbf{H}$ is not $\mathbf{I}$, hopefully quite close to it. The latter would imply that the variances of $\mathbf{R}$ are close to zero.

The distribution of the regression coefficients will be used for inference and when interpreting a fitted regression model (similarly as in the case of the simple regression). The marginal distributions of the individual coefficients $\widehat{\beta}_i$ are determined by the distribution (10.12):

$$\widehat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 v_{ii}) \quad \text{with } v_{ii} = \big((\mathbf{X}^\top\mathbf{X})^{-1}\big)_{ii}, \quad i = 0, \dots, p, \tag{10.15}$$

(again direct consequence of Property 8.6). Hence $(\widehat{\beta}_i - \beta_i)\big/\sqrt{\sigma^2 v_{ii}} \sim \mathcal{N}(0, 1)$. As σ^2 is unknown and we use a plug-in estimate, typically the unbiased estimate

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n-p-1}\boldsymbol{r}^\top\boldsymbol{r}, \tag{10.16}$$

which is again termed *mean squared error*. Its square root is termed *residual standard error* with $n - p - 1$ degrees of freedom. Finally, we use the same approach when deriving the t-test in Equation (5.4) and obtain

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\widehat{\sigma}^2 v_{ii}}} \sim t_{n-p-1} \tag{10.17}$$

as our statistic for testing $H_0 : \beta_i = \beta_{0,i}$. For testing, we are often interested in $H_0 : \beta_i = 0$ for which the test statistic reduces to the one shown in Test 15. Confidence intervals are constructed along equation (4.33) and summarized subsequently in CI 8.

CI 8: Confidence interval for regression coefficients

For the model (10.3), a sample $(1 - \alpha)$ confidence interval for β_i

$$\Big[\, \widehat{\beta}_i \pm t_{n-p-1,1-\alpha/2}\sqrt{\frac{1}{n-p-1}\boldsymbol{r}^\top\boldsymbol{r}\; v_{ii}} \,\Big] \tag{10.18}$$

with $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ and $v_{ii} = \big((\mathbf{X}^\top\mathbf{X})^{-1}\big)_{ii}$.

In R, the multiple regression is fitted with `lm()` again by adding the additional predictor to the right-hand-side of the formula statement. The summary output is very similar as for the simple regression, we simply add further lines to the coefficient block. The degrees of freedom of the numerator of the F-statistic changes from 1 to $p - 1$. The following example illustrates a regression with $p = 2$ predictors.

Example 10.1 (`abrasion` data)**.** The data comes from an experiment investigating how rubber's resistance to abrasion is affected by the hardness of the rubber and its tensile strength (Cleveland, 1993). Each of the 30 rubber samples was tested for hardness and tensile strength, and then subjected to steady abrasion for a fixed time.

R-Code 10.1 performs the regression analysis based on two predictors. The sample confidence intervals *confint(res)* do not contain zero. Accordingly, the p-values of the three t-tests are small.

We naively assumed a linear relationship between the predictors hardness and strength and the abrasion. The scatterplots in Figure 10.1 give some indication that hardness itself has a linear relationship with abrasion. The scatterplot only depict the marginal relations, i.e., the red curves would be linked to *lm(loss hardness, data=abrasion)* and *lm(loss strength, data=abrasion)* (the latter indeed shows a poor linear relationship). Similarly, a quadratic term for *strength* is not necessary (see, the summary of *lm(loss hardness+strength+I(strength^2), data=abrasion)*). ♣

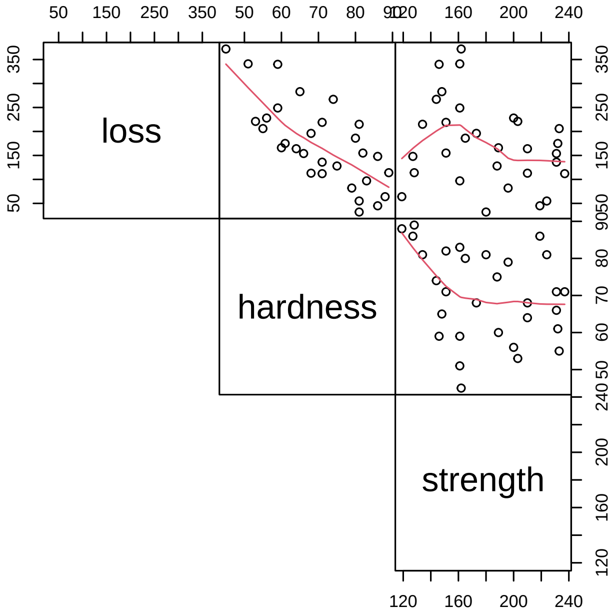


Figure 10.1: Pairs plot of *abrasion* data with red “guide-the-eye” curves. (See R-Code 10.1.)

R-Code 10.1: *abrasion* data: fitting a linear model. (See Figure 10.1.)

```
abrasion <- read.csv('data/abrasion.csv')
str(abrasion)

## 'data.frame': 30 obs. of  3 variables:
##  $ loss    : int  372 206 175 154 136 112 55 45 221 166 ...
##  $ hardness: int  45 55 61 66 71 71 81 86 53 60 ...
##  $ strength: int  162 233 232 231 231 237 224 219 203 189 ...

pairs(abrasion, upper.panel=panel.smooth, lower.panel=NULL, gap=0)
abres <- lm(loss~hardness+strength, data=abrasion)
summary( abres)
```

```
##
## Call:
## lm(formula = loss ~ hardness + strength, data = abrasion)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.38 -14.61   3.82  19.75  65.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  885.161     61.752   14.33  3.8e-14 ***
## hardness      -6.571      0.583  -11.27  1.0e-11 ***
## strength      -1.374      0.194   -7.07  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.5 on 27 degrees of freedom
## Multiple R-squared:  0.84,Adjusted R-squared:  0.828
## F-statistic:   71 on 2 and 27 DF,  p-value: 1.77e-11
confint( abres)
##                2.5 %     97.5 %
## (Intercept) 758.4573 1011.86490
## hardness     -7.7674   -5.37423
## strength     -1.7730   -0.97562
```

10.2 Model Validation

Recall the data analysis workflow shown in Figure 1.1. Suppose we have a *model fit* of a linear model (obtained with `lm()`, as illustrated in the last chapter or in the previous section). *Model validation* essentially verifies if (10.3) is an adequate model for the data to probe the given *Hypothesis to investigate*. The question is not whether a model is correct, but rather if the model is useful ("Essentially, all models are wrong, but some are useful", Box and Draper, 1987 page 424).

Model validation verifies (i) the fixed component (or fixed part) μ_i and (ii) the stochastic component (or stochastic part) ε_i and is typically an iterative process (arrow back to *Propose statistical model* in Figure 1.1).

10.2.1 Basics and Illustrations

Validation is based on (a) graphical summaries, typically (standardized) residuals versus fitted values, individual predictors, summary values of predictors or simply Q-Q plots and (b) summary statistics. The latter are also part of a `summary()` call of a regression object, see, e.g., R-Code 9.6. The residuals are summarized by the range and the quartiles. Below the summary of the coefficients, the following statistics are given

$$\text{Multiple } R^2\text{:}\quad R^2 = 1 - \frac{\text{SS}_E}{\text{SS}_T} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \tag{10.19}$$

$$\text{Adjusted } R^2\text{:}\quad R^2_{\text{adj}} = R^2 - (1 - R^2)\frac{p}{n-p-1}, \tag{10.20}$$

$$\text{Observed value of } F\text{-Test:}\quad \frac{(\text{SS}_T - \text{SS}_E)/p}{\text{SS}_E/(n-p-1)}, \tag{10.21}$$

were SS stands for *sums of squares* and SS_T, SS_E for *total sums of squares* and *sums of squares of the error*, respectively. The statistics are similar to the ones from a simple regression up to taking into account that we have now $p+1$ parameters to estimate.

The last statistic explains how much variability in the data is explained by the model and is essentially equivalent to Test 4 and performs the omnibus test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. When we reject this test, this merely signifies that at least one of the coefficients is significant and thus often not very useful.

A slightly more general version of the F-Test (10.21) is used to compare nested models. Let $\mathcal{M}_0$ be the simpler model with only q out of the p predictors of the more complex model $\mathcal{M}_1$ $(0 \leq q < p)$. The test H_0 : "$\mathcal{M}_0$ is sufficient" is based on the statistic

$$\frac{(\text{SS}_{\text{simple model}} - \text{SS}_{\text{complex model}})/(p-q)}{\text{SS}_{\text{complex model}}/(n-p-1)} \tag{10.22}$$

and often runs under the name ANOVA (analysis of variance). We see an alternative derivation thereof in Chapter 11.

In order to validate the fixed components of a model, it must be verified whether the necessary predictors are in the model. We do not want too many, nor too few. Unnecessary predictors are often identified through insignificant coefficients. When predictors are missing, the residuals show (in the ideal case) structure, indicative for model improvement. In other cases, the quality of the regression is low (F-Test, R^2 (too) small). Example 10.2 below will illustrate the most important elements.

Example 10.2. We construct synthetic data in order to better illustrate the difficulty of detecting a suitable model. Table 10.1 gives the actual models and the five fitted models. In all cases we use a small dataset of size $n = 50$ and predictors x_1 and $x_2 = x_1^2$) that we construct from a uniform distribution. Further, we set $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.25^2)$. R-Code 10.2 and the corresponding Figure 10.2 illustrate how model deficiencies manifest.

We illustrate how residual plots may or may not show missing or unnecessary predictors. Because of the 'textbook' example, the adjusted R^2 values are very high and the p-value of the F-Test is – as often in practice – of little value.

Since the output of *summary()* is quite long, here we show only elements from it. This is achieved with the functions *print()* and *cat()*. For Examples 2 to 5 the output has been constructed by a function call to *subset_of_summary()* constructing the output as the first example.

The plots should supplement a classical graphical analysis through *lm(res)*. ♣

Table 10.1: Fitted models for five examples of 10.2. The true model is always $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$.

Example	fitted model	
1	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$	correct model
2	$Y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$	missing predictor x_1^2
3	$Y_i = \beta_0 + \beta_2 x_2 + \varepsilon_i$	missing predictor x_1
4	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$	unnecessary predictor x_3

R-Code 10.2: Illustration of missing and unnecessary predictors for an artificial dataset. (See Figure 10.2.)

```
set.seed( 18)
n <- 50
x1 <- runif( n);     x2 <- x1^2;     x3 <- runif( n)
eps <- rnorm( n, sd=0.16)
y <- -1 + 3*x1 + 2.5*x1^2 + 1.5*x2 + eps
# Example 1: Correct model
sres <- summary( res <- lm( y ~ x1 + I(x1^2) + x2 ))
print( sres$coef, digits=2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.0      0.045     -23  7.9e-27
## x1               3.0      0.238      13  9.9e-17
## I(x1^2)          3.9      0.242      16  6.9e-21
cat("Adjusted R-squared: ", formatC( sres$adj.r.squared),
  "  F-Test: ", pf(sres$fstatistic[1], 1, n-2, lower.tail = FALSE))
## Adjusted R-squared:  0.9963   F-Test:  4.6376e-53
plotit( res$fitted, res$resid, "fitted values")  # Essentially equivalent to:
# plot( res, which=1, caption=NA, sub.caption=NA, id.n=0) with ylim=c(-1,1)
# i.e, plot(), followed by a panel.smooth()
plotit( x1, res$resid, bquote(x[1]))
plotit( x2, res$resid, bquote(x[2]))
# Example 2: Missing predictor x1^2
sres <- subset_of_summary( lm( y ~ x1 ))
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.5      0.081     -19  5.4e-24
## x1               6.7      0.151      45  8.6e-41
## Adjusted R-squared:  0.9761   F-Test:  8.5684e-41
# Example 3: Missing predictor x1
sres <- subset_of_summary( lm( y ~ x2 ))
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.54      0.052     -10  6.2e-14
## x2              6.88      0.125      55  5.1e-45
## Adjusted R-squared:  0.9841   F-Test:  5.0725e-45
# Example 4: Too many predictors x3
sres <- subset_of_summary( lm( y ~ x1 + x2 + x3))
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.976      0.060   -16.2  1.3e-20
## x1             2.976      0.239    12.5  2.3e-16
## x2             3.938      0.242    16.3  1.0e-20
## x3            -0.069      0.065    -1.1  2.9e-01
## Adjusted R-squared:  0.9963   F-Test:  6.6892e-49
# The results are in sync with:
# tmp <- cbind( y, x1, "x1^2"=x1^2, x2, x3)
# pairs( tmp, upper.panel=panel.smooth, lower.panel=NULL, gap=0)
# cor( tmp)
```

It is important to understand that the stochastic part ε_i does not only represent measurement error. In general, the error is the remaining "variability" (also noise) that is not explained through the predictors ("signal").

With respect to the stochastic part ε_i, the following points should be verified:

1. constant variance: if the (absolute) residuals are displayed as a function of the predictors, the estimated values, or the index, no structure should be discernible. The observations can often be transformed in order to achieve constant variance. Constant variance is also called homoscedasticity and the terms heteroscedasticity or variance heterogeneity are used otherwise.

   More precisely, for heteroscedacity we relax Model (10.3) to $\varepsilon_i \overset{\text{indep}}{\sim} \mathcal{N}(0, \sigma_i^2)$, i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{V})$. In case the diagonal matrix $\mathbf{V}$ is known, we can use so-called weighted least squares (WLS) by considering the argument `weights` in R (`weights=1/diag(V)`).

2. independence: correlation between the residuals should be negligible.

If data are taken over time, observations might be serially correlated or dependent. That means, $\text{Corr}(\varepsilon_{i-1}, \varepsilon_i) \neq 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ which is not a diagonal matrix. This is easy to test and to visualize through the residuals, illustrated in Example 10.3.

If $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} = \sigma^2\mathbf{R}$ where $\mathbf{R}$ is a known correlation matrix a so-called generalized least squares (GLS) approach can be used. Correlated data are extensively discussed in time series analysis and in spatial analysis. We refer to follow-up lectures for a more detailed exhibition.

3. symmetric distribution: it is not easy to find evidence against this assumption. If the distribution is strongly right- or left-skewed, the scatter plots of the residuals will have structure. Transformations or generalized linear models may help. We have a quick look at a generalized linear model in Section 10.4.

Example 10.3 (continuation of Example 10.1). R-Code 10.3 constructs a few diagnostic plots shown in Figure 10.3.

A quadratic term for *strength* is not necessary. However, the residuals appear to be slightly correlated (see bottom right panel, *cor(abres$resid[-1],abres$resid[-30])* is 0.53). We do not have further information about the data here and cannot investigate this aspect further. ♣

R-Code 10.3: *abrasion* data: model validation. (See Figure 10.3.)

```r
# Fitted values
plot( loss~hardness, ylim=c(0,400), yaxs='i', data=abrasion)
points( abres$fitted~hardness, col=4, data=abrasion)
plot( loss~strength, ylim=c(0,400), yaxs='i', data=abrasion)
points( abres$fitted~strength, col=4, data=abrasion)
# Residuals vs ...
plot( abres$resid~abres$fitted)
lines( lowess( abres$fitted, abres$resid), col=2)
abline( h=0, col='gray')
plot( abres$resid~hardness, data=abrasion)
lines( lowess( abrasion$hardness, abres$resid), col=2)
abline( h=0, col='gray')
plot( abres$resid~strength, data=abrasion)
lines( lowess( abrasion$strength, abres$resid), col=2)
abline( h=0, col='gray')
plot( abres$resid[-1]~abres$resid[-30])
abline( h=0, col='gray')
```

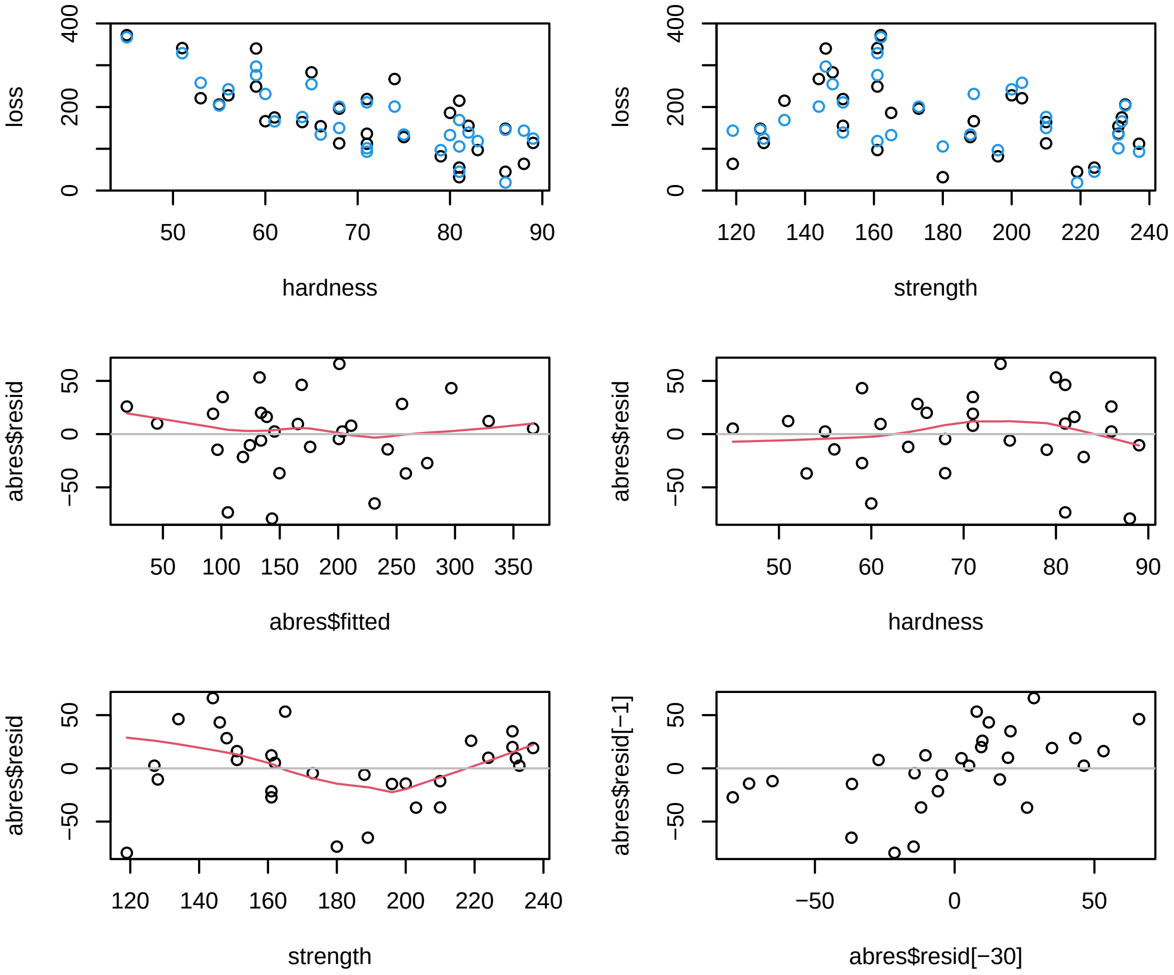


Figure 10.3: `abrasion` data: model validation. Top row shows the loss (black) and fitted values (blue) as a function of hardness (left) and strength (right). Middle and bottom panels are different residual plots. (See R-Code 10.3.)

10.3 Model Selection

An information criterion in statistics is a tool for model selection. It follows the idea of *Occam's razor*, in that a model should not be unnecessarily complex. It balances the goodness of fit of the estimated models with its complexity, measured by the number of parameters. There is a penalty for the number of parameters, otherwise complex models with numerous parameters would be preferred.

The coefficient of determination R^2 is thus not a proper information criterion: every additional predictor that is added to the model will potentially reduce SS_M. The adjusted R^2 is somewhat a information criterion as the second term increases with predictors that do not contribute to reduce the residual sums of squares.

To introduce two well known information criterion, we assume that the distribution of the observations follows a known distribution with an unknown parameter $\boldsymbol{\theta}$ with p components. Hence, we can write down the likelihood function of the observations. In maximum likelihood estimation, the larger the likelihood function $L(\widehat{\boldsymbol{\theta}})$ or, equivalently, the smaller the negative log-likelihood function $-\ell(\widehat{\boldsymbol{\theta}})$, the better the model is. The oldest criterion was proposed as "an information criterion" in 1973 by Hirotugu Akaike (Akaike, 1973) and is known today as the *Akaike information criterion* (AIC):

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p. \tag{10.23}$$

In regression models with normally distributed errors, the maximized log-likelihood is $n\log(\widehat{\sigma}^2)$ up to some constant and so the first term describes the goodness of fit.

It is important to not that additive constants are not relevant. Also solely the difference of two AICs make sense but not the reduction in proportion or something.

Remark 10.1. When deriving the AIC for the regression setting, we have to use likelihood estimates. $\widehat{\boldsymbol{\beta}}_{\text{ML}} = \widehat{\boldsymbol{\beta}}_{\text{LS}}$ and $\widehat{\sigma}^2_{\text{ML}} = \boldsymbol{r}^\top\boldsymbol{r}/n$. The log-likelihood is then shown to be $\ell(\widehat{\boldsymbol{\beta}}_{\text{ML}}, \widehat{\sigma}^2_{\text{ML}}) = -n/2(\log(2*pi) - \log(n) + \log(\boldsymbol{r}^\top\boldsymbol{r}) + 1)$. As additive constants are not relevant for the AIC, we can express the AIC in terms of the residuals only. ♣

The disadvantage of AIC is that the penalty term is independent of the sample size. The *Bayesian information criterion* (BIC)

$$\text{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + \log(n)\,p \tag{10.24}$$

penalizes the model more heavily based on both the number of parameters p and sample size n, and its use is recommended.

We illustrate the use of information criteria with an example based on a classical dataset.

Example 10.4 (`LifeCycleSavings` data)**.** Under the life-cycle savings hypothesis developed by Franco Modigliani, the savings ratio (aggregate personal savings divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old (see, e.g., Modigliani, 1966). The data provided by `LifeCycleSavings` are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations and contains information from 50 countries about these five variables:

- `sr` aggregate personal savings,
- `pop15` % of population under 15,
- `pop75` % of population over 75,
- `dpi` real per-capita disposable income,
- `ddpi` % growth rate of dpi.

Scatter plots are shown in Figure 10.4. R-Code 10.4 fits a multiple linear model, selects models through comparison of various goodness of fit criteria (AIC, BIC) and shows the model validation plots for the model selected using AIC. The `step()` function is a convenient way for selecting relevant predictors. Figure 10.4 gives four of the most relevant diagnostic plots, obtained by passing a fitted object to `plot()` (compare with the manual construction of Figure 10.1).

Different models may result from different criteria: when using BIC for model selection, `pop75` drops out of the model. ♣

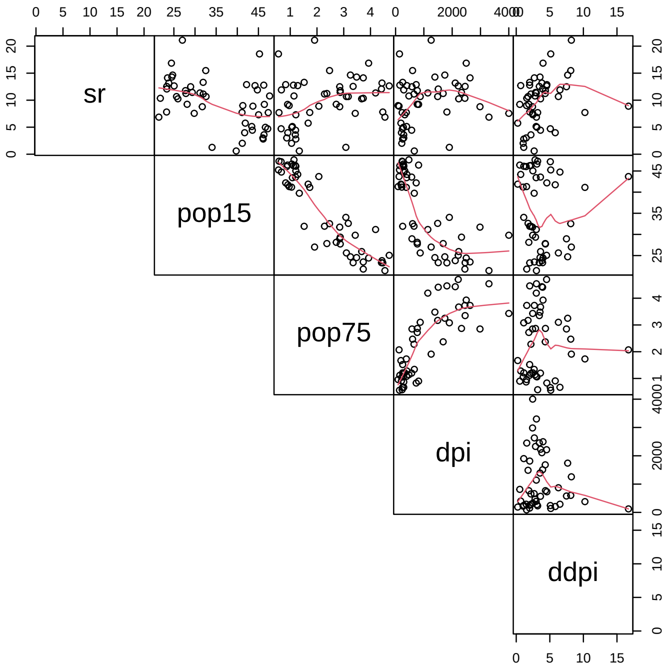


Figure 10.4: Scatter plots of `LifeCycleSavings` with red “guide-the-eye” curves. (See R-Code 10.4.)

R-Code 10.4: `LifeCycleSavings` data: EDA, linear model and model selection. (See Figures 10.4 and 10.5.)

```
data( LifeCycleSavings)
head( LifeCycleSavings)
##                sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
pairs(LifeCycleSavings, upper.panel=panel.smooth, lower.panel=NULL, gap=0)
lcs.all <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
summary( lcs.all)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.242 -2.686 -0.249  2.428  9.751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.566087   7.354516    3.88  0.00033 ***
## pop15       -0.461193   0.144642   -3.19  0.00260 **
## pop75       -1.691498   1.083599   -1.56  0.12553
## dpi         -0.000337   0.000931   -0.36  0.71917
## ddpi         0.409695   0.196197    2.09  0.04247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.8 on 45 degrees of freedom
## Multiple R-squared:  0.338,Adjusted R-squared:  0.28
## F-statistic: 5.76 on 4 and 45 DF,  p-value: 0.00079
```

```
lcs.aic <- step( lcs.all)       # AIC is default choice
```

```
## Start:  AIC=138.3
## sr ~ pop15 + pop75 + dpi + ddpi
##
##         Df Sum of Sq RSS AIC
## - dpi    1       1.9 653 136
## <none>                651 138
## - pop75  1      35.2 686 139
## - ddpi   1      63.1 714 141
## - pop15  1     147.0 798 146
##
## Step:  AIC=136.45
## sr ~ pop15 + pop75 + ddpi
##
##         Df Sum of Sq RSS AIC
## <none>                653 136
## - pop75  1      47.9 701 138
## - ddpi   1      73.6 726 140
## - pop15  1     145.8 798 144
```

```
summary( lcs.aic)$coefficients
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 28.12466    7.18379  3.9150 0.00029698
## pop15       -0.45178    0.14093 -3.2056 0.00245154
## pop75       -1.83541    0.99840 -1.8384 0.07247270
## ddpi         0.42783    0.18789  2.2771 0.02747818
plot( lcs.aic)                   # 4 plots to assess the models
summary( step( lcs.all, k=log(50), trace=0))$coefficients  # now BIC
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 15.59958   2.334394  6.6825 2.4796e-08
## pop15       -0.21638   0.060335 -3.5863 7.9597e-04
## ddpi         0.44283   0.192401  2.3016 2.5837e-02
```

10.4 Extensions of the Linear Model

Naturally, the linearity assumption (of the individual parameters) in the linear model is central. There are situations, where we have a nonlinear relationship between the observations and the parameters. In this section we enumerate some of the alternatives, such that at least relevant statistical terms can be associated.

10.4.1 Logistic Regression

The response Y_i in Model (10.3) is Gaussian, albeit not iid. If this is not the case, for example the observations y_i are proportions, then we need another approach. The idea of this new approach is to model a function of the expected value of Y_i as a linear function of some predictors, for example $g(\mathrm{E}(Y_i)) = \beta_0 + \beta_1 x_i$. The function $g(\cdot)$ is such that for any value x_i, $g^{-1}(\beta_0 + \beta_1 x_i)$ is constrained to $[0,]$. To illustrate the concept we look at a simple logistic regression. Example 10.5 is based on widely discussed data.

Example 10.5 (*`orings`* data)**.** In January 1986 the space shuttle Challenger exploded shortly after taking off, killing all seven crew members aboard. Part of the problem was with the rubber seals, the so-called o-rings, of the booster rockets. Due to low ambient temperature, the seals started to leak causing the catastrophe. The data set *`data( orings, package="faraway")`* contains the number of defects in the six seals in 23 previous launches (Figure 10.6). The question we ask here is whether the probability of a defect for an arbitrary seal can be predicted for an air temperature of 31°F (as in January 1986). See Dalal *et al.* (1989) for a detailed statistical account or simply https://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster.

The variable of interest is a probability (failure of a rubber seal), that we estimate based on binomial data (failures of o-rings) but a linear model cannot guarantee $\widehat{p}_i \in [0, 1]$ (see linear fit in Figure 10.6). In this and similar cases, logistic regression is appropriate. The logistic regression

R-Code 10.5 *orings* data and estimated probability of defect dependent on air temperature. (See Figure 10.6.)

```
data( orings, package="faraway")
str(orings)
## 'data.frame': 23 obs. of  2 variables:
##  $ temp  : num  53 57 58 63 66 67 67 67 68 69 ...
##  $ damage: num  5 1 1 1 0 0 0 0 0 0 ...
plot( damage/6~temp, xlim=c(21,80), ylim=c(0,1), data=orings, pch='+',
     xlab="Temperature [F]", ylab='Probability of damage') # data
abline(lm(damage/6~temp, data=orings), col='gray') # regression line

glm1 <- glm( cbind(damage,6-damage)~temp, family=binomial, data=orings)
points( orings$temp, glm1$fitted, col=2) # fitted values
ct <- seq(20, to=85, length=100)         # vector to predict
p.out <- predict( glm1, new=data.frame(temp=ct), type="response")
lines(ct, p.out)
abline( v=31, col='gray', lty=2)         # actual temp. at start
```

models the probability of a defect as

$$p = \mathrm{P}(\text{defect}) = \frac{1}{1+\exp(-\beta_0-\beta_1 x)}\,, \tag{10.25}$$

where x is the air temperature. Through inversion one obtains a linear model for the log odds

$$g(p) = \log\Bigl(\frac{p}{1-p}\Bigr) = \beta_0 + \beta_1 x, \tag{10.26}$$

where $g(\cdot)$ is generally called the link function. In this special case, the function $g^{-1}(\cdot)$ is called the logistic function. ♣

10.4.2 Generalized Linear Regression

The case of logistic regression can be extended to the so-called *generalized linear model* (GLM). A GLM contains as a special case the classical linear regression and the logistic regression.

The underlying idea is to model a transformation of the expected value of Y_i as a linear function of some predictors. Lets assume that Y_i are independent and from a certain class of distribution and denote $\mathrm{E}(Y_i) = \mu_i$ then $g(\mathrm{E}(Y_i)) = \boldsymbol{x}_i\boldsymbol{\beta}$. In the linear regression setting, the response is Gaussian with $g(x) = x$, in the logistic regression, the response is Bernoulli (or equivalently Binomial) with $g(x) = \log(p/(1-p))$. Alternative possible distributions are Poisson, log-normal, or gamma (the latter is an extension of the exponential or chi-squared distribution).

The advantage of a GLM approach is the common fitting and estimation procedure unified in the `glm()` function, where one simply specifies the distribution and link function via the argument `family=...(link="...")`. While the fitting via R is straightforward, the interpretation of the parameters and model validation are more involved compared to linear regression.

10.4.3 Transformation of the Response

For a Poisson random variable, the variance increases with increasing mean. Similarly, it is possible that the variance of the residuals increases with an increasing number of observations. Instead of "modeling" increasing variances, transformations of the response variables often render the variances of the residuals sufficiently constant. For example, instead of a linear model for Y_i, a linear model for $\log(Y_i)$ is constructed.

Typical transformations are $\log(x)$, $\log(x+1)$, $\sqrt{x}$ or a general power transformation x^λ. We may pick "any" other reasonable transformation. There are formal approaches to determine an optimal transformation of the data, notably the function `boxcox()` from the package *MASS* (see Problem 10.2). However, in practice $\log(\cdot)$ and $\sqrt{\cdot}$ are used dominantly.

It is important to realize that if the original data stems from a "truly" linear model, any non-linear transformation leads to an approximation. On the other hand, a log-transformation of the true model

$$Y_i = \beta_0\, x_i^{\beta_1}\, \mathrm{e}^{\varepsilon_i} \tag{10.27}$$

leads to a linear relationship

$$\log(Y_i) = \log(\beta_0) + \beta_1 \log(x_i) + \varepsilon_i. \tag{10.28}$$

10.4.4 Nonlinear and Non-parametric Regression

A natural generalization of model (10.3) is to relax the linearity assumption on $\boldsymbol{\beta}$ and write

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}), +\varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{10.29}$$

where $f(\boldsymbol{x}, \boldsymbol{\beta})$ is a (sufficiently well behaved) function depending on a vector of covariates $\boldsymbol{x}$ and the parameter vector $\boldsymbol{\beta}$. To estimate the latter we use a least squares approach

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})\big)^2. \tag{10.30}$$

For most functions f, we do not have closed forms for the resulting estimates and iterative approaches are needed. That is, starting from an initial condition we improve the solution by small steps. The correction factor typically depends on the gradient of f. Such algorithms are variants of the so-called *Gauss–Newton algorithm*. The details of these are beyond the scope of this document.

There are some prototype functions $f(\cdot, \boldsymbol{\beta})$ that are often used:

$$Y_i = \beta_1 \exp\big(-(x_i/\beta_2)^{\beta_3}\big) + \varepsilon_i \qquad \text{Weibull model,} \tag{10.31}$$

$$Y_i = \begin{cases} \beta_1 + \varepsilon_i, & \text{if } x_i < \beta_3 \\ \beta_2 + \varepsilon_i, & \text{if } x_i \geq \beta_3, \end{cases} \qquad \text{break point models.} \tag{10.32}$$

The exponential growth model is a special case of the Weibull model.

There is no guarantee that an optimal solution exists (global minimum). Using nonlinear least squares as a black box approach is dangerous and should be avoided. In R, the function `nls()` (nonlinear least-squares) can be used and general optimizer functions are `nlm()` and `optim()`.

An alternative way to express a response in a nonlinear way is through a virtually arbitrary flexible function, similar to the (red) guide-the-eye curves in many of the scatter plots. These smooth curves are constructed through a so-called "non-parametric" regression approach.

10.5 Bibliographic Remarks

The book by Faraway (2006) nicely summarizes extensions of the linear model.

10.6 Exercises and Problems

Problem 10.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Derive the distributions of the fitted values and the residuals, i.e., (10.13) and (10.14).

b) Identify the "simple" and "complex model" in the setting of a simple regression such that (10.22) reduces to (9.21).

c) Show that the AIC for a standard multiple regression model with $p+1$ regression coefficients is proportional to $n\log(\boldsymbol{r}^\top\boldsymbol{r})+2p$ and derive the BIC for the same model.

Problem 10.2 (Box–Cox transformation) In this problem, we motivate the so-called Box–Cox transformation. Suppose that we have a random variable Y with strictly positive mean $\mu>0$ and standard deviation $\sigma=c\mu^\alpha$, where $c>0$ and α arbitrary. That means the standard deviation of Y is proportional to a power of the mean.

a) Define the random variable $X=Y^\lambda$, $\lambda\neq 0$, and show that the standard deviation of X is approximately proportional to $\mu^{\lambda-1+\alpha}$. What value λ of the transformation leads to a constant variance?

b) Suppose we observe from $Y_1,\dots,Y_n$ with respective means $\mu_1,\dots,\mu_n$. Describe a way to estimate the transformation parameter λ.

c) The dataset *`SanMartinoPPts`* from the package *`hydroTSM`* contains daily precipitation at station San Martino di Castrozza (Trento Italy) over 70 years. Determine the optimal transformation for monthly totals. Justify that a square root transformation is adequate. How do you expect the transformation to change when working with annual or daily data?

Hint: monthly totals may be obtained via *`hydroTSM::daily2monthly(SanMartinoPPts, FUN=sum)`*.

Problem 10.3 (Multiple linear regression 1) The data *`stackloss.txt`* are available on the course web page. The data represents the production of nitric acid in the process of oxidizing ammonia. The response variable, stack loss, is the percentage of the ingoing ammonia that escapes unabsorbed. Key process variables are the airflow, the cooling water temperature (in degrees C), and the acid concentration (in percent).

Construct a regression model that relates the three predictors to the response, stack loss. Check the adequacy of the model.

Exercise and data are from B. Abraham and J. Ledolter, *Introduction to Regression Modeling*, 2006, Thomson Brooks/Cole.

Hints:

- Look at the data. Outliers?
- Try to find a "optimal" model. Exclude predictors that do not improve the model fit.
- Use model Diagnostics, use $t-$, F-tests and (adjusted) R^2 values to compare different models.
- Which data points have a (too) strong influence on the model fit? (`influence.measures()`)
- Are the predictors correlated? In case of a high correlation, what are possible implications?

Problem 10.4 (Multiple linear regression 2) The file `salary.txt` contains information about average teacher salaries for 325 school districts in Iowa. The variables are

`District`	name of the district
`districtSize`	size of the district: 1 = small (less than 1000 students) 2 = medium (between 1000 and 2000 students) 3 = large (more than 2000 students)
`salary`	average teacher salary (in dollars)
`experience`	average teacher experience (in years)

a) Produce a pairs plot of the data and briefly describe it.
Hint: `districtSize` is a categorical random variable with only three possible outcomes.

b) For each of the three district sizes, fit a linear model using `salary` as the dependent variable and experience as the covariate. Is there an effect of experience? How can we compare the results?

c) We now use all data jointly and use `districtSize` as covariate as well. However, `districtSize` is not numerical, rather categorical and thus we set `mydata$districtSize <- as.factor( mydata$districtSize)` (with appropriate dataframe name). Fit a linear model using `salary` as the dependent variable and the remaining data as the covariates. Is there an effect of experience and/or district size? How can we interpret the parameter estimates?

Problem 10.5 (Missing or unnecessary predictors) Construct synthetic data according to $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon_i$ with $n = 50$, predictors x_1, x_2 and x_3 drawn from a uniform distribution and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.25^2)$. Fit models according to the table below and discuss the model deficiencies.

Hint: You may consider R-Code 10.2 and Figure 10.2.

Example	fitted model	
1	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon_i$	correct model
2	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$	missing predictor x_1^2
3	$Y_i = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2 + \varepsilon_i$	missing predictor x_1
4	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i$	missing predictor x_2
5	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \varepsilon_i$	unnecessary predictor x_3

Problem 10.6 (BMJ Endgame) Discuss and justify the statements about ‘Multiple regression’ given in doi.org/10.1136/bmj.f4373.

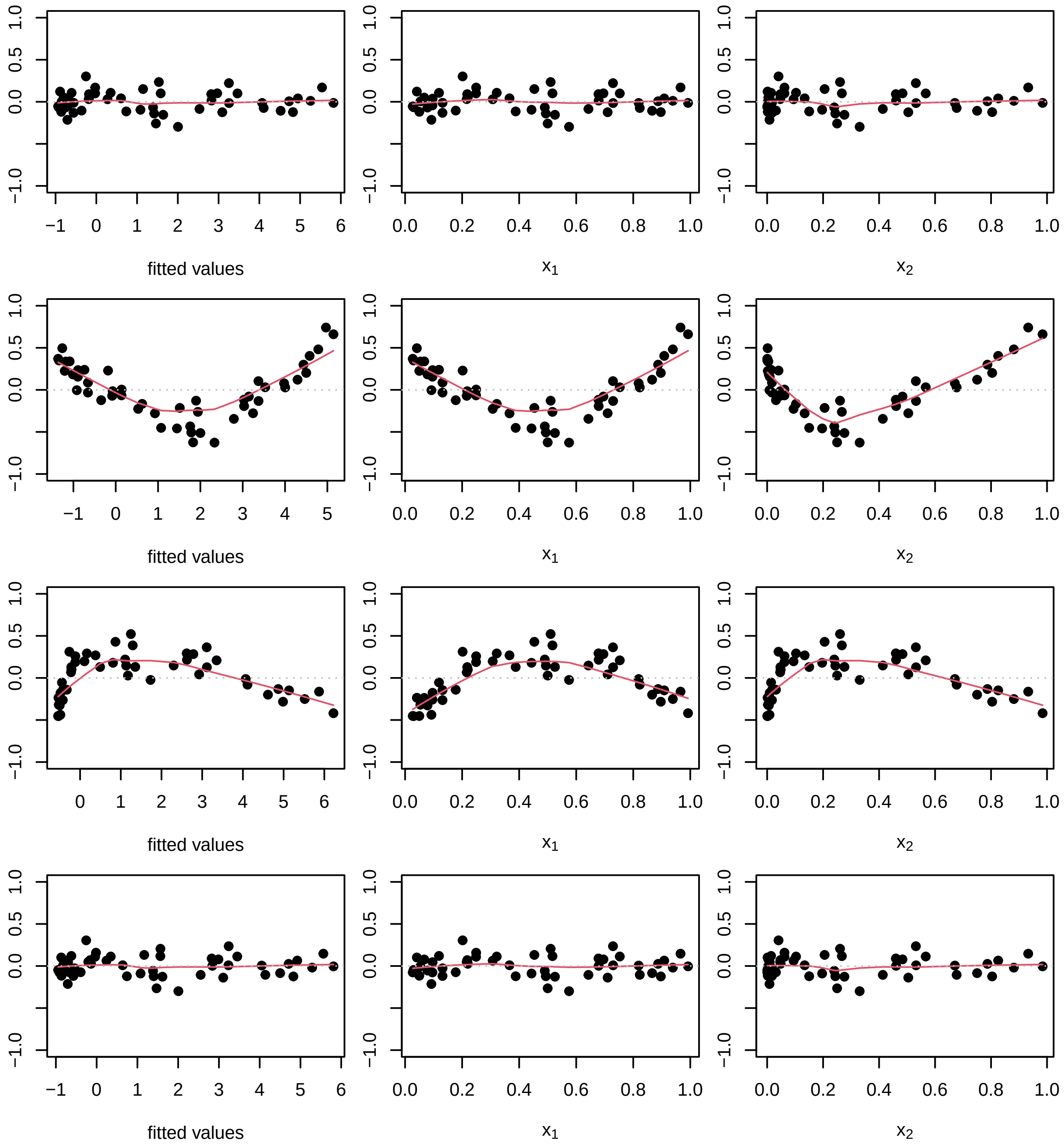


Figure 10.2: Residual plots. Residuals versus fitted values (left column), predictor x_1 (middle) and x_2 (right column). The rows correspond to the different fitted models. The panels in the left column have different scaling of the x-axis. (See R-Code 10.2.)

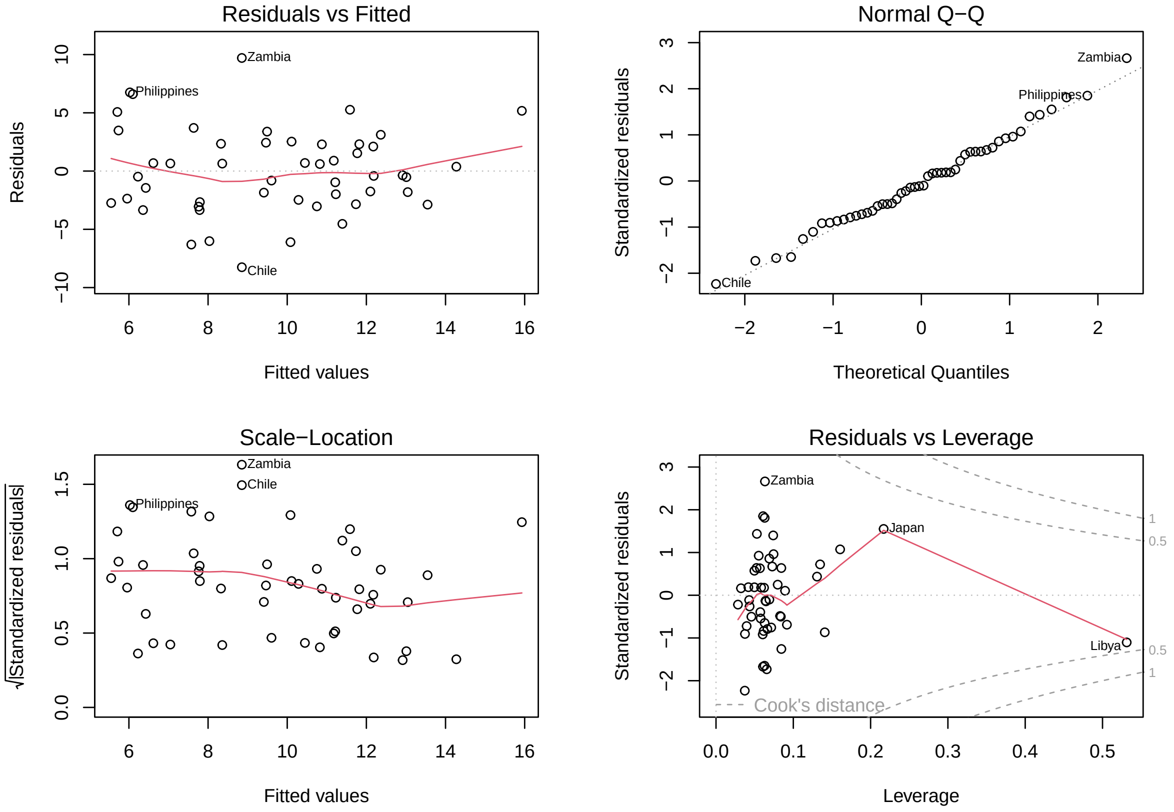


Figure 10.5: *LifeCycleSavings* data: model validation. (See R-Code 10.4.)

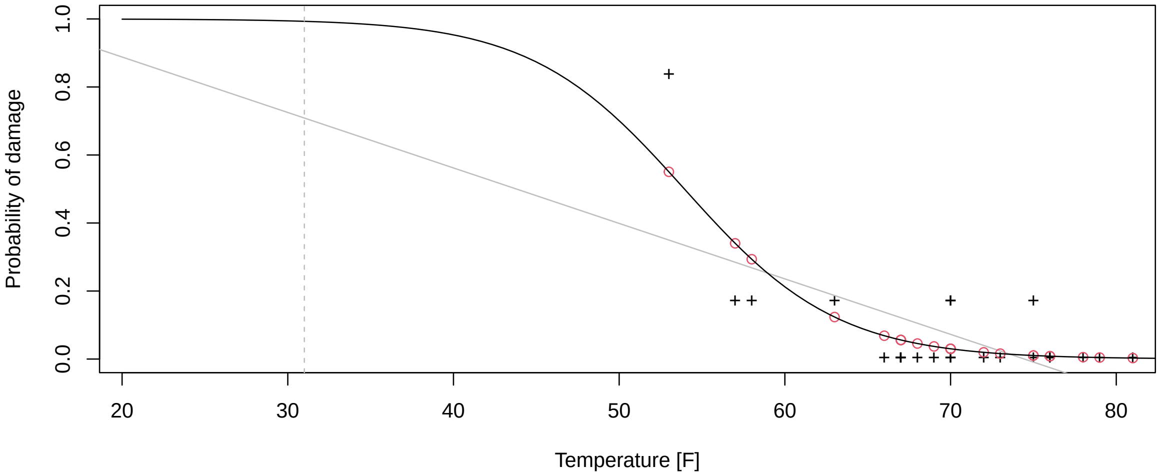


Figure 10.6: `orings` data (proportion of damaged orings, black crosses) and estimated probability of defect (red dots) dependent on air temperature. Linear fit is given by the gray solid line. Dotted vertical line is the ambient launch temperature at the time of launch. (See R-Code 10.5.)

Chapter 11

Analysis of Variance

Learning goals for this chapter:

- Understanding the link between t-tests, linear model and ANOVA
- Define an ANOVA model
- Understand the concept of sums of squares decomposition
- Interpret and discuss the output of a two-way ANOVA
- Define an ANCOVA model for a specific setting

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter11.R.

In this Chapter we will further elaborate on a sums of squares decomposition of linear models. For the ease of presentation, we focus on qualitative predictors, called factors. The simplest setting is comparing the means of I independent samples with a common variance term. This is much in the spirit of Test 2 discussed in Chapter 4, where we compared the means of two independent samples with each other.

Instead of comparing the samples pairwise (which would amount to $\binom{I}{2}$ tests and would require adjustments due to multiple testing as discussed in Section 5.5.2) we introduce a "better" method. Although we are concerned with comparing means we frame the problem to comparing variances, termed analysis of variance (ANOVA). We focus on a linear model approach to ANOVA. Due to historical reasons, the notation is slightly different than what we have seen in the last two chapters; but we try to link and unify as much as possible.

11.1 One-Way ANOVA

The model of this section is tailored to compare I different groups where the variability of the observations around the mean is the same in all groups. That means that there is one common variance parameter and we pool the information across all observations to estimate it.

More formally, the model consists of I groups, in the ANOVA context called factor levels, $i = 1, \dots, I$, and every level contains a sample of size n_i. Thus, in total we have $N = n_1 + \dots + n_I$ observations. The model is given by

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} && (11.1)\\ &= \mu + \beta_i + \varepsilon_{ij}, && (11.2) \end{aligned}$$

where we use the indices to indicate the group and within group observation. Similarly as in the regression models of the last chapters, we again assume $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Formulation (11.1) represents the individual group means directly, whereas formulation (11.2) models an overall mean and deviations from the mean.

However, model (11.2) is overparameterized (I levels and $I+1$ parameters) and an additional constraint on the parameters is necessary. Often, the sum-to-zero-contrast or treatment contrast, written as:

$$\sum_{i=1}^{I} \beta_i = 0 \qquad \text{or} \qquad \beta_1 = 0, \tag{11.3}$$

are used.

We are inherently interested in whether there exists a difference between the groups and so our null hypothesis is $H_0 : \beta_1 = \beta_2 = \dots = \beta_I = 0$. Note that the hypothesis is independent of the constraint. To develop the associated test, we proceed in several steps. We first link the two group case to the notation from the last chapter. In a second step we intuitively derive estimates in the general setting. Finally, we state the test statistic.

11.1.1 Two Level ANOVA Written as a Regression Model

Model (11.2) with $I = 2$ and treatment constraint $\beta_1 = 0$ can be written as a regression problem

$$Y_i^* = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*, \qquad i = 1, \dots, N \tag{11.4}$$

with Y_i^* the components of the vector $(Y_{11}, Y_{12}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2})^\top$ and $x_i = 0$ if $i = 1, \dots, n_1$ and $x_i = 1$ otherwise. We simplify the notation and spare ourselves from writing the index denoted by the asterisk with

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \boldsymbol{\varepsilon} \tag{11.5}$$

and thus we have as least squares estimate

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} N & n_2 \\ n_2 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}^\top & \mathbf{1}^\top \\ \mathbf{0}^\top & \mathbf{1}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} && (11.6)\\ &= \frac{1}{n_1 n_2} \begin{pmatrix} n_2 & -n_2 \\ -n_2 & N \end{pmatrix} \begin{pmatrix} \sum_{ij} y_{ij} \\ \sum_j y_{2j} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1}\sum_j y_{1j} \\ \frac{1}{n_2}\sum_j y_{2j} - \frac{1}{n_1}\sum_j y_{1j} \end{pmatrix}. && (11.7) \end{aligned}$$

Thus the least squares estimates of μ and β_2 in (11.2) for two groups are the mean of the first group and the difference between the two group means.

The null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ in Model (11.2) is equivalent to the null hypothesis $H_0 : \beta_1^* = 0$ in Model (11.4) or to the null hypothesis $H_0 : \beta_1 = 0$ in Model (11.5). The latter is of course based on a t-test for a linear association (Test 15) and coincides with the two-sample t-test for two independent samples (Test 2).

Estimators can also be derived in a similar fashion under other constraints or for more factor levels.

11.1.2 Sums of Squares Decomposition

Historically, we often had the case $n_1 = \cdots = n_I = J$, representing a *balanced* setting. In this case, there is another simple approach to deriving the estimators under the sum-to-zero constraint. We use dot notation in order to show that we work with averages, for example,

$$\overline{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad \text{and} \quad \overline{y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}. \tag{11.8}$$

Based on $Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ we use the following approach to derive estimates

$$y_{ij} = \overline{y}_{\cdot\cdot} + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}) + (y_{ij} - \overline{y}_{i\cdot}). \tag{11.9}$$

With the least squares method, $\widehat{\mu}$ and $\widehat{\beta}_i$ are chosen such that

$$\sum_{i,j}(y_{ij} - \widehat{\mu} - \widehat{\beta}_i)^2 = \sum_{i,j}(\overline{y}_{\cdot\cdot} + \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} + y_{ij} - \overline{y}_{i\cdot} - \widehat{\mu} - \widehat{\beta}_i)^2 \tag{11.10}$$

$$= \sum_{i,j}\big((\overline{y}_{\cdot\cdot} - \widehat{\mu}) + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} - \widehat{\beta}_i) + (y_{ij} - \overline{y}_{i\cdot})\big)^2 \tag{11.11}$$

is minimized. We evaluate the square of this last equation and note that the cross terms are zero since

$$\sum_{j=1}^{J}(y_{ij} - \overline{y}_{i\cdot}) = 0 \quad \text{and} \quad \sum_{i=1}^{I}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} - \widehat{\beta}_i) = 0 \tag{11.12}$$

(the latter due to sum-to-zero constraint) and thus $\widehat{\mu} = \overline{y}_{\cdot\cdot}$ and $\widehat{\beta}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$. Hence, writing $r_{ij} = y_{ij} - \overline{y}_{i\cdot}$, we have

$$y_{ij} = \widehat{\mu} + \widehat{\beta}_i + r_{ij}. \tag{11.13}$$

The observations are orthogonally projected in the space spanned by μ and β_i. This orthogonal projection allows for the division of the sums of squares of the observations (mean corrected to be precise) into the sums of squares of the model and sum of squares of the error component. These sums of squares are then weighted and compared. The representation of this process in table form and the subsequent interpretation is often equated with the analysis of variance, denoted ANOVA.

Remark 11.1. This orthogonal projection also holds in the case of a classical regression framework, of course. Using (10.13) and (10.14), we have

$$\widehat{\boldsymbol{y}}^\top \boldsymbol{r} = \boldsymbol{y}^\top \mathbf{H}^\top (\mathbf{I} - \mathbf{H})\boldsymbol{y} = \boldsymbol{y}^\top (\mathbf{H} - \mathbf{H}\mathbf{H})\boldsymbol{y} = 0, \tag{11.14}$$

because the hat matrix $\mathbf{H}$ is symmetric ($\mathbf{H}^\top = \mathbf{H}$) and idempotent ($\mathbf{HH} = \mathbf{H}$). ♣

The decomposition of the sums of squares can be derived with help from (11.9). No assumptions about constraints or n_i are made

$$\sum_{i,j}(y_{ij} - \overline{y}_{\cdot\cdot})^2 = \sum_{ij}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} + y_{ij} - \overline{y}_{i\cdot})^2 \tag{11.15}$$

$$= \sum_{ij}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 + \sum_{i,j}(y_{ij} - \overline{y}_{i\cdot})^2 + \sum_{i,j} 2(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})(y_{ij} - \overline{y}_{i\cdot}), \tag{11.16}$$

where the cross term is again zero because $\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot}) = 0$. Hence we have the decomposition of the sums of squares

$$\underbrace{\sum_{i,j}(y_{ij} - \overline{y}_{\cdot\cdot})^2}_{\text{Total}} = \underbrace{\sum_{i,j}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}_{\text{Model}} + \underbrace{\sum_{i,j}(y_{ij} - \overline{y}_{i\cdot})^2}_{\text{Error}} \tag{11.17}$$

or $\text{SS}_T = \text{SS}_A + \text{SS}_E$. We choose deliberately SS_A instead of SS_M as this will simplify subsequent extensions. Using the least squares estimates $\widehat{\mu} = \overline{y}_{\cdot\cdot}$ and $\widehat{\beta}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$, this equation can be read as

$$\frac{1}{N}\sum_{i,j}(y_{ij} - \widehat{\mu})^2 = \frac{1}{N}\sum_{i} n_i(\widehat{\mu + \beta_i} - \widehat{\mu})^2 + \frac{1}{N}\sum_{i,j}(y_{ij} - \widehat{\mu + \beta_i})^2 \tag{11.18}$$

$$\widehat{\text{Var}(y_{ij})} = \frac{1}{N}\sum_{i} n_i\widehat{\beta}_i^{\,2} + \widehat{\sigma^2}, \tag{11.19}$$

(where we could have used some divisor other than N). The test statistic for the statistical hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I = 0$ is based on the idea of decomposing the variance into variance between groups and variance within groups, just as illustrated in (11.19), and comparing them. Formally, this must be made more precise. A good model has a small estimate for $\widehat{\sigma^2}$ in comparison to that for the second sum. We now develop a quantitative comparison of the sums.

A raw comparison of both variance terms is not sufficient, the number of observations must be considered: SS_E increases as N increases also in light of a high quality model. In order to weight the individual sums of squares, we divide them by their degrees of freedom, e.g., instead of SS_E we will use $\text{SS}_E/(N - I)$ and instead of SS_A we will use $\text{SS}_A/(I - 1)$, which we will term mean squares. Under the null hypothesis, the mean squares are chi-square distributed and thus their quotients are F distributed. Hence, an F-test as illustrated in Test 4 is needed again. Historically, such a test has been "constructed" via a table and is still represented as such. This so-called ANOVA table consists of columns for the sums of squares, degrees of freedom, mean squares and F-test statistic due to variance between groups, within groups, and the total variance. Table 11.1 illustrates such a generic ANOVA table, numerical examples are given in Example 11.1 later in this section. Note that the third row of the table represents the sum of the first two rows. The last two columns are constructed from the first two ones.

Table 11.1: Table for one-way analysis of variance.

Source	Sum of squares (SS)	Degrees of freedom (DF)	Mean squares (MS)	Test statistic
Between groups (Factors, levels, ...)	$\text{SS}_A = \sum_{i,j}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$I - 1$	$\text{MS}_A = \dfrac{\text{SS}_A}{I-1}$	$F_{\text{obs}} = \dfrac{\text{MS}_A}{\text{MS}_E}$
Within groups (Error)	$\text{SS}_E = \sum_{i,j}(y_{ij} - \overline{y}_{i\cdot})^2$	$N - I$	$\text{MS}_E = \dfrac{\text{SS}_E}{N-I}$	
Total	$\text{SS}_T = \sum_{i,j}(y_{ij} - \overline{y}_{\cdot\cdot})^2$	$N - 1$		

The expected values of the mean squares (MS) are

$$\text{E}(\text{MS}_A) = \text{E}\big(\text{SS}_A/(I-1)\big) = \sigma^2 + \frac{\sum_i n_i \beta_i^2}{I-1}, \quad \text{E}(\text{MS}_E) = \text{E}\big(\text{SS}_E/(N-I)\big) = \sigma^2. \tag{11.20}$$

Note that only the latter is intuitive (Problem 11.1.**b**).

Thus under $H_0 : \beta_1 = \cdots = \beta_I = 0$, $\text{E}(\text{MS}_A) = \text{E}(\text{MS}_E)$ and hence the ratio MS_A/MS_E is close to one, but typically larger. We reject H_0 for large values of this ratio. Test 16 summarizes the procedure. The test statistic of Table 11.1 naturally agrees with that of Test 16. Observe that when $\text{MS}_A \le \text{MS}_E$, i.e., $F_{\text{obs}} \le 1$, H_0 is never rejected. Details about F-distributed random variables are given in Section 3.2.3.

Test 16: Performing a one-way analysis of variance

Question: Of the means $\overline{y}_1, \overline{y}_2, \dots, \overline{y}_I$, are at least two significantly different?

Assumptions: The I populations are normally distributed with homogeneous variances. The samples are independent.

Calculation: Construct a one-way ANOVA table. The quotient of the mean squares of the factor and the error are needed:

$$F_{\text{obs}} = \frac{\text{SS}_A/(I-1)}{\text{SS}_E/(N-I)} = \frac{\text{MS}_A}{\text{MS}_E}$$

Decision: Reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I$ if $F_{\text{obs}} > F_{\text{crit}}$, where F_{crit} is the $1-\alpha$ quantile of an F-distribution with $I-1$ gives the degrees of freedom "between" and $N-I$ the degrees of freedom "within"

Calculation in R: `summary( lm(...))` for the value of the test statistic or `anova( lm(...))` for the explicit ANOVA table.

Example 11.1 discusses a simple analysis of variance.

Example 11.1 (`retardant` data)**.** Many substances related to human activities end up in wastewater and accumulate in sewage sludge. The present study focuses on hexabromocyclododecane (HBCD) detected in sewage sludge collected from a monitoring network in Switzerland. HBCD's main use is in expanded and extruded polystyrene for thermal insulation foams, in building and construction. HBCD is also applied in the backcoating of textiles, mainly in furniture upholstery. A very small application of HBCD is in high impact polystyrene, which is used for electrical and electronic appliances, for example in audio visual equipment. Data and more detailed background information are given in Kupper *et al.* (2008) where it is also argued that loads from different types of monitoring sites showed that brominated flame retardants ending up in sewage sludge originate mainly from surface runoff, industrial and domestic wastewater.

HBCD is harmful to one's health, may affect reproductive capacity, and may harm children in the mother's womb.

In R-Code 11.1 the data are loaded and reduced to Hexabromocyclododecane. First we use constraint $\beta_1 = 0$, i.e., Model (11.5). The estimates naturally agree with those from (11.7). Then we use the sum-to-zero constraint and compare the results. The estimates and the standard errors changed (and thus the p-values of the t-test). The p-values of the F-test are, however, identical, since the same test is used.

The R command `aov` is an alternative for performing ANOVA and its use is illustrated in R-Code 11.2. We prefer, however, the more general `lm` approach. Nevertheless we need a function which provides results on which, for example, Tukey's honest significant difference (HSD) test can be performed with the function `TukeyHSD`. The differences can also be calculated from the coefficients in R-Code 11.1. The p-values are smaller because multiple tests are considered. ♣

R-Code 11.1: `retardant` data: ANOVA with `lm` command and illustration of various contrasts.

```
tmp <- read.csv('./data/retardant.csv')
retardant <- read.csv('./data/retardant.csv', skip=1)
names( retardant) <- names(tmp)
HBCD <- retardant$cHBCD
str( retardant$StationLocation)
##  chr [1:16] "A11" "A12" "A15" "A16" "B11" "B14" "B16" "B25" "C2" "C4" ...
type <- as.factor( rep(c("A","B","C"), c(4,4,8)))

lmout <- lm( HBCD ~ type )
summary(lmout)
##
## Call:
## lm(formula = HBCD ~ type)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -87.6  -44.4  -26.3   22.0  193.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     75.7       42.4    1.78    0.098 .
## typeB           77.2       60.0    1.29    0.220
## typeC          107.8       51.9    2.08    0.058 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.8 on 13 degrees of freedom
## Multiple R-squared:  0.249,Adjusted R-squared:  0.134
## F-statistic: 2.16 on 2 and 13 DF,  p-value: 0.155
```

```
options( "contrasts")
```

```
## $contrasts
##         unordered           ordered
## "contr.treatment"      "contr.poly"
```

```
# manually construct the estimates:
c( mean(HBCD[1:4]), mean(HBCD[5:8])-mean(HBCD[1:4]),
   mean(HBCD[9:16])-mean(HBCD[1:4]))
```

```
## [1]  75.675  77.250 107.788
```

```
# change the constrasts to sum-to-zero
options(contrasts=c("contr.sum","contr.sum"))
lmout1 <- lm( HBCD ~ type )
#  summary(lmout1)  # as above, except the coefficents are different:
print( summary(lmout1)$coef, digits=3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    137.4       22.3    6.15 3.51e-05
## type1          -61.7       33.1   -1.86 8.55e-02
## type2           15.6       33.1    0.47 6.46e-01
```

```
beta <- as.numeric(coef(lmout1))
# Construct 'contr.treat' coefficients:
c( beta[1]+beta[2], beta[3]-beta[2], -2*beta[2]-beta[3])
```

```
## [1]  75.675  77.250 107.787
```

R-Code 11.2 `retardant` data: ANOVA with `aov` and multiple testing of the means.

```
aovout <- aov( HBCD ~ type )
options("contrasts")

## $contrasts
## [1] "contr.sum" "contr.sum"

coefficients( aovout)  # coef( aovout) is suffient as well.

## (Intercept)       type1       type2
##     137.354     -61.679      15.571

summary(aovout)

##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2  31069   15534    2.16   0.15
## Residuals   13  93492    7192

TukeyHSD( aovout)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HBCD ~ type)
##
## $type
##        diff      lwr    upr   p adj
## B-A  77.250  -81.085 235.59 0.42602
## C-A 107.787  -29.335 244.91 0.13376
## C-B  30.537 -106.585 167.66 0.82882
```

11.2 Two-Way and Complete Two-Way ANOVA

Model (11.2) can be extended for additional factors. Adding one factor to a one-way model leads us to a two-way model

$$Y_{ijk} = \mu + \beta_i + \gamma_j + \varepsilon_{ijk} \tag{11.21}$$

with $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, n_{ij}$ and $\varepsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The indices again specify the levels of the first and second factor as well as the count for that configuration. As stated, the model is over parameterized and additional constraints are again necessary, in which case

$$\sum_{i=1}^{I} \beta_i = 0, \quad \sum_{j=1}^{J} \gamma_j = 0 \quad \text{or} \quad \beta_1 = 0, \quad \gamma_1 = 0 \tag{11.22}$$

are often used.

When the n_{ij} in each group are not equal, the decomposition of the sums of squares is not necessarily unique. In practice this is often unimportant since, above all, the estimated coefficients are compared with each other (“contrasts”). We recommend always using the command `lm(...)`. In case we do compare sums of squares there are resulting ambiguities and factors need to be included in decreasing order of "natural" importance.

For the sake of illustration, we consider the *balanced* case of $n_{ij} = K$, called complete two-way ANOVA. More precisely, the model consists of $I \cdot J$ groups and every group contains K samples and $N = I \cdot J \cdot K$. The calculation of the estimates are easier than in the unbalanced case and are illustrated as follows.

As in the one-way case, we can derive least squares estimates

$$y_{ijk} = \underbrace{\overline{y}_{\dots}}_{\widehat{\mu}} + \underbrace{\overline{y}_{i\cdot\cdot} - \overline{y}_{\dots}}_{\widehat{\beta}_i} + \underbrace{\overline{y}_{\cdot j\cdot} - \overline{y}_{\dots}}_{\widehat{\gamma}_j} + \underbrace{y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\dots}}_{r_{ijk}} \tag{11.23}$$

and separate the sums of squares

$$\underbrace{\sum_{i,j,k}(y_{ijk} - \overline{y}_{\dots})^2}_{\text{SS}_T} = \underbrace{\sum_{i,j,k}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\dots})^2}_{\text{SS}_A} + \underbrace{\sum_{i,j,k}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\dots})^2}_{\text{SS}_B} + \underbrace{\sum_{i,j,k}(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\dots})^2}_{\text{SS}_E}. \tag{11.24}$$

We are interested in the statistical hypotheses $H_0 : \beta_1 = \dots = \beta_I = 0$ and $H_0 : \gamma_1 = \dots = \gamma_J = 0$. The test statistics of both of these tests are given in the last column of Table 11.2. The test statistic $F_{\text{obs},A}$ is compared with the quantiles of the F distribution with $I - 1$ and $N - I - J + 1$ degrees of freedom. Similarly, for $F_{\text{obs},B}$ we use $J - 1$ and $N - I - J + 1$ degrees of freedom. The tests are equivalent to that of Test 16.

Table 11.2: Table of the complete two-way ANOVA.

Source	SS	DF	MS	Test statistic
Factor A	$\text{SS}_A = \sum_{i,j,k}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\dots})^2$	$I - 1$	$\text{MS}_A = \dfrac{\text{SS}_A}{I-1}$	$F_{\text{obs},A} = \dfrac{\text{MS}_A}{\text{MS}_E}$
Factor B	$\text{SS}_B = \sum_{i,j,k}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\dots})^2$	$J - 1$	$\text{MS}_B = \dfrac{\text{SS}_B}{J-1}$	$F_{\text{obs},B} = \dfrac{\text{MS}_B}{\text{MS}_E}$
Error	$\text{SS}_E =$ $\sum_{i,j,k}(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\dots})^2$	$\text{DF}_E =$ $N{-}I{-}J{+}1$	$\text{MS}_E = \dfrac{\text{SS}_E}{\text{DF}_E}$	
Total	$\text{SS}_T = \sum_{i,j,k}(y_{ijk} - \overline{y}_{\dots})^2$	$N - 1$		

Model (11.21) is additive: "More of both leads to even more". It might be that there is a certain canceling or saturation effect. To model such a situation, we need to include an interaction $(\beta\gamma)_{ij}$ in the model to account for the non-linear effects:

$$Y_{ijk} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \varepsilon_{ijk} \tag{11.25}$$

with $\varepsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and corresponding ranges for the indices. In addition to constraints (11.22) we require

$$\sum_{i=1}^{I} (\beta\gamma)_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{J} (\beta\gamma)_{ij} = 0 \quad \text{for all } i \text{ and } j \tag{11.26}$$

or analogous treatment constraints are often used. As in the previous two-way case, we can derive the least squares estimates

$$y_{ijk} = \underbrace{\overline{y}_{...}}_{\widehat{\mu}} + \underbrace{\overline{y}_{i..} - \overline{y}_{...}}_{\widehat{\beta}_i} + \underbrace{\overline{y}_{.j.} - \overline{y}_{...}}_{\widehat{\gamma}_j} + \underbrace{\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}}_{\widehat{(\beta\gamma)}_{ij}} + \underbrace{y_{ijk} - \overline{y}_{ij.}}_{r_{ijk}} \tag{11.27}$$

and a decomposition in sums of squares is straightforward.

Note that for $K = 1$ a model with interaction does not make sense as the error of the two-way model is the interaction.

Table 11.3 shows the table of a complete two-way ANOVA. The test statistics $F_{\text{obs},A}$, $F_{\text{obs},B}$ and $F_{\text{obs},AB}$ are compared with the quantiles of the F distribution with $I-1$, $J-1$, $(I-1)(J-1)$ and $N-IJ$ degrees of freedom, respectively.

Table 11.3: Table of the complete two-way ANOVA with interaction.

Source	SS	DF	MS	Test statistic
Factor A	$\text{SS}_A = \sum_{i,j,k} (\overline{y}_{i..} - \overline{y}_{...})^2$	$I-1$	$\text{MS}_A = \frac{\text{SS}_A}{\text{DF}_A}$	$F_{\text{obs},A} = \frac{\text{MS}_A}{\text{MS}_E}$
Factor B	$\text{SS}_B = \sum_{i,j,k} (\overline{y}_{.j.} - \overline{y}_{...})^2$	$J-1$	$\text{MS}_B = \frac{\text{SS}_B}{\text{DF}_B}$	$F_{\text{obs},B} = \frac{\text{MS}_B}{\text{MS}_E}$
Interaction	$\text{SS}_{AB} =$ $\sum_{i,j,k} (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$	$(I-1)\times$ $(J-1)$	$\text{MS}_{AB} = \frac{\text{SS}_{AB}}{\text{DF}_{AB}}$	$F_{\text{obs},AB} = \frac{\text{MS}_{AB}}{\text{MS}_E}$
Error	$\text{SS}_E = \sum_{i,j,k} (y_{ijk} - \overline{y}_{ij.})^2$	$N-IJ$	$\text{MS}_E = \frac{\text{SS}_E}{\text{DF}_E}$	
Total	$\text{SS}_T = \sum_{i,j,k} (y_{ijk} - \overline{y}_{...})^2$	$N-1$		

11.3 Analysis of Covariance

We now combine regression and analysis of variance elements, i.e., continuous predictors and factor levels, by adding “classic” predictors to models (11.2) or (11.21). Such models are sometimes called ANCOVA, e.g., example

$$Y_{ijk} = \mu + \beta_1 x_i + \gamma_j + \varepsilon_{ijk}, \tag{11.28}$$

with $j = 1, \dots, J$, $k = 1, \dots, n_j$ and $\varepsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Additional constraints are again necessary. Keeping track of indices and Greek letters quickly gets cumbersome and one often resorts to R formula notation. For example, if the predictor x_i is in the variable `Population` and γ_j is in the variable `Treatment`, in form of a factor then

$$\texttt{y ~ Population + Treatment} \tag{11.29}$$

is representing (11.28). An intersection is indicated with `:`, for example `Treatment:Month`. The `*` operator denotes 'factor crossing': `a*b` is interpreted as `a + b + a:b`. See Section Details of `help(formula)`.

ANOVA tables are constructed similarly and individual rows thereof are associated to the individual elements in a formula specification. Hence, our unified approach via `lm(...)`. Notice that for the estimates the order of the variable in formula (11.29) does not play a role, for the decomposition of sums of squares it does. Different statistical software packages have different approaches and thus may lead to minor differences.

11.4 Example

Example 11.2 (`UVfilter` data)**.** Octocrylene is an organic UV Filter found in sunscreen and cosmetics. The substance is classified as a contaminant and dangerous for the environment by the EU under the CLP Regulation. Sunscreens containing i.a. octocrylene has been forbidden in some areas as a protective measure for their coral reefs.

Because the substance is difficult to break down, the environmental burden of octocrylene can be estimated through the measurement of its concentration in sludge from waste treatment facilities.

The study Plagellat *et al.* (2006) analyzed octocrylene (`OC`) concentrations from 24 different purification plants (consisting of three different types of `Treatment`), each with two samples (`Month`). Additionally, the catchment area (`Population`) and the amount of sludge (`Production`) are known. Treatment type `A` refers to small plants, `B` medium-sized plants without considerable industry and `C` medium-sized plants with industry.

R-Code 11.3 prepares the data and fits first a one-way ANOVA based on `Treatment` only, followed by a two-way ANOVA based on `Treatment` and `Month` (with and without interactions). Note that the setup is not balanced with respect to treatment type.

Adding the factor `Month` improves considerably the model fit: increase of the adjusted R^2 from 40% to 50%) and the standard errors of the treatment effects estimates are further reduced.

The interaction is not significant, as the corresponding p-value is above 14%. Based on Figure 11.1 this is not surprising. First, the seasonal effect of groups A and B are very similar and second, the variability in group C is too large. ♣

R-Code 11.3: `UVfilter` data: one-way ANOVA using `lm`.

```r
UV <- read.csv( './data/chemosphere.csv')
UV <- UV[,c(1:6,10)]     # reduce to OT
```

```
str(UV, strict.width='cut')

## 'data.frame': 24 obs. of  7 variables:
##  $ Treatment : chr  "A" "A" "A" "A" ...
##  $ Site_code : chr  "A11" "A12" "A15" "A16" ...
##  $ Site      : chr  "Chevilly" "Cronay" "Thierrens" "Prahins" ...
##  $ Month     : chr  "jan" "jan" "jan" "jan" ...
##  $ Population: int  210 284 514 214 674 5700 8460 11300 6500 7860 ...
##  $ Production: num  2.7 3.2 12 3.5 13 80 150 220 80 250 ...
##  $ OT        : int  1853 1274 1342 685 1003 3502 4781 3407 11073 3324 ...

with( UV, table(Treatment, Month))

##          Month
## Treatment jan jul
##         A   5   5
##         B   3   3
##         C   4   4

options( contrasts=c("contr.sum", "contr.sum"))
lmout <- lm( log(OT) ~ Treatment, data=UV)
summary( lmout)

##
## Call:
## lm(formula = log(OT) ~ Treatment, data = UV)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.952 -0.347 -0.136  0.343  1.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.122      0.116   70.15  < 2e-16 ***
## Treatment1    -0.640      0.154   -4.16  0.00044 ***
## Treatment2     0.438      0.175    2.51  0.02049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.555 on 21 degrees of freedom
## Multiple R-squared:  0.454,Adjusted R-squared:  0.402
## F-statistic: 8.73 on 2 and 21 DF,  p-value: 0.00174

lmout <- lm( log(OT) ~ Treatment + Month, data=UV)
summary( lmout)

##
```

```
## Call:
## lm(formula = log(OT) ~ Treatment + Month, data = UV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7175 -0.3452 -0.0124  0.1691  1.2236
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.122      0.106   76.78  < 2e-16 ***
## Treatment1    -0.640      0.141   -4.55  0.00019 ***
## Treatment2     0.438      0.160    2.74  0.01254 *
## Month1        -0.235      0.104   -2.27  0.03444 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.507 on 20 degrees of freedom
## Multiple R-squared:  0.566,Adjusted R-squared:  0.501
## F-statistic: 8.69 on 3 and 20 DF,  p-value: 0.000686
summary( aovout <- aov( log(OT) ~ Treatment * Month, data=UV))
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment        2   5.38   2.688   11.67 0.00056 ***
## Month            1   1.32   1.325    5.75 0.02752 *
## Treatment:Month  2   1.00   0.499    2.17 0.14355
## Residuals       18   4.15   0.230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD( aovout, which=c('Treatment'))
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(OT) ~ Treatment * Month, data = UV)
##
## $Treatment
##         diff      lwr     upr   p adj
## B-A  1.07760  0.44516 1.71005 0.00107
## C-A  0.84174  0.26080 1.42268 0.00446
## C-B -0.23586 -0.89729 0.42556 0.64105
boxplot( log(OT)~Treatment, data=UV, col=7, boxwex=.5)
at <- c(0.7, 1.7, 2.7, 1.3, 2.3, 3.3)
boxplot( log(OT)~Treatment+Month, data=UV, add=T, at=at, xaxt='n', boxwex=.2)
```

```r
with( UV, interaction.plot( Treatment, Month, log(OT), col=2:3))
```

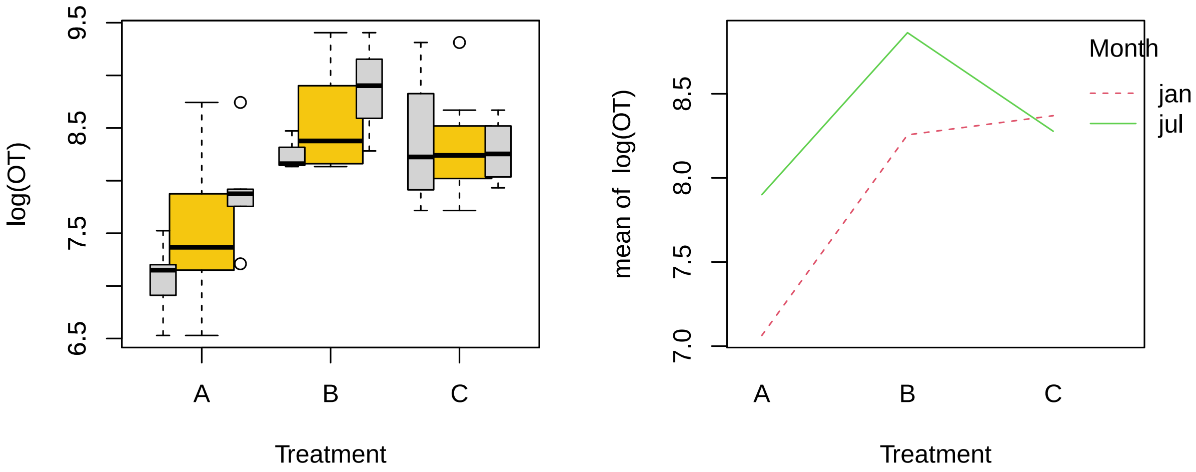


Figure 11.1: *UVfilter* data: box plots sorted by treatment and interaction plot. (See R-Code 11.3.)

11.5 Bibliographic Remarks

Almost all books covering linear models have a section about ANOVA.

11.6 Exercises and Problems

Problem 11.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Show the second and third equality of

$$R^2_{\text{adj}} = R^2 - (1 - R^2)\frac{p}{n - p - 1} = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} = 1 - \frac{\text{SS}_E/\text{dof}_E}{\text{SS}_T/\text{dof}_T}$$

b) Show the results of (11.20).

Problem 11.2 (ANOVA) We consider the data *chemosphere_OC.csv* available on the course web page. The data describe the octocrylene (*OC*) concentration sampled from 12 wastewater treatment plants in Switzerland. Further variables in the dateset are: *Behandlung* (treatment of the wastewater), *Monat* (month when the sample was collected), *Einwohner* (number of inhabitant connected to the plant), *Produktion* (sludge production (metric tons of dry matter per year), everything that doesn't enter the water system after treatment).

Octocrylene is an organic UV filter and is used in sunscreens and as additive in cosmetics for daily usage. The substance is classified as irritant and dangerous for the environment (EU classification of dangerous substances).

The data are published in C. Plagellat, T. Kupper, R. Furrer, L. F. de Alencastro, D. Grandjean, J. Tarradellas *Concentrations and specific loads of UV filters in sewage sludge originating from a monitoring network in Switzerland*, Chemosphere 62 (2006) 915–25.

a) Describe the data. Do a visual inspection to check for differences between the treatment types and between the months of data aquirement. Use an appropriate plot function to do so. Describe your results.

Hint: Also try the function `table()`

b) Fit a one-way ANOVA with `log(OC)` as response variable and `Behandlung` as explanatory variable.

Hint: use `lm` and perform an `anova` on the output. Don't forget to check model assumptions.

c) Extend the model to a two-way ANOVA by adding `Monat` as a predictor. Interpret the summary table.

d) Test if there is a significant interaction between `Behandlung` and `Monat`. Compare the result with the output of `interaction.plot`

e) Extend the model from (b) by adding `Produktion` as an explanatory variable. Perform an `anova` on the model output and interpret the summary table. (Such a model is sometimes called Analysis of Covariance, ANCOVA).

Switch the order of your explanatory variables and run an anova on both model outputs. Discuss the results of `Behandlung + Produktion` and `Produktion + Behandlung`. What causes the differences?

Problem 11.3 (ANOVA table) Calculate the missing values in the following table:

```
library( aov( yield ~ block + N + P + K, digits=npk))
##              Df Sum Sq Mean Sq F value  Pr(>F)
## block        5  343.3
## N            1                          0.00366
## P                 8.4    8.40
## K            1   95.2           5.946
## Residuals   15          16.01
```

How many observations have been taken in total? Do we have a balanced or complete four-way setting?

Problem 11.4 (ANCOVA) The dataset `rats` of the package `faraway` consists of two-way ANOVA design with factors poison (three levels I, II and III) and treatment (four levels A, B, C and D). To study the toxic agents, 4 rats were exposed to each pair of factors. The response was survival time in tens of hours.

a) Test the effect of poison and treatment.

b) Would a transformation of the response variable improve the model?

Problem 11.5 (ANCOVA) The perceived stress scale (PSS) is the most widely used psychological instrument for measuring the perception of stress. It is a measure of the degree to which situations in one's life are appraised as stressful.

The dataset `PrisonStress` from the package `PairedData` gives the PSS measurements for 26 people in prison at the entry and at the exit. Part of these people were physically trained during their imprisonment.

a) Describe the data. Do a visual inspection to check for differences between the treatment types and PSS.

b) Propose a statistical model.

Problem 11.6 (BMJ Endgame) Discuss and justify the statements about 'One way analysis of variance' given in doi.org/10.1136/bmj.e2427.

Chapter 12

Design of Experiments

Learning goals for this chapter:

- ◇ Understand the issues and principles of Design of Experiments
- ◇ Compute sample size for an experiment
- ◇ Compute power of test
- ◇ Describe different setups of an experiment

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter12.R.

Design of Experiments (DoE) is a relatively old field of statistics. Pioneering work has been done almost 100 years ago by Sir Ronald Fisher and co-workers at Rothamsted Experimental Station, England, where mainly agricultural questions have been discussed. The topic has been taken up by the industry after the second world war to, e.g., optimize production of chemical compounds, work with robust parameter designs. In recent decades, advances are still been made on for example using the abundance of data in machine learning type discovery or in preclinical and clinical research were the sample sizes are often extremely small.

In this chapter we will selectively cover different aspects of DoE, focusing on sample size calculations, power and randomization. Additionally, we also cover a few domain specific concepts and terms that are often used in the context of setting up experiments for clinical or preclinical trials.

12.1 Basic Idea and Terminology

Often statisticians are consulted *after* the data has been collected. Moreover, if the data does not “show” what has been hypothesized, frantic visits to “data clinics” are done. Along the same lines, Sir Ronald Fisher once stated “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” (Fisher, 1938, Presidential Address to the First Indian Statistical Congress).

Here, the term 'experiment' describes a controlled procedure that is (hopefully) carefully designed to test one (or very few) scientific hypothesis. In the context of this chapter, the hypothesis is often about the effect of independent variables on one dependent variable, i.e., the outcome measure. In terms of our linear model equation (10.2), what is the effect of one or several of the $x_{i\ell}$ on the Y_i. Designing the experiment implies the choice of the independent variables (we need to account for possible confounders, or effect modifiers), the values thereof (fixed at particular "levels" or randomly chosen) and sample size. Again in terms of our model, we need to include and well specify all necessary predictors $x_{i\ell}$ that have an effect on Y_i. Finally, we need to determine the sample size n such that the desired effects – if they exist – are statistically significant.

The prime paradigm is of DoE is

Maximize primary variance, minimize error variance and control for secondary variance.

which translates to maximize the signal we are investigating, minimize the noise we are not modeling and control for uncertainties with carefully chosen independent variables.

In the context of DoE we often want to compare the effect of a treatment (or procedure) on an outcome. Examples that have been discussed in previous chapters are: "Is there a progression of pododermatitis at the hind paws over time?", "Is a diuretic medication during pregnancy reducing the risk of pre-eclampsia?", "How much can we increase hardness of metal springs with lower temperatures of quenching baths?", "Is residual octocrylene in waste water sludge linked to particular waste water types?".

To design an experiment, it is very important to differentiate between *exploratory* or *confirmatory research* questions. An exploratory experiment tries to discover as much as possible about the sample or the phenomenon under investigation, given time and resource constraints. Whereas in a confirmatory experiment we want to verify, to confirm, or to validate a result, which was often derived from an earlier exploratory experiment. Table 12.1 summarizes both approaches in a two-valued setting. Some of the design elements will be further discussed in later sections of this chapter. The binary classification should be understood within each domain: few observations in one domain could be many in another one. In both situations and all scientific domains, however, a proper statistical analysis is crucial.

Table 12.1: Differences between exploratory or confirmatory experiments.

Design feature	Exploratory	Confirmatory
Design space	Large	Small
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardized
Treatments	Many	Few
Outcomes	Many	Few
Hypotheses	Loose and flexible	Narrow and predefined
Statistical tests	Generate hypotheses	Confirm/reject hypotheses
Inferences about population	Not possible	Possible with rigorous designs

12.2 Sample Size Calculations

When planning an experiment, we should always carefully evaluate sample size n, that is require to be able to properly conclude our hypothesis. In many cases sample size needs to be determined before starting the experiment: organizing (time-wise) the experiment, acquire necessary funds, filing study protocols or submitting a license to an ethic commission. As a general rule, we choose *as many as possible but as few as necessary* samples to balance statistical and economic interest.

12.2.1 Experimental and Response Units

Suppose that we test the effect of a dietary treatment for female rabbits (say, with and without a vitamin additive) on the weight of the litter within two housing boxes. Each doe (i.e., female reproductive rabbit) in the box receives the same treatment, i.e., the treatment is not applied to a single individual subject and could not be individually controlled for. All does form a single, so-called *experimental unit*. The outcomes or responses are measured on the *response units*, which are typically "smaller" than the experimental units. In our example, we would weight the litter of each doe in the housing box individually, but aggregate or average these to a single number. As a side note, this average often justifies the use of a normal response model when an experimental unit consists of several response units.

Formally, experimental units are entities which are independent of each other and to which it is possible to assign a treatment or intervention independently of the other units. The experimental unit is the unit which has to be replicated in an experiment. Below, when we talk about a sample, we are talking about a sample of experimental units. We do not discuss the choice of the response units here as it is most often situation specific were a statistician has little to contribute.

12.2.2 Case of Single Confidence Interval

In this section, we link the sample size to the width of three different confidence intervals. Specifically, we discuss the necessary number of observations required such that the width of the empirical confidence interval has a predefined size.

To start, assume that we are in the setting of a simple z confidence interval at level $(1-\alpha)$ with known σ, as seen Equation (4.31). If we want to ensure an empirical interval width ω, we need

$$n \approx 4z^2_{1-\alpha/2}\frac{\sigma^2}{\omega^2} \tag{12.1}$$

observations. We use an approximation relation because sample size is determined as an integer. In this setting, the right-hand-side of (12.1) does not involve the data and thus the width of the confidence interval is in any case guaranteed. Note that in order to reduce the width by a factor two, we need to quadruple the sample size.

The same approach is used when estimating a proportion. We use, for example, the Wald confidence interval (6.10) to get

$$n \approx 4z^2_{1-\alpha/2}\frac{\widehat{p}(1-\widehat{p})}{\omega^2}, \tag{12.2}$$

which corresponds to (12.1) with the the plug-in estimate $\widehat{p}(1-\widehat{p})$ for $\widehat{\sigma}^2$ in the setting of a Bernoulli random variable and central limit approximation for X/n. Of course, $\widehat{p}$ is not known a priori and we often take the conservative choice of $\widehat{p} = 1/2$ as the function $x(1-x)$ is maximized over $(0,1)$ at $x = 1/2$. Thus, without any prior knowledge on p we may choose conservatively $n \approx (z_{1-\alpha/2}/\omega)^2$. Alternatively, the sample size calculation can be done based on a Wilson confidence interval (6.11), where a quadratic equation needs to be solved to obtain n (see Problem 12.1.**a**).

If we are estimating a Pearson's correlation coefficient, we can use CI 6 to link interval width ω with n. Here, we use an alternative approach and would like to determine sample size such that the interval does not contain the value zero, i.e., the width is just smaller than $2r$. The derivation relies on the duality of tests and confidence intervals (see Section 5.4). Recall Test 14 for Pearson's correlation coefficient. From Equation (9.3) we construct the critical value for the test (boundary of the rejection region, see Figure 5.3) and based on that we can calculate the minimum sample size necessary to detect a correlation $|r| \geq r_\text{crit}$ as significant:

$$t_\text{crit} = r_\text{crit}\frac{\sqrt{n-2}}{\sqrt{1-r^2_\text{crit}}} \quad\Longrightarrow\quad r_\text{crit} = \frac{t_\text{crit}}{\sqrt{n-2+t^2_\text{crit}}}. \tag{12.3}$$

Figure 12.1 illustrates the least significant correlation for specific sample sizes. Specifically, with sample size $n < 24$ correlations between -0.4 and 0.4 are not significant and for a correlation of ±0.25 to be significant, we require $n > 62$ at level $\alpha = 5\%$ (see R-Code 12.1).

R-Code 12.1 Significant correlation for specific sample sizes (See Figure 12.1.)

```
rcrit <- function(n, alpha=.05) {
  tcrit <- qt( 1 - alpha/2, df=n-2)
  return( tcrit / sqrt( n-2 + tcrit^2))
}
curve( rcrit(x), from=3, to=200, xlab="n", ylab="rcrit", ylim=c(0,1), yaxs='i')
round( c(rcrit(25), uniroot( function(x) rcrit(x)-0.25, c(50,100))$root), 2)
## [1]  0.40 62.02
abline( v=63, h=0.25, col='gray')
```

12.2.3 Case of *t*-Tests

Sample sizes are most often determined to be able to "detect" an alternative hypothesis with a certain probability. That means we need to link the sample size with the power $1-\beta$ of a particular statistical test.

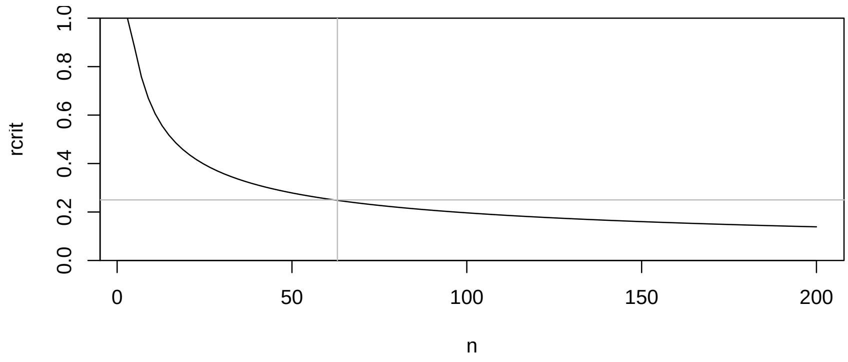


Figure 12.1: Significant correlation for specific sample sizes (at level $\alpha = 5\%$). For a sample correlation of 0.25, n needs to be larger than 62 as indicated with the gray lines. For a particular n, correlations above the line are significant, below are not. (See R-Code 12.1.)

As a simple example, we consider a one-sided z-test with $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$. The Type II error probability for the true mean μ_1 is

$$\beta = \beta(\mu_1) = \text{P}(H_0 \text{ not rejected given } \mu = \mu_1) = \dots = \Phi\Big(z_{1-\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\Big). \tag{12.4}$$

Suppose we want to "detect" the alternative with probability $1 - \beta(\mu_1)$, i.e., reject the null hypothesis with probability $1 - \beta$ when the true mean is μ_1. Hence, plugging the values in (12.4) and solving for n we have approximate sample size

$$n \approx \Big(\sigma \frac{z_{1-\alpha} + z_{1-\beta}}{\mu_0 - \mu_1}\Big)^2. \tag{12.5}$$

Hence, the sample size depends on the Type I and II error probabilities as well as the standard deviation and the difference of the means. The latter three quantities are often combined to the standardized effect size

$$\delta = \frac{\mu_0 - \mu_1}{\sigma}, \tag{12.6}$$

If the standard deviation is not known, an estimate can be used. An estimate based version of δ is often termed *Cohen's d*.

For a two-sided test, a similar expression is found where $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. For a one-sample t-test (Test 1) the right hand side of (12.5) is analogue with the quantiles $t_{n-1,1-\alpha}$ and $t_{n-1,1-\beta}$ respectively. Note that now the right hand side depends on n as well as on the quantiles. To determine n, we start with a reasonable value for n to obtain the quantiles, calculate the resulting n and repeat the two steps for at least one more iteration. In R, the function `power.t.test()` uses a numerical approach.

In the case of two independent samples, the degrees of freedom in the t-quantiles need to be adjusted from $n-1$ to $n-2$. Cohen's d is defined as $(\bar{x}_1 - \bar{x}_2)/s_p$, where s_p^2 is an estimate of the pooled variance (e.g., as given in Test 2).

For t-tests in the behavioral sciences, Cohen (1988) defined small, medium and large (standardized) effect sizes as $d = 0.2$, 0.5 and 0.8, respectively. These are often termed the *conventional* *effect sizes* but depend on the type of test, see also the function `cohen.ES()` of the R package `pwr`. In preclinical studies, effect sizes are typically larger.

Example 12.1. In the setting of a two-sample t-test with equal group sizes, we need at level $\alpha = 5\%$ and power $1 - \beta = 80\%$ in each group 26, 64 and 394 observations for a large, medium and small effect size, respectively, see, e.g., `power.t.test( d=0.2, power=0.8)` or `pwr.t.test( d=0.2, power=0.8)` from the `pwr` package.

For unequal sample sizes, the sum of both group sizes is a bit larger compared to equal sample sizes (balanced setting). For a large effect size, we would, for example, require $n_1 = 20$ and $n_2 = 35$, leading to three more observations compared to the balanced setting, (`pwr.t2n.test(n1=20, d=0.8, power=.8)` from the `pwr` package). ♣

Many R packages and other web interfaces allow to calculate sample sizes for many different settings of comparing means. Of course, care is needed when applying such toolboxes whether they use the same parametrizations, definitions etc.

12.3 Blocking, Randomization and Biases

In many experiments the subjects or generally the experimental units are inherently heterogeneous with respect to factors that we are not of interest. This heterogeneity may imply variability in the data, masking the effect we would like to study. *Blocking* is a technique for dealing with this nuisance heterogeneity. Hence, we distinguish between the treatment factors that we are interested in and the nuisance factors which have some effect on the response but are not of interest to us.

The term blocking originated in agricultural experiments where it refers to a set of plots of land that have very similar characteristics with respect to crop yield, in other words they are homogeneous.

If a nuisance factor is known and controllable, we use blocking and control for it by including a blocking factor in the experiment. Typical, blocking factors are factory, production batch, or sex. These are controllable in the sense that we are able to choose which factor to include

If a nuisance factor is known and uncontrollable, we may use the concept of ANCOVA, i.e., to remove the effect of the factor. Suppose that age has an effect on the treatment. It is not possible to control for age and creating age batches may not be efficient either. Hence we include age in our model. This approach is less efficient than blocking as we do correct for the design compared to design the experiment to account for the factor.

Unfortunately, there are also unknown and uncontrollable nuisance factors. To protect for these we use proper randomization such that their impact is balanced in all groups. Hence, we can see randomization as a insurance against systematic biases due to nuisance factors.

12.3.1 Randomization

Randomization is the mechanism to assign a treatment to an experimental unit by pure chance. Randomization ensures that – on average – the only systematic difference between the groups is the treatment, all other effects that are not accounted for are averaged out. Proper randomization also protects against spurios correlations in the observations.

The randomization procedure should be a truly randomized procedure for all assignments, ultimately performed by a genuine random number generator. There are several procedures, simple randomization, balanced or constrained randomization, stratified randomization etc. and of course, the corresponding sample sizes are determined a priori.

Simple randomization randomly assigns the type of treatment to an experimental unit. For example, to assign 12 subjects to three groups we use `sample(x=3, size=12, replace=TRUE)`. This procedure has the disadvantage of leading to a possibly unbalanced design and thus should not be used in practice.

A practical randomization scheme is a *completely randomized design* (CRD) in which we randomly assign the type of treatment to an experimental unit constrained that in all groups the same number of subjects are assigned to a specific treatment (conditional on appropriate sample size). This can be achieved by numbering the subjects, followed by random drawing of the corresponding numbers and finnaly putting the corresponding subjects in the appropriate four groups: `matrix(sample(x=12, size=12, replace=FALSE), nrow=3)`. The four groups themselves should also be randomized: `sample(c(’A’,’B’,’C’))`. In a single call, we have `set.seed(1); matrix(sample(x=12, size=12, replace=FALSE), nrow=3, dimnames=list(sample(c(’A’,’B`

When setting up an experiment, we need to carefully eliminate possible factors or variables that influence both, the dependent variable and independent variable and thus causing a spurious association. For example, smoking causes lung cancer and yellow fingers. It is not possible to conclude that yellow fingers cause lung cancer. Similarly, there is a strong association between the number of fire men at the incident and the filed insurance claim (the more men the larger the claim). The confounding here is the size of the fire, of course. In less striking words, we often have to take into account slope, aspect, altitude when comparing the yield of different fields or the sex, age, weight, pre-existing conditions when working with human or animal data. Not taking into account these factors will induce biases in the results.

In the case of discrete confounders it is possible to split your sample into subgroups according to these pre-defined factors. These subgroups are often called blocks (when controllable) or strata (when not). To randomize, *randomized complete block design* (RCBD), a stratified randomization, is used. In RCBD each block receives the same amount of experimental units and each block is like a small CRD experiment.

Suppose we have six male and six female subjects. The randomization is done for both the male and female groups according to `cbind( matrix(sample(x=6, size=6, replace=FALSE), nrow=3), matrix(sample(x=7:12, size=6, replace=FALSE), nrow=3))`.

Example 12.2. Suppose we are studying the effect of fertilizer type on plant growth. We have access to 24 plant pots arranged in a four by six alignment inside a glass house. We have three different fertilizer and one control group.

To allocate the fertilizers to the plants we number the plants row by row (starting top right). We assign randomly the 24 pots randomly to four groups of four. The fertilizers are then additionally randomly assigned to the different groups. This results in a CRD design (left panel of Figure 12.2).

Suppose that the glass house has one open wall. This particular setup affects plant growth unequally, because of temperature differences between the open and the opposite side. To account for the difference we block the individual rows and we assign randomly one fertilizer to each plant in every row. This results in a RCBD design (right panel of Figure 12.2). ♣

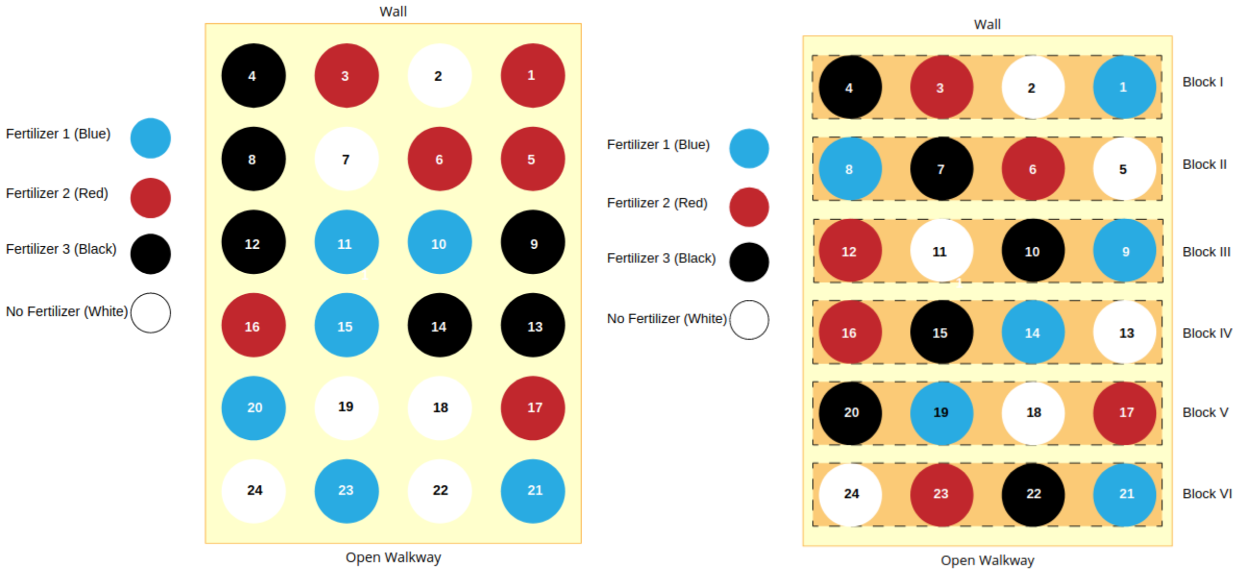


Figure 12.2: Different randomization of 24 plants. CRD (left), RCBD (right). Source online.stat.psu.edu/stat502/lesson/6.

RCBD are used in practice and might be seen on fields of agricultural education and research stations, as illustrated in Figure 12.3.

Remark 12.1. In certain situations there are two different EUs in the same experiment. As illustration suppose we are studying the effect of irrigation amount and fertilizer type on crop

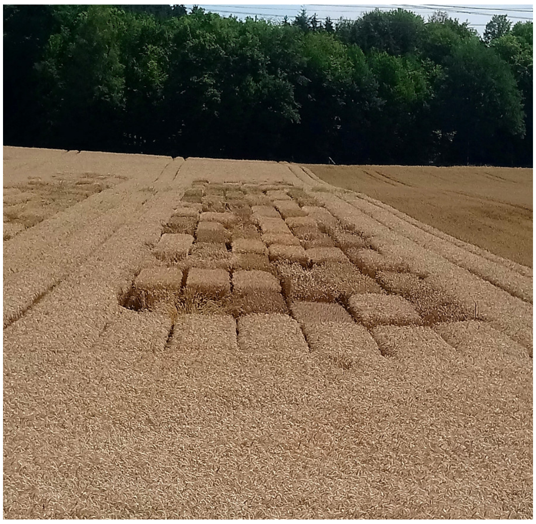


Figure 12.3: Grain field with experimental plots. (Photo: R. Furrer).

yield. If irrigation is more difficult to vary on a small scale and fields are large enough to be split, a *split plot design* becomes appropriate. Irrigation levels are assigned to whole plots by CRD and fertilizer is assigned to subplots using RCBD (irrigation is the block).

Split-plot experiments are not straight-forward to model statistically and we refer to followup lectures. ♣

12.3.2 Biases

Similar to the bias of an estimator we denote any systematic error of an experiment as a *bias*. Biases are introduced at the design of the study, at the experiment itself and at the analysis phase. Inherently, these biases are not part of the statistical model and thus induce biases in the estimates and/or inflate variance of the estimates.

There are many different types of biases and we explain these in its simplest from, assuming an experimental setup of two groups comparing one treatment to a control.

- *Selection bias* occurs when the two groups differ systematically next to the effect that is analyzed.

  When studies are not carefully implemented, it is possible to associate smaller risk for dementia for smokers compared to non-smokers. Here, selection bias occurs by smokers having a lower life-expectancy compared to non-smokers and thus having fewer dementia cases (Hernán *et al.*, 2008)

- *Confirmation bias* occurs when the experimenter or analyziser searches for a confirmation in the experiment.

  In a famous study, students were told that there exist two different types of rats, "maze bright" and "maze dull" rats where the former are genetically more apt to navigate in a maze. The result of an experiment by the students showed that the maze bright ones did perform systematically better than the maze dull ones (Rosenthal and Fode, 1963).

- *Confounding* is a bias that occurs due to another factor that distorts the relationship between treatment and outcome.

  There are many classical examples of confounding bias. One is that coffee drinkers have a larger risk of lung cancers. Such studies typically neglect that there is a larger proportion of smokers that are coffee drinkers, than non-smoking coffee drinkers (Galarraga and Boffetta, 2016).

- in certain cohort studies, subjects do not receive the treatment immediately after they have entered the study. The time span between entering and receiving the treatment is called the *immortal time* and this needs to be taken into account when comparing to the control group.

  A famous example was a study that wrongly claimed that Oscar laureates live about four years longer than comparable actors without (Redelmeier and Singh, 2001). It is not sufficient to group the actors in two groups (received Oscar or not) and then to match the actors in both groups.

- *Performance bias* is when a care giver or analyzer treats the subjects of the two groups differently.

- *Attrition bias* occurs when participants do not have the same drop-out rate in the control and treatment groups.

There are many more types of biases, see, e.g., catalogofbias.org/biases/.

12.3.3 Some Particular Terms

This section summarizes different terms that are used in the context of design of experiments.

An *intervention* is a process where a group of subjects (or experimental units) is subjected to such a surgical procedure, a drug injection, or some other form of treatment (intervention).

Control has several different uses in design. First, an experiment is controlled because scientists assign treatments to experimental units. Otherwise, we would have an *observational study*. Second, a control treatment is a "standard" treatment that is used as a baseline or basis of comparison for the other treatments. This control treatment might be the treatment in common use, or it might be a null treatment (no treatment at all). For example, a study of new pain killing drugs could use a standard pain killer as a control treatment, or a study on the efficacy of fertilizer could give some fields no fertilizer at all. This would control for average soil fertility or weather conditions.

Placebo is a null treatment that is used when the act of applying a treatment has an effect. Placebos are often used with human subjects, because people often respond to any treatment: for example, reduction in headache pain when given a sugar pill. Blinding is important when placebos are used with human subjects. Placebos are also useful for nonhuman subjects. The apparatus for spraying a field with a pesticide may compact the soil. Thus we drive the apparatus over the field, without actually spraying, as a placebo treatment. In case of several factors, they are combined to form treatments. For example, the baking treatment for a cake involves a given time at a given temperature. The treatment is the combination of time and temperature, but we can vary the time and temperature separately. Thus we speak of a time factor and a temperature factor. Individual settings for each factor are called levels of the factor.

A randomized controlled trial (RCT) is study in which people are allocated at random to receive one of several clinical interventions. One of these interventions is the standard of comparison or control. The control may be a standard practice, a placebo, a sham treatment or no intervention at all. Someone who takes part in a RCT is called a participant or subject. RCTs seek to measure and compare the outcomes after the participants received their intervention. Because the outcomes are measured, RCTs are quantitative studies.

In sum, RCTs are quantitative, comparative, controlled experiments in which investigators study two or more interventions in a series of individuals who receive them in random order. The RCT is one of the simplest and most powerful tools in clinical research but often relatively expensive.

Confounding occurs when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment. The two factors or treatments are said to be confounded. Except in very special circumstances, confounding should be avoided. Consider planting corn

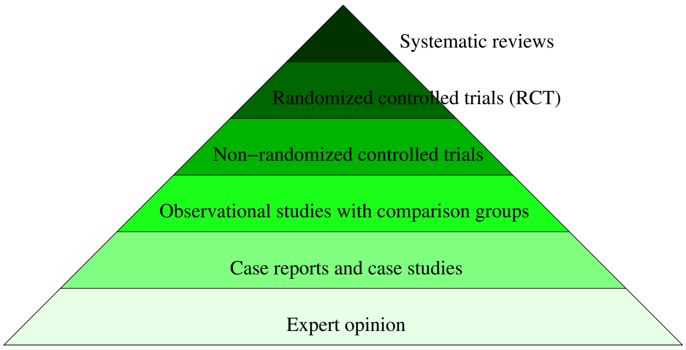

Systematic reviews
Randomized controlled trials (RCT)
Non–randomized controlled trials
Observational studies with comparison groups
Case reports and case studies
Expert opinion

Figure 12.4: (One possible) presentation of the evidence-based medicine pyramid.

variety A in Minnesota and corn variety B in Iowa. In this experiment, we cannot distinguish location effects from variety effects: the variety factor and the location factor are confounded.

Blinding occurs when the evaluators of a response do not know which treatment was given to which unit. Blinding helps prevent bias in the evaluation, even unconscious bias from well-intentioned evaluators. Double blinding occurs when both the evaluators of the response and the subject (experimental units) do not know the assignment of treatments to units. Blinding the subjects can also prevent bias, because subject responses can change when subjects have expectations for certain treatments.

Before a new drug is admitted to the market, many steps are necessary: starting from a discovery based step toward highly standardized clinical trials (type I, II and III). At the very end, there are typically randomized controlled trials, that by design (should) eliminate all possible confounders.

At later steps, when searching for an appropriate drug, a decision may be available based on "evidence": what has been used in the past, what has been shown to work (for similar situations). This is part of evidence-based medicine. Past information may be of varying quality, ranging from ideas opinions to case studies to RCTs or systematic reviews. Figure 12.4 represents a so-called *evidence-based medicine* pyramid which reflects the quality of research designs (increasing) and quantity (decreasing) of each study design in the body of published literature (from bottom to top). For other scientific domains, similar pyramids exist, with bottom and top typically remaining the same.

12.4 DoE in the Classical Framework

DoE in the Fisher sense is heavily ANOVA driven by his analysis of the crop experiments at Rothamsted Experimental Station and thus in many textbooks DoE is equated to the discussion of ANOVA. Here, we have separated the statistical analysis in Chapter 11 from the conceptual setup of the experiment in this chapter.

In a typical ANOVA setting we should strive to have the same amount of observations in each cell (for all settings of levels). Such a setting is called a balanced design (otherwise it is unbalanced). If every treatment has the same number of observations, effect of unequal variances are mitigated.

In a simple regression setting, the standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ depend on $1/\sum_i(x_i - \bar{x})^2$, see expressions for the estimates (9.12) and (9.13). Hence, to reduce the variability of the estimates, we should increase $\sum_i(x_i - \bar{x})^2$ as much as possible. Specifically, suppose the interval $[\,a, b\,]$ represents a natural range for the predictor, then we should choose half of the predictors as a and the other half as b.

This last argument justifies a discretization of continuous (and controllable) predictor variables in levels. Of course this implies that we expect a linear relationship. If the relationship is not linear, such a discretization may be devastating.

12.4.1 Sample Size in a One-Way and Two-Way ANOVA

To illustrate the difficulty of sample size calculation in an ANOVA context we consider a one-way example. Recall the one-way model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{12.7}$$

see (11.1) for detailed assumptions. We reject $H_0 : \mu_1 = \mu_1 = \cdots = \mu_I$ when $(\text{SS}_A/(I-1))/(\text{SS}_E/(N-I)) = \text{MS}_A/\text{MS}_E$ is larger than the $1-\alpha$-quantile of an F-distribution with $I-1$ and $N-I$ the degrees of freedom.

The difficulty is in specifying the alternative. One often chooses the setting where all but two group means are equal and the two deviate by $\Delta/2$.

In R, the function `Fpower1()` from the package `daewr` can be used to calculate the power and resulting sample size.

In the framework of two-way ANOVA, the same difficulties hold. In R, the function `Fpower2()` from the package `daewr` can be used.

12.4.2 Sums of Squares for Unbalanced Two-Way ANOVA

If we have a balanced two-way ANOVA, the sums of squares partition is additive due to "orthogonality" induced by the equal number in each cell and thus we have

$$\text{SS}_T = \text{SS}_A + \text{SS}_B + \text{SS}_{AB} + \text{SS}_E \tag{12.8}$$

(see also Equation (11.24)). In the unbalanced setting this is not the case and the decomposition depends on the order which we introduce the factors in the model. That means, the ANOVA table of `aov(y~f1+f2)` is not the same as `aov(y~f2+f1)`. We should consider the ANOVA table as sequential: each additional factor (row in the table), we reduce the remaining variability. Hence, we should rather write

$$\text{SS}_T = \text{SS}_A + \text{SS}_{B|A} + \text{SS}_{AB|A,B} + \text{SS}_E\,, \tag{12.9}$$

where the term $\text{SS}_{B|A}$ indicates the sums of squares of factor B after correction of factor A. and similarly, term $\text{SS}_{AB|A,B}$ indicates the sums of squares of the interaction AB after correction of factors A and B.

This concept of sums of squares after correction is not new. We have encountered this type of correction already: SS_T is actually calculated after correcting for the overall mean.

Equation (12.9) represents the sequential sums of squares decomposition, called *Type I sequential SS*: SS_A and $\mathrm{SS}_{B|A}$ and $\mathrm{SS}_{AB|A,B}$. It is possible to show that $\mathrm{SS}_{B|A} = \mathrm{SS}_{A,B} - \mathrm{SS}_A$, where the former is the classical sums of squares of a model without interactions. An ANOVA table such as given in Table 11.3 yields different p-values for $H_0 : \beta_1 = \cdots = \beta_I = 0$ and $H_0 : \gamma_1 = \cdots = \gamma_J = 0$ if the order of the factors is exchanged. This is often a disadvantage and for the F-test the so-called *Type II partial SS*, being $\mathrm{SS}_{A|B}$ and $\mathrm{SS}_{B|A}$ can be used. As there is no interaction involved, we should use Type II only if the interaction is not significant (in which case it is to be preferred over Type I). Alternatively, *Type III partial SS*, $\mathrm{SS}_{A|B,AB}$ and $\mathrm{SS}_{B|A,AB}$, may be used.

In R, the output of `aov`, or `anova` are Type I sequential SS. To obtain the other types, manual calculations may be done or using the function `Anova(..., type=i)` from the package `car`.

Example 12.3. Consider Example 11.2 in Section 11.4 but we eliminate the first observation and the design is unbalanced in both factors. R-Code 12.2 calculates the Type I sequential SS for the same order as in R-Code 11.3. Type II partial SS are subsequently slightly different.

Note that the design is balanced for the factor `Month` and thus simply exchanging the order does not alter the SS here. ♣

R-Code 12.2 Type I and II SS for `UVfilter` data without the first observation.

```
require( car)
lmout2 <- lm( log(OT) ~ Month + Treatment, data=UV, subset=-1) # omit 1st!
print( anova( lmout2), signif.stars=FALSE)
## Analysis of Variance Table
##
## Response: log(OT)
##           Df Sum Sq Mean Sq F value Pr(>F)
## Month      1   1.14   1.137    4.28  0.053
## Treatment  2   5.38   2.692   10.12  0.001
## Residuals 19   5.05   0.266
print( Anova( lmout2, type=2), signif.stars=FALSE)  # type=2 is default
## Anova Table (Type II tests)
##
## Response: log(OT)
##           Sum Sq Df F value Pr(>F)
## Month       1.41  1    5.31  0.033
## Treatment   5.38  2   10.12  0.001
## Residuals   5.05 19
```

12.5 Bibliographic Remarks

Devore (2011) derives sample size calculations for many classical tests in an accessible fashion.

An online lecture about DoE is available at https://online.stat.psu.edu/stat503/home.

There are many different online apps to calculate sample sizes, for example https://w3.math.uzh.ch/shiny/git/reinhard.furrer/SampleSizeR/. For specific operating systems the free software G*Power calculates power and sample sizes for various settings (www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html.

For in-vivo studies, The Experimental Design Assistant https://eda.nc3rs.org.uk/ is a very helpful tool. Classical approaches are PREPARE guidelines (planning guidelines before the study) https://norecopa.no/PREPARE and ARRIVE guidelines (reporting guidelines after the study) https://www.nc3rs.org.uk/arrive-guidelines.

12.6 Exercises and Problems

Problem 12.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Compare sample sizes when using Wilson and Wald type confidence intervals for an proportion.

b) In the context of a simple regression, the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\text{Var}(\widehat{\beta}_0) = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\Big)\,, \qquad \text{Var}(\widehat{\beta}_1) = \sigma^2 \frac{1}{\sum_i (x_i - \overline{x})^2}\,.$$

Assume that $x_1, \dots, x_n \in [a, b]$, with n even. Show that the variances are minimized by choosing half at a and the other half at b. What if n is odd?

Problem 12.2 (Study design) Consider a placebo-controlled trial for a treatment B (compared to a placebo A). The clinician proposes, to use ten patients, who first receive the placebo A and after a long enough period treatment B. Your task is to help the clinician to find an optimal design with at most 20 treatments and with at most 20 patients available.

a) Describe alternative designs, argue regarding which aspects those are better or worse than the original.

b) Give an adequate statistical test for each of your suggestions.

Problem 12.3 (Sample size calculation) Suppose we compare the mean of some treatment in two equally sized groups. Let z_γ denote the γ-quantile of the standard normal distribution. Furthermore, the following properties are assumed to be known or fixed:

- clinically relevant difference $\Delta = \mu_1 - \mu_0$, we can assume without loss of generality that $\Delta > 0$
- standard deviation of the treatment effect σ (same in both groups)
- Power $1 - \beta$.

- Type I error rate α.

a) Write down the suitable test statistic and its distributions under the null hypothesis.

b) Derive an expression for the power using the test statistic.

c) Prove analytically that the required sample size n in each group is at least

$$n = \frac{2\sigma^2(z_{1-\beta} + z_{1-\alpha/2})^2}{\Delta^2}$$

Problem 12.4 (Sample size and group allocation) A randomized clinical trial to compare treatment A to treatment B is being conducted. To this end 20 patients need to be allocated to the two treatment arms.

a) Using R, randomize the 20 patients to the two treatments with equal probability. Repeat the randomization in total a 1000 times retaining the difference in group size and visualize the distribution of the differences with a histogram.

b) In order to obtain group sizes that are closer while keeping randomization codes secure a random permuted block design with varying block sizes 2 and 4 and respective probabilities 0.25 and 0.75 is now to be used. Here, for a given length, each possible block of equal numbers of As and Bs is chosen with equal probability. Using R, randomize the 20 patients to the two treatments using this design. Repeat the randomization in total a 1000 times retaining the difference in group size. What are the possible values this difference my take? How often did these values occur?

Problem 12.5 (BMJ Endgame) Discuss and justify the statements about ‘Sample size: how many participants are needed in a trial?’ given in doi.org/10.1136/bmj.f1041.

Chapter 13

Bayesian Approach

Learning goals for this chapter:

- Describe the fundamental differences between the Bayesian and frequentist approach
- Describe the Bayesian terminology
- Explain how to compute posterior probability
- Interpret posterior probability and a posterior credible interval
- Explain the idea of the Bayes factor and link it to model selection

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter13.R.

In statistics there exist two different philosophical approaches to inference: frequentist and Bayesian inference. Past chapters dealt with the frequentist approach; now we introduce the Bayesian approach. Here, we consider the parameter as a random variable with a suitable distribution, which is chosen a priori, i.e., before the data is collected and analyzed. The goal is to update this prior knowledge after observation of the data in order to draw conclusions (with the help of the so-called posterior distribution).

As an example, suppose I have misplaced my cordless phone in my room. I should start search in rooms where the chance of finding it is highest. Based on past experience, the chance that I might have left the phone in the living room is much larger than on a shelf in the pantry. But there is not much of a difference between the sofa of living room and counter in the kitchen. If I use the locator button on the base station, the phone starts to beep and I can use this information to update and tailor my search.

13.1 Motivating Example and Terminology

Bayesian statistics is often introduced by recalling the so-called *Bayes theorem*, which states for two events A and B

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}, \qquad \text{for } \mathrm{P}(B) \neq 0, \tag{13.1}$$

and is shown by using twice Equation (2.5). Bayes theorem is often used in probability theory to calculate probabilities along an event tree, as illustrated in the arch-example below.

Example 13.1. A patient sees a doctor and gets a test for a (relatively) rare disease. The prevalence of this disease is 0.5%. As typical, the screening test is not perfect and has a sensitivity of 99%, i.e., true positive rate; properly identified the disease in a sick patient, and a specificity of 98%, i.e., true negative rate; a healthy person is correctly identified disease free. What is the probability that the patient has the disease provided the test is positive?

Denoting the events D = 'Patient has disease' and $+$ = 'test is positive' we can use Equation (13.1) to calculate

$$\begin{aligned}\mathrm{P}(D \mid +) &= \frac{\mathrm{P}(+ \mid D)\,\mathrm{P}(D)}{\mathrm{P}(+)} = \frac{\mathrm{P}(+ \mid D)\,\mathrm{P}(D)}{\mathrm{P}(+ \mid D)\,\mathrm{P}(D) + \mathrm{P}(+ \mid D^c)\,\mathrm{P}(D^c)}\\ &= \frac{99\% \cdot 0.5\%}{99\% \cdot 0.5\% + 2\% \cdot 99.5\%} = 20\%.\end{aligned} \tag{13.2}$$

Note that for the denominator we have used the so-called *law of total probability* to get an expression for $\mathrm{P}(+)$. ♣

An interpretation of the previous example from a frequentist view is in terms of proportion of outcomes (in a repeated sampling framework). In the Bayesian approach, we view the probabilities as "degree of belief", where we have some proposition (event D in Example 13.1) and new evidence (event $+$ in Example 13.1). More specifically, $\mathrm{P}(D)$ represents the prior believe of our proposition, $\mathrm{P}(+ \mid D)/\,\mathrm{P}(+)$ is the support of the evidence for the proposition and $\mathrm{P}(D \mid +)$ is the posterior believe of the proposition after having accounted for the new evidence $+$.

Extending Bayes' theorem to the setting of two continuous random variables X and Y along the definition of the conditional density (8.14) we have

$$f_{X\mid Y=y}(x \mid y) = \frac{f_{Y\mid X=x}(y \mid x)\, f_X(x)}{f_Y(y)}, \text{ for all } y \text{ such that } f_Y(y) > 0. \tag{13.3}$$

In the context of Bayesian inference the random variable X will now be a parameter, typically of the distribution of Y:

$$f_{\Theta\mid Y=y}(\theta \mid y) = \frac{f_{Y\mid \Theta=\theta}(y \mid \theta)\, f_\Theta(\theta)}{f_Y(y)}, \text{ for all } y \text{ such that } f_Y(y) > 0. \tag{13.4}$$

Hence, current knowledge about the parameter is expressed by a probability distribution on the parameter: the prior distribution. The model for our observations is called the likelihood. We use our observed data to update the prior distribution and thus obtain the posterior distribution.

In the next section, we discuss examples where the parameter is the success probability of a trial and the mean in a normal distribution.

Notice that $\text{P}(B)$ in (13.1), $\text{P}(+)$ in (13.2), or $f_Y(y)$ in (13.3) and (13.4) serves as a normalizing constant, i.e., it is independent of A, D, x or the parameter θ, respectively. Thus, we often write the posterior without this normalizing constant

$$f_{\Theta|Y=y}(\theta \mid y) \propto f_{Y|\Theta=\theta}(y \mid \theta) \times f_\Theta(\theta), \tag{13.5}$$

(or in short form $f(\theta \mid y) \propto f(y \mid \theta)f(\theta)$ if the context is clear). The symbol "$\propto$" means "proportional to". For simplicity, we will omit the additional constraint that $f(y) > 0$.

Finally, we can summarize the most important result in Bayesian inference. The posterior density is proportional to the likelihood multiplied by the prior density, i.e.,

$$\text{Posterior density} \;\propto\; \text{Likelihood} \;\times\; \text{Prior density} \tag{13.6}$$

In a nutshell, advantages of using a Bayesian framework are:

- formal way to incorporate priori knowledge;
- intuitive interpretation of the posterior;
- much easier to model complex systems;
- no n-dependency of 'significance' and the p-value.

As nothing comes for free, there are also some disadvantages:

- more 'elements' have to be specified for a statistical model;
- in virtually all cases, a Bayesian approach is computationally more demanding.

Until recently, there were clear fronts between frequentists and Bayesians. Luckily, these differences have vanished.

13.2 Bayesian Inference

We illustrate the concept of Bayesian inference with two typical examples that are tractable.

13.2.1 Bayesian Estimates

Example 13.2. (beta-binomial model) If we observe y successes (out of n), a frequentist setting assumes $Y \sim \mathcal{Bin}(n, p)$ and uses as estimate $\widehat{p} = y/n$ (Section 6.1).

In the Bayesian framework we assume the success probability p as a random variable with an associated distribution. We require the support of the associated density to be the interval $(0, 1)$. One example is the uniform distribution $\mathcal{U}(0, 1)$ or the so-called Beta distribution, see Section 13.5.1. The density of a Beta random variable is given by

$$f(p) = c \cdot p^{\alpha-1}(1-p)^{\beta-1}, \qquad p \in [0, 1], \alpha > 0, \beta > 0, \tag{13.7}$$

with normalization constant c. We write $P \sim \mathcal{B}eta(\alpha, \beta)$. Figure 13.6 shows densities for various pairs (α, β).

If we investigate the probability of a lamb being male, then it is highly unlikely that $p < 0.1$ or $p > 0.9$. This additional knowledge about the parameter p would be reflected by using a prior $P \sim \mathcal{B}eta(5, 5)$, for example.

The posterior density is then proportional to

$$\propto \binom{n}{y} p^y (1-p)^{n-y} \times c \cdot p^{\alpha-1}(1-p)^{\beta-1} \tag{13.8}$$

$$\propto p^y p^{\alpha-1}(1-p)^{n-y}(1-p)^{\beta-1} = p^{y+\alpha-1}(1-p)^{n-y+\beta-1}, \tag{13.9}$$

which can be recognized as a beta distribution $\mathcal{B}eta(y + \alpha, n - y + \beta)$.

Figure 13.1 illustrates the case of $y = 10$, $n = 13$ with prior $\mathcal{B}eta(5, 5)$. The posterior mode is now between the prior mode (0.5) and the frequentist estimate $\widehat{p}$.

The expected value of a beta distributed random variable $\mathcal{B}eta(\alpha, \beta)$ is $\alpha/(\alpha + \beta)$ (here the prior distribution). The posterior expected value is thus

$$\mathrm{E}(P \mid Y = y) = \frac{y + \alpha}{n + \alpha + \beta}. \tag{13.10}$$

Specifically, the mean changed from 0.5 to $(10 + 5)/(13 + 5 + 5) \approx 0.65$. ♣

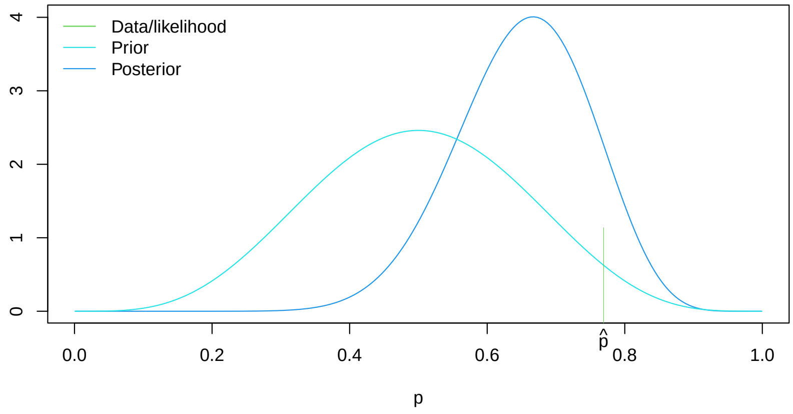


Figure 13.1: Beta-binomial model with prior density (cyan), data/likelihood (green) and posterior density (blue).

In the previous example, we use $P \sim \mathcal{B}eta(\alpha, \beta)$ and fix α and β during model specification, which is why they are called hyper-parameters.

The beta distribution $\mathcal{B}eta(1, 1)$, i.e., $\alpha = 1$, $\beta = 1$, is equivalent to a uniform distribution $\mathcal{U}(0, 1)$. The uniform distribution for the probability p, however, does not mean "information-free". As a result of Equation (13.10), a uniform distribution as prior is "equivalent" to two experiments, of which one is a success. That means, we can see the prior as two pseudo-observations.

In the next example, we have the data and the parameter both continuous.

Example 13.3. (normal-normal model) Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We assume σ is known. The mean μ is the only parameter of interest, for which we assume the prior $\mathcal{N}(\eta, \tau^2)$. Thus, we have the Bayesian model:

$$\begin{aligned} Y_i \mid \mu &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad i = 1, \dots, n, && (13.11) \\ \mu &\sim \mathcal{N}(\eta, \tau^2). && (13.12) \end{aligned}$$

where σ^2, η and τ^2 are considered as hyper-parameters. Notice that we have again slightly abused the notation by using μ as the realization in (13.11) and as the random variable in (13.12). Since the context determines the meaning, we use this simplification for the parameters in the Bayesian context. The posterior density is then

$$\begin{aligned} f(\mu \mid y_1, \dots, y_n) \propto f(y_1, \dots, y_n \mid \mu) \times f(\mu) &= \prod_{i=1}^{n} f(y_i \mid \mu) \times f(\mu) && (13.13) \\ &\propto \prod_{i=1}^{n} \exp\Big(-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\Big) \exp\Big(-\frac{1}{2}\frac{(\mu - \eta)^2}{\tau^2}\Big) && (13.14) \\ &\propto \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu - \eta)^2}{\tau^2}\Big), && (13.15) \end{aligned}$$

where the constants $(2\pi\sigma^2)^{-1/2}$ and $(2\pi\tau^2)^{-1/2}$ do not need to be considered. Through further manipulation (of the square in μ) one obtains

$$\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\left(\mu - \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{n\bar{y}}{\sigma^2} + \frac{\eta}{\tau^2}\right)\right)^2\right) \tag{13.16}$$

and thus the posterior distribution is

$$\mathcal{N}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{n\bar{y}}{\sigma^2} + \frac{\eta}{\tau^2}\right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right). \tag{13.17}$$

In other words, the posterior expected value

$$\mathrm{E}(\mu \mid y_1, \dots, y_n) = \eta\,\frac{\sigma^2}{n\tau^2 + \sigma^2} + \bar{y}\,\frac{n\tau^2}{n\tau^2 + \sigma^2} = \eta\,\omega + \bar{y}\,(1 - \omega) \tag{13.18}$$

is a weighted mean of the prior mean η and the mean of the likelihood $\bar{y}$. The weights $\omega = \sigma^2/(n\tau^2 + \sigma^2)$ depend on the variance parameters and on n. With a smaller prior variance τ, more weight is given to the prior, for example. The larger n is, the less weight there is on the prior mean, since $\sigma^2/(n\tau^2 + \sigma^2) \to 0$ for $n \to \infty$. Typically, the prior is fixed but if more data is collected, the posterior mean will be closer to the mean of the data and the prior has a weaker "influence" on the posterior.

Figure 13.2 (based on R-Code 13.1) illustrates the setting of this example with $\bar{y} = 2.1$, $n = 4$ and the hyper-parameters $\sigma^2 = 1$, $\eta = 0$ and $\tau^2 = 2$. Here, the likelihood is with respect to $\overline{Y}$, i.e., the likelihood is a function of the parameter μ, given by the density of $\overline{Y}$, a Gaussian with mean $\bar{y}$ and variance σ^2/n, see Problem 13.1.**a**. ♣

As a summary statistic of the posterior distribution the posterior mode is often used. Naturally, the posterior median and posterior mean (i.e., expectation of the posterior distribution) are intuitive alternatives. In the case of the previous example, the posterior mode is the same as the posterior mean.

R-Code 13.1 Normal-normal model. (See Figure 13.2.)

```
# Information about data:
ybar <- 2.1;      n <- 4;     sigma2 <- 1
# information about prior:
priormean <- 0;   priorvar <- 2
# Calculating the posterior variance and mean:
postvar <- 1/( n/sigma2 + 1/priorvar)
postmean <- postvar*( ybar*n/sigma2 + priormean/priorvar )
# Plotting follows:
y <- seq(-2, to=4, length=500)
plot( y, dnorm( y, postmean, sqrt( postvar)), type='l', col=4,
     ylab='Density', xlab=bquote(mu))
lines( y, dnorm( y, ybar, sqrt( sigma2/n)), col=3)
lines( y, dnorm( y, priormean, sqrt( priorvar)), col=5)
legend( "topleft", legend=c("Data/likelihood", "Prior", "Posterior"),
       col=c(3, 5, 4), bty='n', lty=1)
```

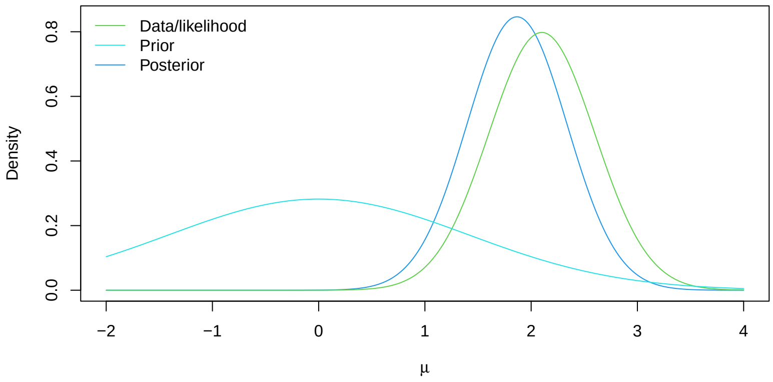


Figure 13.2: Normal-normal model with prior (cyan), data/likelihood (green) and posterior (blue). (See R-Code 13.1.)

13.2.2 Bayesian Confidence intervals

Interval estimation in the frequentist approach results in confidence intervals. But sample confidence intervals need to be interpreted with care, in a context of repeated sampling. A sample $(1-\alpha)\%$ confidence interval $[b_u, b_o]$ contains the true parameter with a frequency of $(1-\alpha)\%$ in infinite repetitions of the experiment. With a Bayesian approach, we can now make statements about the parameter with probabilities. In Example 13.3, based on Equation (13.17)

$$\mathrm{P}\left(v^{-1}m - z_{1-\alpha/2}v^{-1/2} \leq \mu \leq v^{-1}m + z_{1-\alpha/2}v^{-1/2}\right) = 1-\alpha, \tag{13.19}$$

with $v = n/\sigma^2 + 1/\tau^2$ and $m = n\bar{y}/\sigma^2 + \eta/\tau^2$. That means that the bounds $v^{-1}m \pm z_{1-\alpha/2}v^{-1/2}$ can be used to construct a Bayesian counterpart to a confidence interval.

Definition 13.1. The interval R with

$$\int_R f(\theta \mid y_1, \dots, y_n)\, \mathsf{d}\theta = 1 - \alpha \tag{13.20}$$

is called a $(1-\alpha)\%$ credible interval for θ with respect to the posterior density $f(\theta \mid y_1, \dots, y_n)$ and $1-\alpha$ is the credible level of the interval. ♢

The definition states that the parameter θ, now seen as a random variable whose posterior density is given by $f(\theta \mid y_1, \dots, y_n)$, is contained in the $(1-\alpha)\%$ credible interval with probability $(1-\alpha)$.

Example 13.4 (continuation of Example 13.3)**.** The interval $[\,0.94, 2.79\,]$ is a 95% credible interval for the parameter μ. ♣

Since the credible interval for a fixed α is not unique, the "narrowest" is often used. This is the so-called HPD interval (highest posterior density interval). HPD intervals and credible intervals in general are often determined numerically.

Example 13.5 (continuation of Example 13.2)**.** The 2.5% and 97.5% quantiles of the posterior (13.9) are 0.45 and 0.83, respectively. A HPD is given by the bounds 0.46 and 0.84. The differences are not pronounced as the posterior density is fairly symmetric. Hence, the widths of both are almost identical: 0.377 and 0.375.

The frequentist sample 95% CI is $[0.5, 0.92]$, with width 0.42, see Equation (6.9). ♣

13.2.3 Predictive Distribution

In the classical regression framework, the estimated regression line represented the mean of an unobserved new location. To fully assess the uncertainty of the prediction, we had to take into account the uncertainty of the estimates and argued that the prediction is given by a t-distribution (see, CI 7).

In the Bayesian setting, the likelihood $f(y_{\text{new}} \mid \theta)$ can be seen as a the density of the predictive distribution. That means, the distribution of an unobserved new observation y_{new}. As the classical regression framework, using $f(y_{\text{new}} \mid \widehat{\theta}\,)$, with $\widehat{\theta}$ some Bayesian estimate of the parameter (e.g., posterior mean or posterior mode). The better approach is based on the posterior predictive distribution, defined as follows.

Definition 13.2. The *posterior predictive distribution* of a Bayesian model with likelihood $f(y \mid \theta)$ and prior $f(\theta)$ is

$$f(y_{\text{new}} \mid y_1, \dots, y_n) = \int f(y_{\text{new}} \mid \theta) f(\theta \mid y_1, \dots, y_n)\, \mathsf{d}\theta. \tag{13.21}$$ ♢

In the previous equation, $f(y_{\text{new}} \mid \theta, y_1, \dots, y_n)$ represents the likelihood and thus there is no dependency on the data. Hence $f(y_{\text{new}} \mid \theta, y_1, \dots, y_n) = f(y_{\text{new}} \mid \theta)$.

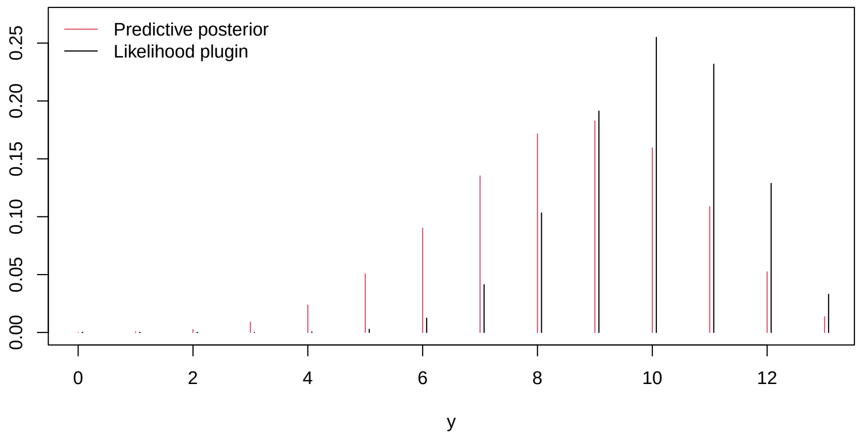


Figure 13.3: Predictive posterior distribution for the beta-binomial model (red) and likelihood with plugin parameter $\widehat{p} = 10/13$ (black).

Example 13.6 (continuation of Example 13.2)**.** In the context of the beta-binomial model, the posterior predictive distribution is constructed based on the single observation y only

$$\begin{aligned} f(y_{\text{new}} \mid y) &= \int_0^1 f(y_{\text{new}} \mid p) \times f(p \mid y)\, \mathsf{d}p \\ &= \binom{n}{y_{\text{new}}} c \int_0^1 p^{y_{\text{new}}}(1-p)^{n-y_{\text{new}}} \times p^{y+\alpha-1}(1-p)^{n-y+\beta-1}\, \mathsf{d}p, \end{aligned} \tag{13.22}$$

where c is the normalizing constant for the posterior. The integral itself gives us the normalizing constant of a $\mathcal{B}eta(y_{\text{new}} + y + \alpha, 2n - y_{\text{new}} - y + \beta)$ distribution. We do not recognize this distribution per se. As illustration, Figure 13.3 shows the posterior predictive distribution based on the observation $y = 10$ and prior $\mathcal{B}eta(5, 5)$. The prior implies that the posterior predictive distribution is much more centered compared to the likelihood with plugin parameter $\widehat{p} = 10/13$ (i.e., the binomial distribution $\mathcal{B}in(13, 10/13)$).

R-Code 13.2 Predictive distribution with the beta-binomial model. (See Figure 13.3.)

```
library(LearnBayes)
n <- 13
y <- 0:n
pred.probs <- pbetap(c( 10+5, 13-10+5), n, y)  # prior Beta(5,5)
plot(y, pred.probs, type="h", ylim=c(0,.27), col=2, ylab='')
lines( y+0.07, dbinom(y, size=n, prob=10/13), type='h')
legend("topleft", legend=c("Predictive posterior", "Likelihood plugin"),
       col=2:1, lty=1, bty='n')
```

Even in the simple case of the beta-binomial model, it is not straightforward to derive the predictive posterior distribution. Quite often, more detailed integration knowledge is required. In the case of the normal-normal model as introduced in Example 13.3, it is possible to show that the posterior predictive distribution is again normal $\mathcal{N}(\mu_{\text{post}}, \sigma^2 + \sigma^2_{\text{post}})$, where μ_{post}, σ^2_{post} are the posterior mean and posterior variance as given in (13.17).

13.2.4 Bayes Factors

The Bayesian counterpart to hypothesis testing is done through a comparison of posterior probabilities. For example, consider two specific models specified by two hypotheses H_0 and H_1. By Bayes theorem,

$$\underbrace{\frac{\mathrm{P}(H_0 \mid y_1,\dots,y_n)}{\mathrm{P}(H_1 \mid y_1,\dots,y_n)}}_{\text{Posterior odds}} = \underbrace{\frac{\mathrm{P}(y_1,\dots,y_n \mid H_0)}{\mathrm{P}(y_1,\dots,y_n \mid H_1)}}_{\text{Bayes factor }(\mathrm{BF}_{01})} \times \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{Prior odds}}, \tag{13.23}$$

that means that the Bayes factor BF_{01} summarizes the evidence of the data for the hypothesis H_0 versus the hypothesis H_1. The Bayes factor is any positive number. However, it has to be mentioned that a Bayes factor needs to exceed 3 to talk about substantial evidence for H_0. For strong evidence we typically require Bayes factors larger than 10. More precisely, Jeffreys (1983) differentiates

$$1 < \begin{array}{l}\text{barely worth}\\ \text{mentioning}\end{array} < 3 < \text{substantial} < 10 < \text{strong} < 30 < \begin{array}{l}\text{very}\\ \text{strong}\end{array} < 100 < \text{decisive}$$

For values smaller than one, we would favor H_1 and the situation is similar by inverting the ratio, as also illustrated in the following example.

Example 13.7. We consider the setup of Example 13.2 and compare the models with $p = 1/2$ and $p = 0.8$ when observing 10 successes among the 13 trials. To calculate the Bayes factor, we need to calculate $\mathrm{P}(Y = 10 \mid p)$ for $p = 1/2$ and $p = 0.8$. Hence, the Bayes factor is

$$\mathrm{BF}_{01} = \frac{\binom{13}{10}0.5^{10}(1-0.5)^3}{\binom{13}{10}0.8^{10}(1-0.2)^3} = \frac{0.0349}{0.2457} = 0.1421, \tag{13.24}$$

which is somewhat substantial ($1/0.1421 \approx 7$) in favor of H_1. This is not surprising, as the observed proportion is $\widehat{p} = 10/13 = 0.77$ close to $p = 0.8$ corresponding to H_1. ♣

In the example above, the hypotheses H_0 and H_1 are understood in the sense of $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. The situation for an unspecified alternative $H_1 : \theta \neq \theta_0$ is much more interesting and relies on using the prior $f(\theta)$ and integrating out the parameter θ:

$$f(y_1,\dots,y_n \mid H_1 : \theta \neq \theta_0) = \int f(y_1,\dots,y_n \mid \theta) f(\theta)\,\mathrm{d}\theta, \tag{13.25}$$

illustrated as follows.

Example 13.8 (continuation of Example 13.7)**.** For the situation $H_1 : p \neq 0.5$ using the prior $\mathcal{Beta}(5,5)$, we have

$$\begin{aligned}\mathrm{P}(Y = 13 \mid H_1) &= \int_0^1 \mathrm{P}(Y = 13 \mid p) f(p)\,\mathrm{d}p\\ &= \int_0^1 \binom{13}{10} p^{10}(1-p)^3 \cdot c\,p^4(1-p)^4\,\mathrm{d}p = 0.0704,\end{aligned} \tag{13.26}$$

where we used `integrate( function(p) dbinom(10,13,prob=p)*dbeta(p, 5,5),0,1)`. Thus, $\mathrm{BF}_{01} = 0.0349/0.0704 = 0.4957$. Hence, the Bayes factor is approximately 2 in favor of H_1, barely worth calculating the value. Under a uniform prior, the support for H_1 only marginally increases (from 2.017 to 2.046). ♣

Example 13.9 (continuation of Example 13.3, data from Example 5.8)**.** We now look at a Bayesian extension of the frequentist t-test. For simplicity we assume the one sample setting without deriving the explicit formulas. The package *BayesFactor* provides functionality to calculate Bayes factors for different settings.

We take the pododermatitis scores (see R-Code 5.1). R-Code **??** shows that the Bayesfactor comparing the null model $\mu = 3.33$ against the alternative $\mu \neq 3.33$ is approximately 14. Here, we have used the standard parameter setting which includes the specification of the prior and the prior variance. The prior variance can be specified with the argument *rscale* with default $0.707 = \sqrt{2}$. Increasing this variance leads to a flatter prior and thus to a smaller Bayes factor. Default priors are typically very reasonable and we come back to the choice of the priors in the next section. ♣

R-Code 13.3 Bayes factor within the normal-normal model.

```
library(BayesFactor)
ttestBF(PDHmean, mu=3.33)    # data may be reloaded.
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 14.557 ±0%
##
## Against denominator:
##   Null, mu = 3.33
## ---
## Bayes factor type: BFoneSample, JZS
```

Bayes factors are popular because they are linked to the BIC (Bayesian Information Criterion) and thus automatically penalize model complexity. Further, they also work for non-nested models.

13.3 Choice and Effect of Prior Distribution

The choice of the prior distribution belongs to the modeling process just like the choice of the likelihood distribution. Naturally, the prior should be fixed before the data has been collected.

The examples in the last section were such, that the posterior and prior distributions did belong to the same class. Naturally, this is no coincidence and such prior distributions are called *conjugate prior* distributions.

With other prior distributions we may obtain posterior distributions that we no longer "recognize" and normalizing constants must be explicitly calculated. An alternative approach to integrating the posterior density is discussed in Chapter 14.

Beside conjugate priors, there are many more classes that are typically discussed in a full Bayesian lecture. We prefer to classify the effect of the prior instead. Although not universal and quite ambigiuous, we differentiate between *informative*, *weakly informative* and *uninformative* priors. The first describes a prior that is specific for the data at hand where with different data the prior typically chances. The prior has potentially a substantial effect on the posterior. Weakly informative priors do not have a strong influence on the posterior but they may substantially contribute to the model in situations where the model is ill posed. Finally, an uninformative prior is such, that the likelihood relates essentially to the posterior in terms of some criterion like posterior mean.

Uninformative priors are not classical prior distributions. In Example 13.2 we would require a "beta density" with $\alpha = \beta = 0$ in order to have a posterior mean that is equivalent to the likelihood estimate (see Equation (13.10)). However, for $\alpha = \beta = 0$ the normalizing constant of (13.7) is not finite as $\int_0^1 p^{-1}(1-p)^{-1}\,\mathsf{d}p$ diverges. Similarly, in Example 13.3, in order that $\mathrm{E}(\mu \mid y_1, \dots, y_n) = \bar{y}$, we need $\tau \to \infty$, that means, the prior of μ is "completely constant" (see Equation (13.18)). As we do not have a bounded range for μ, we have again not a "proper density", i.e., a so-called *improper prior*.

Without going into details, it is possible that for certain improper priors the posterior is a letigimate density.

For large n the difference between a Bayesian and likelihood estimate is not pronounced. As a matter of fact, it is possible to show that the posterior mode converges to the likelihood estimate as the number of observations increase.

Example 13.10. We consider again the normal-normal model and compare the posterior density for various n with the likelihood. We keep $\bar{y} = 2.1$, independent of n. As shown in Figure 13.4, the maximum likelihood estimate does not depend on n ($\bar{y}$ is kept constant by design). However, the uncertainty decreases (standard error is $\sigma/\sqrt{n}$). For increasing n, the posterior approaches the likelihood density. In the limit, there is no difference between the posterior and the likelihood. The R-Code follows closely R-Code 13.1. ♣

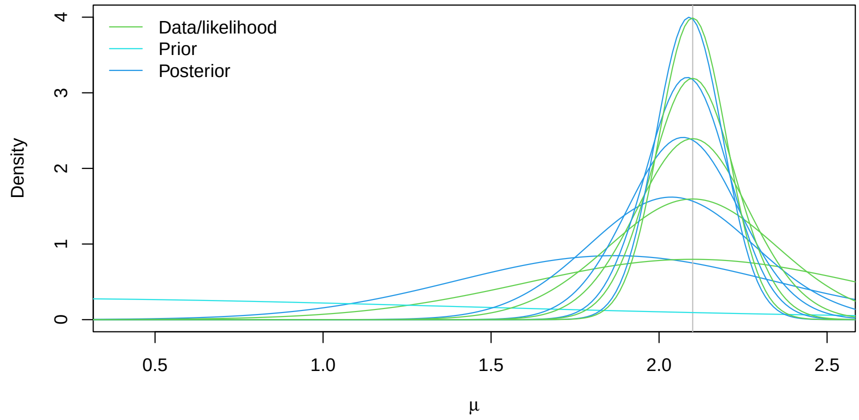


Figure 13.4: Normal-normal model with prior (cyan), data/likelihood (green) and posterior (blue) for increasing n ($n = 4, 36, 64, 100$). Prior is $\mathcal{N}(0, 2)$ and $\bar{y} = 2.1$ in all cases.

13.4 Regression in a Bayesian Framework

In this section we introduce a Bayesian approach to simple linear and logistic regression. More complex models are deferred to Chapter 14. Here, we will discuss the conceptual ideas and use software tools as a black box approach. The underlying computational principles will be discussed in Chapter 14.

The simplest Bayesian regression model is as follows

$$Y_i \mid \boldsymbol{\beta}, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2), \qquad i = 1, \dots, n, \tag{13.27}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}(\boldsymbol{\eta}, \sigma^2\mathbf{T}), \tag{13.28}$$

where σ^2, $\boldsymbol{\eta}$ and $\mathbf{T}$ are hyper-parameters. The model is quite similar to (13.11) and (13.12) with the exception that we have a multivariate prior distribution for $\boldsymbol{\beta}$. Instead of the parameter τ^2 we use $\sigma^2\mathbf{T}$, where $\mathbf{T}$ is a $(p+1) \times (p+1)$ symmetric positive definite matrix. The special form will allow us to factor σ^2 and simplify the posterior. With a few steps, it is possible to show that

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim \mathcal{N}_{p+1}(\mathbf{V}^{-1}\boldsymbol{m}, \sigma^2\mathbf{V}^{-1}) \tag{13.29}$$

with $\mathbf{V}^{-1} = \mathbf{T}^{-1} + \mathbf{X}^\top\mathbf{X}$ and $\boldsymbol{m} = \mathbf{T}^{-1}\boldsymbol{\eta} + \mathbf{X}^\top\boldsymbol{y}$ (see Problem 13.1.**b**).

It is possible to show that the posterior is a weighted average of the prior mean $\boldsymbol{\eta}$ and the classical least squares estimate $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\boldsymbol{y}$ (see Problem 13.1.**c**).

The function *bayesglm()* from the R package *arm* implements an accessible way for simple linear regression and logistic regression. It is simple in the sense that it returns the posterior modes of the estimates in a framework that is similar to a frequentist approach. We need to specify the priors for the regression coefficients (separately for the intercept and the remaining coefficients).

Example 13.11 (Bayesian approach to *orings* data)**.** We revisit Example 10.5 and fit in R-Code 13.4 a Bayesian logistic model to the data.

In the first *bayesglm()* call, we set the prior variances to infinity, resulting in uninformative priors. The posterior mode is identical to the result of a classical *glm()* model fit.

In the second call, we use Gaussian priors for both parameters with mean zero and variance 9. This choice is set by *prior.df=Inf* (i.e., a t-distribution with infinite degrees of freedom), by the default *prior.mean=0*, and by *prior.scale=3*. and similarly for the intercept parameter. The slope parameter is hardly affected by the prior. The intercept is, because with its rather informative choice of the prior variance, the posterior mode is shrunk towards zero.

Note that *summary(baye)* should not be used, as the printed p-values are not relevant in the Bayesian context. The function *display()* is the preferred way. ♣

Remark 13.1. For one particular class, consider several grades of some students. A possible model might be

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{13.30}$$

where α_i represents the performance relative to the overall mean. This performance is not "fixed", it highly depends on the choice of the student and is thus variable. Hence, it would make more

R-Code 13.4 `orings` data and estimated probability of defect dependent on air temperature. (See Figure 13.5.)

```
require(arm)
data( orings, package="faraway")
bayes1 <- bayesglm( cbind(damage,6-damage)~temp, family=binomial, data=orings,
            prior.scale=Inf, prior.scale.for.intercept=Inf) #
coef(bayes1)

## (Intercept)        temp
##    11.66299    -0.21623

# result "similar" to
# coef( glm( cbind(damage,6-damage)~temp, family=binomial, data=orings))

bayes2 <- bayesglm( cbind(damage,6-damage)~temp, family=binomial, data=orings,
            prior.df = Inf, prior.scale=3,
            prior.df.for.intercept=Inf, prior.scale.for.intercept=3) #
arm::display(bayes2)

## bayesglm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
##     data = orings, prior.scale = 3, prior.df = Inf, prior.scale.for.intercept = 3,
##     prior.df.for.intercept = Inf)
##             coef.est coef.se
## (Intercept) 10.54     3.03
## temp        -0.20     0.05
## ---
## n = 23, k = 2
## residual deviance = 17.1, null deviance = 38.9 (difference = 21.8)

plot( damage/6~temp, xlim=c(21,80), ylim=c(0,1), data=orings, pch='+',
     xlab="Temperature [F]", ylab='Probability of damage', cex=1.5) # data
glm1 <- glm( cbind(damage,6-damage)~temp, family=binomial, data=orings)
ct <- seq(20, to=85, length=100)                       # vector to predict
p.out <- predict( glm1, new=data.frame(temp=ct), type="response")
lines(ct, p.out, lwd=3, col=4)
scoefs <- coef(sim(bayes2)) # simulation of coefficients...
for (i in 1:100) {
   lines( ct, invlogit( scoefs[i,1]+scoefs[i,2]*ct), col=rgb(.8,.8,.8,.2))
}
```

sense to consider α_i as random, or, more specifically, $\alpha_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$, and α_i and ε_{ij} independent. Such models are called *mixed effects models* in contrast to the *fixed effects model* as discussed in this book.

From a Bayesian perspective, such a separation is not necessary, as all Bayesian linear models are a mixed-effects model. In the example above, we impose as prior $\alpha_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$. ♣

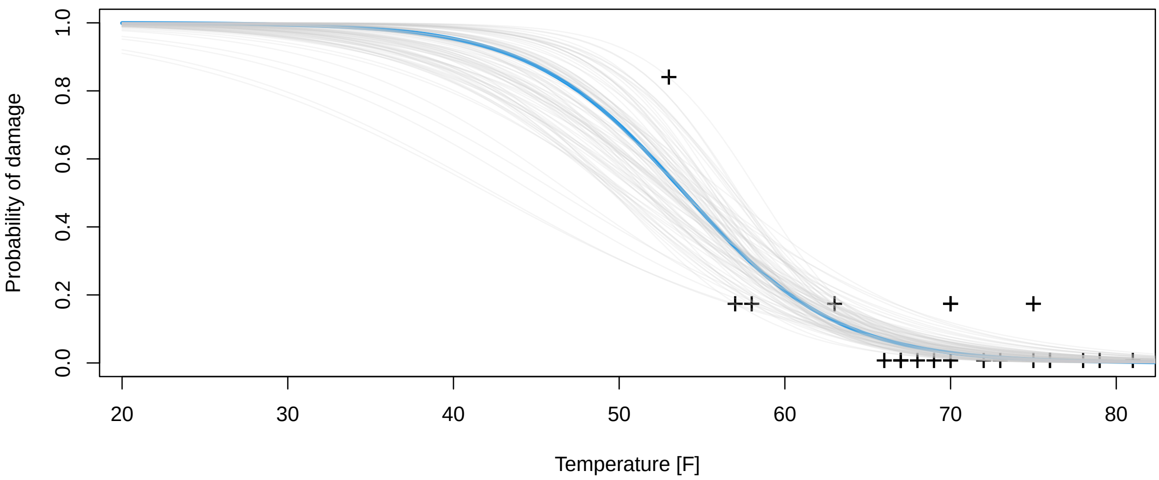


Figure 13.5: `orings` data (proportion of damaged orings, black crosses), estimated probability of defect dependent on air temperature by a logistic regression (blue line). Gray lines are based on draws from the posterior distribution. (See R-Code 13.4.)

13.5 Appendix: Distributions Common in Bayesian Analysis

13.5.1 Beta Distribution

We introduce a random variable with support $[0, 1]$. Hence this random variable is well suited to model probabilities (proportions, fractions) in the context of Bayesian modeling.

A random variable Y with density

$$f_Y(y) = c \cdot y^{\alpha-1}(1-y)^{\beta-1}, \qquad y \in [0,1], \alpha > 0, \beta > 0, \tag{13.31}$$

where c is a normalization constant, is called beta distributed with parameters α and β. We write this as $Y \sim \mathcal{B}eta(\alpha, \beta)$. The normalization constant cannot be written in closed form for all parameters α and β. For $\alpha = \beta$ the density is symmetric around $1/2$, for $\alpha > 1$, $\beta > 1$ the density is unimodal with mode $(\alpha-1)/(\alpha+\beta-2)$ and for $0 < \alpha < 1$, $0 < \beta < 1$ the density has a bathtub shape. For arbitrary $\alpha > 0$, $\beta > 0$ we have:

$$\mathrm{E}(Y) = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \tag{13.32}$$

Figure 13.6 shows densities of the beta distribution for various pairs of (α, β).

13.5.2 Gamma Distribution

Many results involving the variance parameters of a normal distribution are simpler if we would work with the precision, i.e., the inverse of the variance. A prior for the precision should take

R-Code 13.5 Densities of beta distributed random variables for various pairs of (α, β). (See Figure 13.6.)

```r
p <- seq( 0, to=1, length=100)
a.seq <- c( 1:6, .8, .4, .2, 1, .5, 2)
b.seq <- c( 1:6, .8, .4, .2, 4, 4, 4)
col <- c( 1:6, 1:6)
lty <- rep( 1:2, each=6)
plot( p, dbeta( p, 1, 1),  type='l', ylab='Density',  xlab='x',
     xlim=c(0,1.3), ylim=c(0,3))
for ( i in 2:length(a.seq))
    lines( p, dbeta(p, a.seq[i], b.seq[i]), col=col[i], lty=lty[i])
legend("topright",  col=c(NA,col), lty=c(NA, lty), cex=.9, bty='n',
       legend=c(expression( list( alpha, beta)), paste(a.seq, b.seq, sep=',')))
```

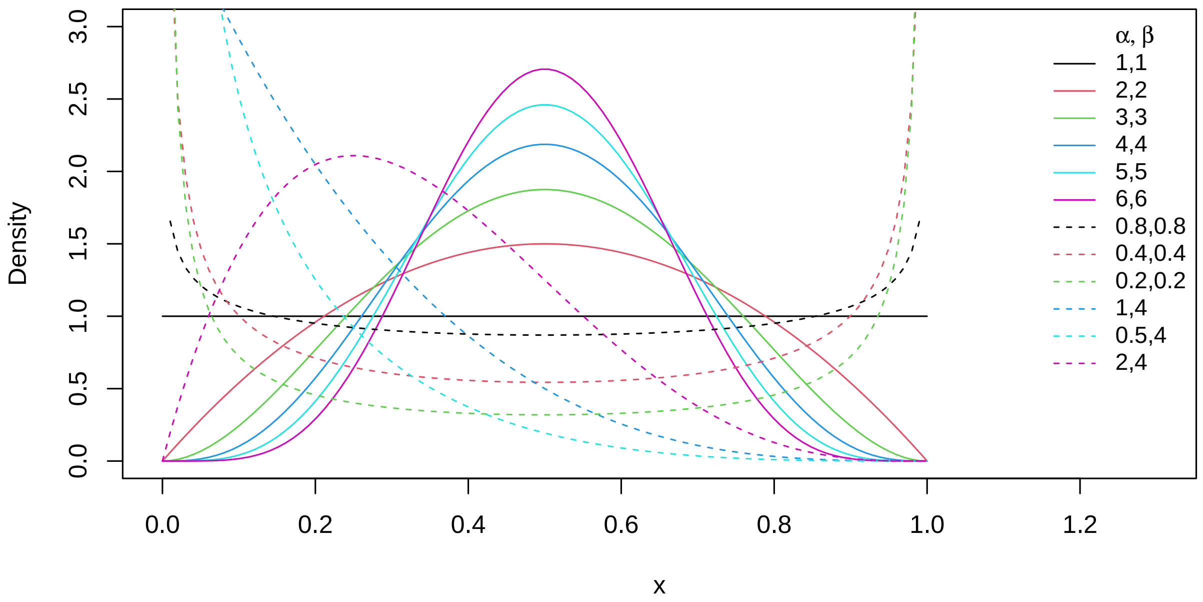


Figure 13.6: Densities of beta distributed random variables for various pairs of (α, β). (See R-Code 13.5.)

any positive value and the gamma distribution is a natural choice because of its conjugacy with the normal likelihood.

A random variable Y with density

$$f_Y(y) = f(y \mid \alpha, \beta) = c \cdot y^{\alpha-1} \exp(-\beta y), \qquad y > 0, \alpha > 0, \beta > 0, \tag{13.33}$$

is called gamma distributed with parameters α and β. We write $Y \sim\sim \mathcal{G}am(\alpha, \beta)$. The normalization constant c cannot be written in closed form for all parameters α and β.

For arbitrary $\alpha > 0$, $\beta > 0$ we have:

$$\mathrm{E}(Y) = \frac{\alpha}{\beta}, \qquad \mathrm{Var}(Y) = \frac{\alpha}{\beta^2}. \tag{13.34}$$

The parameters α and β are also called the shape and rate parameter, respectively. The parameterizations of the density in terms of the scale parameter $1/\beta$ is also frequently used.

13.5.3 Inverse Gamma Distribution

Many results involving the variance parameters of a normal distribution would be simpler if we would work with the precision, i.e., the inverse of the variance. In such cases we choose a so-called inverse-gamma distribution for the parameter $\tau = 1/\sigma^2$.

A random variable Y is said to be distributed according an inverse-gamma distribution if $1/Y$ is distributed according a gamma distribution. For parameters α and β, the density is given by

$$f_Y(y) = f(y \mid \alpha, \beta) = c \cdot y^{-\alpha-1} \exp(-\beta/y), \qquad y > 0, \alpha > 0, \beta > 0, \tag{13.35}$$

and we write $Y \sim \mathcal{IGam}(\alpha, \beta)$. The normalization constant c cannot be written in closed form for all parameters α and β.

For arbitrary $\beta > 0$ and for $\alpha > 1$ and $\alpha > 2$, we have respectively

$$\mathrm{E}(Y) = \frac{\beta}{\alpha - 1}, \qquad \mathrm{Var}(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}. \tag{13.36}$$

13.6 Bibliographic Remarks

An accessible discussion of the Bayesian approach can be found in Held and Sabanés Bové (2014), including a discussion about the choice of prior distribution. A classic is Bernardo and Smith (1994).

The online open access book "An Introduction to Bayesian Thinking" available at https://statswithr.github.io/book/ nicely melds theory and R source code.

The source https://www.nicebread.de/grades-of-evidence-a-cheat-sheet compares different categorizations of evidence based on a Bayes factor and illustrates that the terminology is not universal.

13.7 Exercises and Problems

Problem 13.1 (Theoretical derivations) In this problem we derive some of the theoretical and mathematical results that we have stated in the chapter.

a) Show that the likelihood of Example 13.3 can be written as $c \cdot \exp\big(-n(\bar{y} - \mu)^2/\sigma^2\big)$, where c is the normalizing constant that does not depend on μ.

b) Derive the posterior distribution of $\boldsymbol{\beta} \mid \boldsymbol{y}$.

c) Show that the posterior mean of $\boldsymbol{\beta} \mid \boldsymbol{y}$ can be written as

$$\Big((\sigma^2\mathbf{T})^{-1} + (\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})^{-1}\Big)^{-1}\Big((\sigma^2\mathbf{T})^{-1}\boldsymbol{\eta} + (\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})^{-1}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}\Big)^{-1} \tag{13.37}$$

and give an interpretation with respect to $\sigma^2\mathbf{T}$ and $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Problem 13.2 (Sunrise problem) The sunrise problem is formulated as "what is the probability that the sun rises tomorror?" Laplace formulated this problem by casting it into his *rule of succession* which calculates the probability of a success after having observed y successes out of n trials (the $(n+1)$th is again independent of the previous ones).

a) Formulate the rule of succession in a Bayesian framework and calculate the expected probability for the $n+1$th term.

b) Laplace assumed that the Earth was created about 6000 years ago. If we use the same information, what is the probability that the sun rises tomorrow?

Problem 13.3 (Normal-gamma model) Let $Y_1, Y_2, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1/\kappa)$. We assume that the value of the expectation μ is known (i.e., we treat it as a constant in our calculations), whereas the precision, i.e., inverse of the variance, denoted here with κ, is the parameter of interest.

a) Write down the likelihood of this model.

b) We choose the a gamma prior for the parameter κ, i.e., $\kappa \sim \mathcal{G}am(\alpha, \beta)$. How does this distribution relates to the exponential distribution?

Plot four densities for $(\alpha, \beta) = (1,1)$, $(1,2)$, $(2,1)$ and $(2,2)$. How does a certain choice of α, β be interpreted with respect to our "beliefs" on κ?

c) Derive the posterior distribution of the precision κ.

d) Compare the prior and posterior distributions. Why is the choice in **b**) sensible?

e) Simulate some data with $n = 50$, $\mu = 10$ and $\kappa = 0.25$. Plot the prior and posterior distributions of κ for $\alpha = 2$ and $\beta = 1$.

Problem 13.4 (Bayesian statistics) For the following Bayesian models, derive the posterior distribution and give an interpretation thereof in terms of prior and data.

a) Let $Y \mid \mu \sim \mathcal{N}(\mu, 1/\kappa)$, where κ is the precision (inverse of the variance) and is assumed to be known (hyper-parameter). Further, we assume that $\mu \sim \mathcal{N}(\eta, 1/\nu)$, for fixed hyper-parameters η and $\nu > 0$.

b) Let $Y \mid \lambda \sim \mathcal{P}ois(\lambda)$ with a prior $\lambda \sim \mathcal{G}am(\alpha, \beta)$ for fixed hyper-parameters $\alpha > 0$, $\beta > 0$.

c) Let $Y \mid \theta \sim \mathcal{U}(0, \theta)$ with a prior a shifted Pareto distribution with parameters $\gamma > 0$ and $\xi > 0$, whose density is

$$f(\theta; \gamma, \xi) \propto \theta^{-(\gamma+1)}\mathbb{I}_{\theta>\xi}(\theta). \tag{13.38}$$

Chapter 14

Monte Carlo Methods

Learning goals for this chapter:

- ⋄ Explain how to numerically approximate integrals
- ⋄ Explain how to sample from an arbitrary (univariate) distribution in R
- ⋄ Describe qualitatively Gibbs sampling
- ⋄ Explain the idea of a Bayesian hierarchical model
- ⋄ Able to interpret the output of a MCMC sampler of a simple model

R-Code for this chapter: www.math.uzh.ch/furrer/download/sta120/chapter14.R.

In Chapter 13, the posterior distribution of several examples were similar to the chosen prior distribution albeit with different parameters. Specifically, for binomial data with a beta prior, the posterior is again beta. This was no coincidence; rather, we chose so-called conjugate priors based on our likelihood (distribution of the data).

With other prior distributions, we may have “complicated”, not standard posterior distributions, for which we no longer know the normalizing constant, the expected value or any other moment in general. Theoretically, we could derive the normalizing constant and then in subsequent steps determine the expectation and the variance (via integration) of the posterior. The calculation of these types of integrals is often complex and so here we consider classic simulation procedures as a solution to this problem. In general, so-called *Monte Carlo* simulation is used to numerically solve a complex problem through repeated random sampling.

In this chapter, we start with illustrating the power of Monte Carlo simulation where we utilize, above all, the law of large numbers. We then discuss one method to draw a sample from an arbitrary density and, finally, illustrate a method to derive (virtually) arbitrary posterior densities by simulation. We conclude the chapter with a few realistic examples.

14.1 Monte Carlo Integration

In this section we discuss how to approximate integrals. Let X be a continuous random variable and $f_X(x)$ be its density function and $g(x)$ an arbitrary (sufficiently “well behaved”) function. We aim to evaluate expected value of $g(X)$, i.e.,

$$\mathrm{E}\big(g(X)\big) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathsf{d}x. \tag{14.1}$$

Hence, $g(x)$ cannot be entirely arbitrary, but such that the integral on the right-hand side of (14.1) is well defined. An approximation of this integral is (along the idea of method of moments)

$$\mathrm{E}\big(g(X)\big) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathsf{d}x \approx \widehat{\mathrm{E}\big(g(X)\big)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i), \tag{14.2}$$

where $x_1, \dots, x_n$ is a realization of a random sample with density $f_X(x)$. The method relies on the law of large numbers (see Section 3.3).

Example 14.1. To estimate the expectation of a χ^2_1 random variable we can use `mean( rnorm( 100000)^2)`, yielding 1 with a couple digits of precision, close to what we expect according to Equation (3.6).

Of course, we can use the same approach to calculate arbitrary moments of a χ^2_n or $F_{n,m}$ distribution. ♣

We now discuss this justification in slightly more details. We consider a continuous function g (over the interval $[a, b]$) and the integral $I = \int_a^b g(x)\,\mathsf{d}x$. There exists a value ξ such that $I = (b-a)g(\xi)$ (often termed as the mean value theorem for definite integrals). We do not know ξ nor $g(\xi)$, but we hope that the “average” value of g is close to $g(\xi)$. More formally, let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{U}(a, b)$ which we use to calculate the average (the density of X_i is $f_X(x) = 1/(b-a)$ over the interval $[a, b]$ and zero elsewhere). We now show that on average, our approximation is correct:

$$\begin{aligned}\mathrm{E}(\widehat{I}\,) &= \mathrm{E}\Big((b-a)\frac{1}{n}\sum_{i=1}^{n} g(X_i)\Big) = (b-a)\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(g(X_i)) = (b-a)\,\mathrm{E}\big(g(X)\big)\\ &= (b-a)\int_a^b g(x) f_X(x)\,\mathsf{d}x = (b-a)\int_a^b g(x)\frac{1}{b-a}\,\mathsf{d}x = \int_a^b g(x)\,\mathsf{d}x = I\,.\end{aligned} \tag{14.3}$$

We can generalize this to almost arbitrary densities $f_X(x)$ having a sufficiently large support:

$$\widehat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(x_i)}{f_X(x_i)}, \tag{14.4}$$

where the justification is as in (14.3). The density in the denominator takes the role of an additional weight for each term.

Similarly, to integrate over a rectangle $\mathcal{R}$ in two dimensions (or a cuboid in three dimensions, etc.), we use a uniform random variable for each dimension. More specifically, let $\mathcal{R} = [a, b] \times [c, d]$ then

$$\int_{\mathcal{R}} g(x, y)\,\mathsf{d}x\,\mathsf{d}y = \int_a^b\int_c^d g(x, y)\,\mathsf{d}x\,\mathsf{d}y \approx (b-a)(d-c)\frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i), \tag{14.5}$$

where $x_1, \dots, x_n$ and $y_1, \dots, y_n$, are two samples of $\mathcal{U}(a, b)$ and of $\mathcal{U}(c, d)$, respectively.

To approximate $\int_{\mathcal{A}} g(x, y)\,\mathsf{d}x\,\mathsf{d}y$ for some complex domain $\mathcal{A} \subset \mathbb{R}^2$. We choose a bivariate random vector having a density $f_{X,Y}(x, y)$ whose support contains $\mathcal{A}$. For example we define a rectangle $\mathcal{R}$ such that $\mathcal{A} \subset \mathcal{R}$ and let $f_{X,Y}(x, y) = (b - a)(d - c)$ over $\mathcal{R}$ and zero otherwise. We define the indicator function $\mathbb{I}_{\mathcal{A}}(x, y)$ that is 1 if $(x, y) \in \mathcal{A}$ and zero otherwise. Then we have the general formula

$$\int_{\mathcal{A}} g(x, y)\,\mathsf{d}x\,\mathsf{d}y = \int_a^b \int_c^d \mathbb{I}_{\mathcal{A}}(x, y) g(x, y)\,\mathsf{d}x\,\mathsf{d}y \approx \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\mathcal{A}}(x_i, y_i) \frac{g(x_i, y_i)}{f_{X,Y}(x_i, y_i)}. \tag{14.6}$$

Testing if a point (x_i, y_i) is in the domain $\mathcal{A}$ is typically an easy problem.

We now illustrate this idea with two particular examples.

Example 14.2. Consider the bivariate normal density specified in Example 8.4 and suppose we are interested in evaluating the probability that $\mathrm{P}(X > Y^2)$. To approximate this probability we can draw a large sample of the bivariate normal density and calculate the proportion for which $x_i > y_i^2$, as illustrated in R-Code 14.1 and yielding 10.47%.

In this case, the function g is the density with which we are drawing the data points. Hence, Equation (14.6) reduces to calculate the proportion of the data satisfying $x_i > y_i^2$. ♣

R-Code 14.1 Calculating probability with the aid of a Monte Carlo simulation

```
set.seed( 14)
require(mvtnorm)                             # to sample the bivariate normals
l.sample <- rmvnorm( 10000, mean=c(0,0), sigma=matrix( c(1,2,2,5), 2))
mean( l.sample[,1] > l.sample[,2]^2)  #  calculate the proportion

## [1] 0.1047
```

Example 14.3. The area of the unit circle is π as well as the volume of a cylinder placed at the origin with height one. To estimate π we estimate the volume of the cylinder and we consider $\mathcal{U}(-1, 1)$ for both coordinates of a square that contains the unit circle. The function $g(x, y) = 1$ is the identity function, $\mathbb{I}_{\mathcal{A}}(x, y)$ is the indicator function of the set $\mathcal{A} = \{x^2 + y^2 \leq 1\}$ and $f_{X,Y}(x_i, y_i) = 1/4$ for $0 \neq x, y \neq 1$. We have the following approximation of the number π

$$\pi = \int_{-1}^1 \int_{-1}^1 \mathbb{I}_{\mathcal{A}}(x, y)\,\mathsf{d}x\,\mathsf{d}y \approx 4 \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\mathcal{A}}(x_i, y_i), \tag{14.7}$$

where $x_1, \dots, x_n$ and $y_1, \dots, y_n$, are two independent samples of $\mathcal{U}(-1, 1)$. Equation (14.6) reduces to calculate a proportion again.

It is important to note that the convergence is very slow, see Figure 14.1. It can be shown that the rate of convergence is of the order of $1/\sqrt{n}$. ♣

In practice, more efficient "sampling" schemes are used. More specifically, we do not sample uniformly but deliberately "stratified". There are several reasons to sample randomly stratified but the discussion is beyond the scope of the work here.

R-Code 14.2 Approximation of π with the aid of Monte Carlo integration. (See Figure 14.1.)

```
set.seed(14)
m <- 49                    # calculate for 49 different n
n <- round( 10+1.4^(1:m))  # non-equal spacing
piapprox <- numeric(m)     # to store the approximation
for (i in 1:m) {
   st <- matrix( runif( 2*n[i]), ncol=2)       # bivariate uniform
   piapprox[i] <- 4*mean( rowSums( st^2)<= 1)  # proportion
}
plot( n, abs( piapprox-pi)/pi, log='xy', type='l') # plotting on log-log scale
lines( n, 1/sqrt(n), col=2, lty=2)                 # order of convergence
sel <- 1:7*7                                       # subset for printing
cbind( n=n[sel], pi.approx=piapprox[sel], rel.error=     # summaries
      abs( piapprox[sel]-pi)/pi, abs.error=abs( piapprox[sel]-pi))
```

```
##              n pi.approx  rel.error  abs.error
## [1,]        21    2.4762 0.21180409 0.66540218
## [2,]       121    3.0083 0.04243968 0.13332819
## [3,]      1181    3.1634 0.00694812 0.02182818
## [4,]     12358    3.1662 0.00783535 0.02461547
## [5,]    130171    3.1403 0.00040166 0.00126186
## [6,]   1372084    3.1424 0.00025959 0.00081554
## [7,]  14463522    3.1406 0.00032656 0.00102592
```

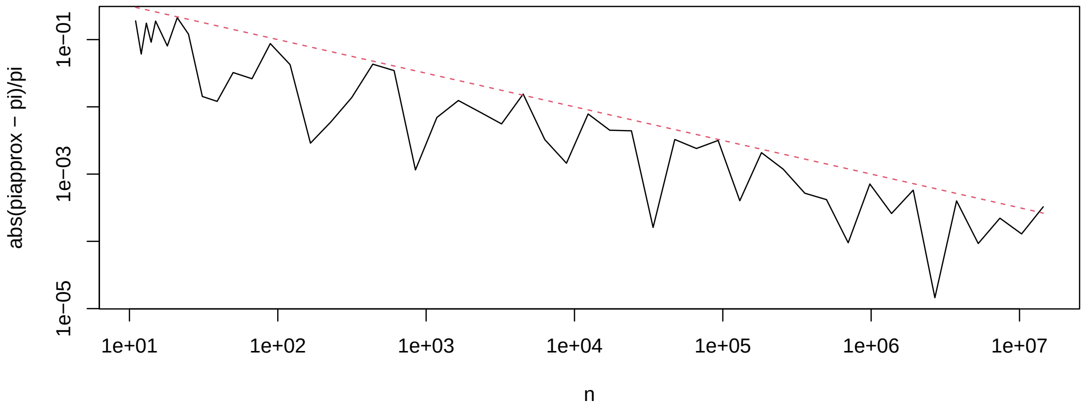


Figure 14.1: Convergence of the approximation for π: the relative error as a function of n (log-log scale). (See R-Code 14.2.)

14.2 Rejection Sampling

We now discuss an approach to sample from a distribution with density $f_Y(y)$ when no direct method exists. There are many of such approaches and we discuss an intuitive but inefficient one here. The approach is called rejection sampling.

In this method, values from a known density $f_Z(z)$ (proposal density) are drawn and through rejection of "unsuitable" values, observations of the density $f_Y(y)$ (target density) are generated. This method can also be used when the normalizing constant of $f_Y(y)$ is unknown and we write $f_Y(y) = c \cdot f^*(y)$.

The procedure is as follows: Step 0: Find an $m < \infty$, so that $f^*(y) \leq m \cdot f_Z(y)$ for all y. Step 1: draw a realization $\tilde{y}$ from $f_Z(y)$ and a realization u from a standard uniform distribution $\mathcal{U}(0,1)$. Step 2: if $u \leq f^*(\tilde{y})/\big(m \cdot f_Z(\tilde{y})\big)$ then $\tilde{y}$ is accepted as a simulated value from $f_Y(y)$, otherwise $\tilde{y}$ is dicarded and no longer considered. We cycle along Steps 1 and 2 until a sufficiently large sample has been obtained. The algorithm is illustrated in the following example.

Example 14.4. The goal is to draw a sample from a $\mathcal{B}eta(6,3)$ distribution with the rejection sampling method. That means, $f_Y(y) = c \cdot y^{6-1}(1-y)^{3-1}$ and $f^*(y) = y^5(1-y)^2$. As proposal density we use a uniform distribution, hence $f_Z(y) = \mathbb{I}_{0 \leq y \leq 1}(y)$. We select $m = 0.02$, which fulfills the condition $f^*(y) \leq m \cdot f_Z(y)$ since `optimize( function(x) x^5*(1-x)^2, c(0, 1), maximum=TRUE)` is roughly 0.152.

An implementation of the example is given in R-Code 14.3. Of course, `f_Z` is always one here. The R-Code can be optimized with respect to speed. It would then, however, be more difficult to read.

Figure 14.2 shows a histogram and the density of the simulated values. By construction the bars of the target density are smaller than the one of the proposal density. In this particular example, we have sample size 285. ♣

R-Code 14.3: Rejection sampling in the setting of a beta distribution. (See Figure 14.2.)

```
set.seed( 14)
n.sim <- 1000
m <- 0.02
fstar <- function(y) y^( 6-1) * (1-y)^(3-1)        # unnormalized target
f_Z <- function(y) ifelse( y >= 0 & y <= 1, 1, 0) # proposal density
result <- sample <- rep( NA, n.sim)                # to store the result
for (i in 1:n.sim){
    sample[i] <- runif(1)                                # ytilde, proposal
    u <- runif(1)                                        # u, uniform
    if( u < fstar( sample[i]) /( m * f_Z( sample[i])) )  # if accept ...
        result[i] <- sample[i]                           #   ... keep
}
mean( !is.na(result))                  # proportion of accepted samples
## [1] 0.285
result <- result[ !is.na(result)]      # eliminate NAs
hist( sample, xlab="y", main="", col="lightblue") # hist of all proposals
hist( result, add=TRUE, col=4)                     #   of the kept ones
curve( dbeta(x, 6, 3), frame =FALSE, ylab="", xlab='y', yaxt="n")
```

```
lines( density( result), lty=2, col=4)
legend( "topleft", legend=c("truth", "smoothed empirical"),
       lty=1:2, col=c(1,4))
```

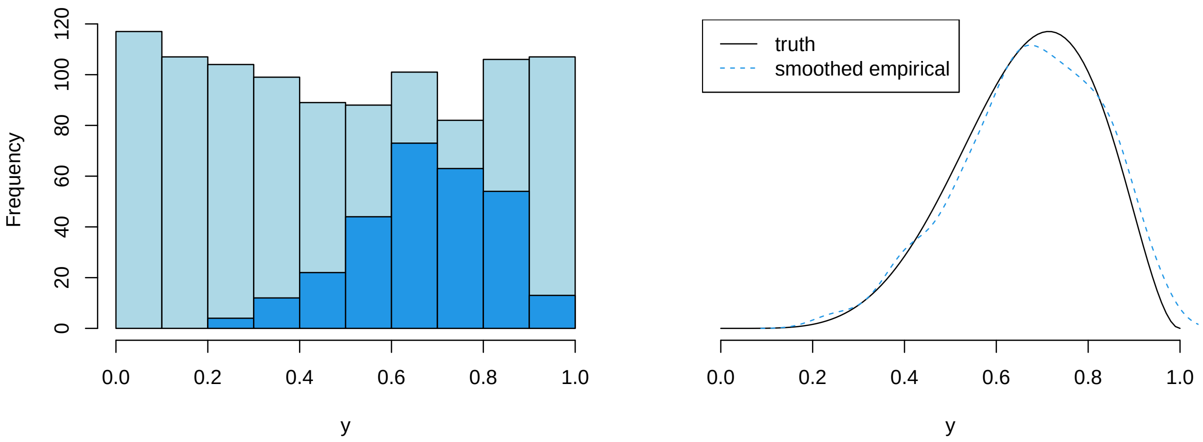


Figure 14.2: Left panel: histogram of the simulated values of $f_Z(y)$ (light blue) and $f_Y(y)$ (dark blue). Right panel: theoretical density (truth) black and the simulated density (smoothed empirical) blue dashed. (See R-Code 14.3.)

For efficiency reasons the constant m should be chosen to be as small as possible to reduce the number of rejections. Nevertheless in practice, rejection sampling is intuitive but often quite inefficient. The next section illustrates an approach well suited for complex Bayesian models.

14.3 Gibbs Sampling

The idea of Gibbs sampling is to simulate the posterior distribution through the use of a so-called *Markov chain*. This algorithm belongs to the family of Markov chain Monte Carlo (MCMC) methods. We illustrate the principle in a Bayesian context with a likelihood that depends on two parameters θ_1 and θ_2. Based on some prior, the joint posterior density is written as $f(\theta_1, \theta_2 \mid y)$ with, for simplicity, a single observation y. The Gibbs sampler reduces the problem to two one-dimensional simulations $f(\theta_1 \mid \theta_2, y)$ and $f(\theta_2 \mid \theta_1, y)$. Starting with some initial value $\theta_{2,0}$ we draw $\theta_{1,1}$ from $f(\theta_1 \mid \theta_{2,0}, y)$, followed by $\theta_{2,1}$ from $f(\theta_2 \mid \theta_{1,1}, y)$ and $\theta_{1,2}$ from $f(\theta_1 \mid \theta_{2,1}, y)$, etc. If all is setup properly, the sample $(\theta_{1,i}, \theta_{2,i})$, $i = 1, \dots, n$, is a sample of the posterior density $f(\theta_1, \theta_2 \mid y)$. Often we omit the first few samples to avoid influence of possibly sub-optimal initial values.

In many cases one does not have to program a Gibbs sampler oneself but can use a pre-programmed sampler. We use the sampler JAGS (Just Another Gibbs sampler) (Plummer, 2003) with the R-Interface package `rjags` (Plummer, 2016).

R-Codes 14.4, 14.5 and 14.6 give a short, but practical overview into MCMC methods with JAGS in the case of a simple Gaussian likelihood. Luckily more complex models can easily be constructed based on the approach shown here.

When using MCMC methods, you may encounter situations in which the sampler does not converge (or converges too slowly). In such a case the posterior distribution *cannot* be approximated with the simulated values. It is therefore important to examine the simulated values for eye-catching patterns. For example, the so-called *trace-plot*, observations in function of the index, as illustrated in the right panel of Figure 14.3 is often used.

Example 14.5. R-Code 14.4 implements the normal-normal model for a single observation, $y = 1$, $n = 1$, known variance, $\sigma^2 = 1.1$, and a normal prior for the mean μ:

$$Y \mid \mu \sim \mathcal{N}(\mu, 1.1), \tag{14.8}$$

$$\mu \sim \mathcal{N}(0, 0.8). \tag{14.9}$$

The basic approach to use JAGS is to first create a file containing the Bayesian model definition. This file is then transcribed into a model graph (function `jags.model()`) from which we can finally draw samples (`coda.samples()`).

Defining a model for JAGS is quite straightforward, as the notation is very close to the one fro, R. Some care is needed when specifying variance parameters. In our notation, we typically use the variance σ^2, as in $\mathcal{N}(\,\cdot\,,\sigma^2)$; in R we have to specify the standard deviation σ as parameter `sd` in the function `dnorm(..., sd=sigma)`; and in JAGS we have to specify the precision $1/\sigma^2$ in the function `dnorm(..., precision=1/sigma2)`, see also LeBauer *et al.* (2013).

The resulting samples are typically plotted with smoothed densities, as seen in the left panel of Figure 14.3 with prior and likelihood, if possible. The posterior seems affected similarly by likelihood (data) and prior, the mean is close to the average of the prior mean and the data. More precisely, the prior is slightly tighter as its variance is slightly smaller (0.8 vs. 1.1), thus the posterior mean is slightly closer to the prior mean than to y. The setting here is identical to Example 13.3 and thus the posterior has again a normal distribution with $\mathcal{N}\big(0.8/(0.8+1.1), 0.8\cdot 1.1/(0.8+1.1)\big)$, see Equation (13.17). ♣

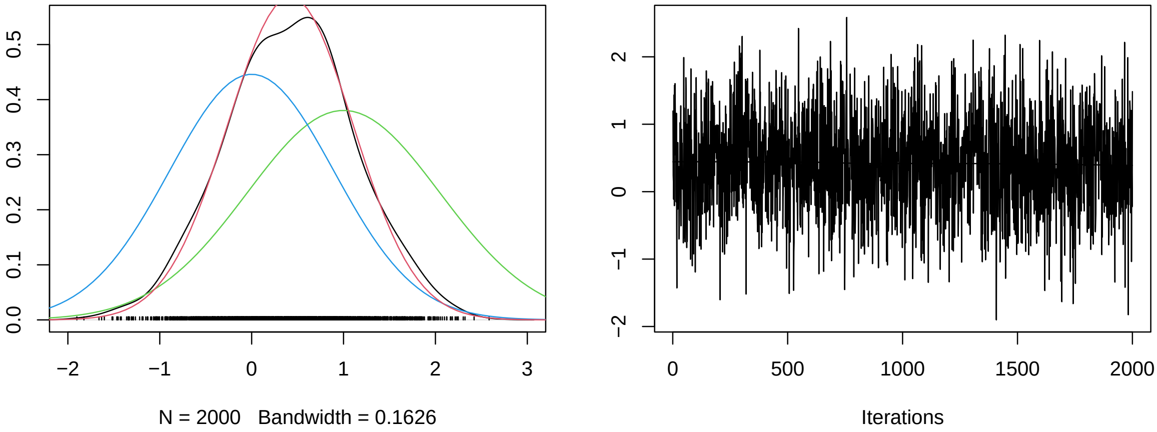


Figure 14.3: Left: empirical densities: MCMC based posterior (black), exact (red), prior (blue), likelihood (green). Right: trace-plot of the posterior $\mu \mid y = 1$. (See R-Code 14.4.)

R-Code 14.4 JAGS sampler for normal-normal model, with $n = 1$.

```
require( rjags)
writeLines("model {                    # File with  Bayesian model definition
              y ~ dnorm( mu, 1/1.1)    # here Precision = 1/Variance
              mu ~ dnorm( 0, 1/0.8)    # Precision again!
            }",    con="jags01.txt")   # arbitrary file name
jagsModel <- jags.model( "jags01.txt", data=list( 'y'=1))  # transcription
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 8
##
## Initializing model
postSamples <- coda.samples( jagsModel, 'mu', n.iter=2000) # draw samples

plot( postSamples, trace=FALSE, main="", auto.layout=FALSE, xlim=c(-2, 3))
y <- seq(-3, to=4, length=100)
lines( y, dnorm( y, 1, sqrt(1.1)), col=3)             # likelihood
lines( y, dnorm( y, 0, sqrt(0.8)   ), col=4)          # prior
lines( y, dnorm( y, 1/1.1 * (1.1*0.8/(0.8 + 1.1)),
                sqrt(1.1*0.8/(0.8 + 1.1))), col=2)    # posterior
plot( postSamples, density=FALSE, main="", auto.layout=FALSE)
```

Example 14.6. R-Code 14.5 extends the normal-normal model to $n = 10$ observations with known variance:

$$Y_1, \dots, Y_n \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1.1), \tag{14.10}$$

$$\mu \sim \mathcal{N}(0, 0.8). \tag{14.11}$$

We draw the data in R via *rnorm(n, 1, sqrt(1.1))* and proceed similarly as in R-Code 14.4. Figure 14.4 gives the empirical and exact densities of the posterior, prior and likelihood and shows a trace-plot as a basic graphical diagnostic tool. The density of the likelihood is $\mathcal{N}(\bar{y}, 1.1/\sqrt{n})$, the prior density is based on (14.11) and the posterior density is based on (13.17). The latter simplifies considerably because we have $\eta = 0$ in (14.11).

As the number of observations increases, the data gets more "weight". From (13.18), the weight increases from $0.8/(0.8 + 1.1) \approx 0.42$ to $0.8n/(0.8n + 1.1) \approx 0.88$. Thus, the posterior is "closer" to the likelihood but slightly more peaked. As both the variance of the data and the variance of the priors are comparable, the prior has a comparable impact on the posterior as if we would possess an additional observation with value zero. ♣

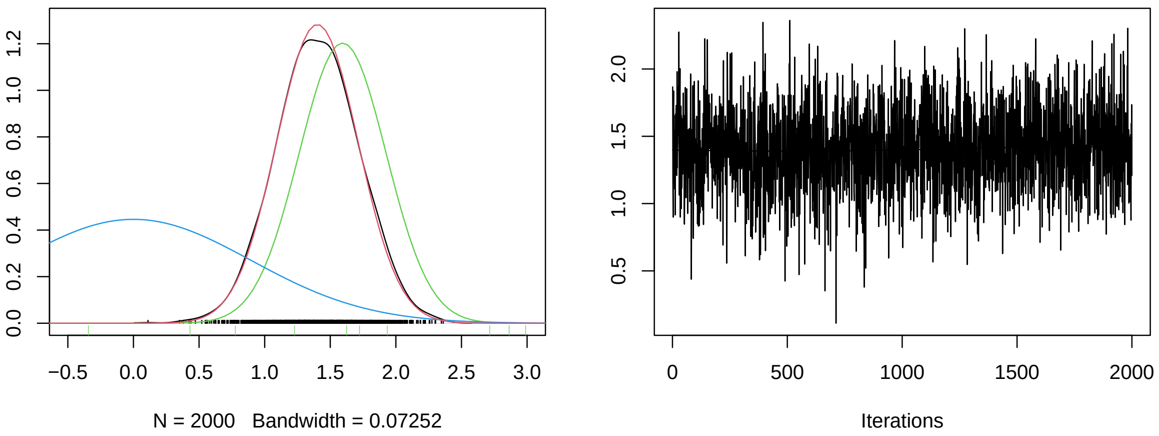


Figure 14.4: Left: empirical densities: MCMC based posterior $\mu \mid y = y_1, \ldots, y_n$, $n = 10$ (black), exact (red), prior (blue), likelihood (green). Black and green ticks are posterior sample and observations, respectively. Right: trace-plot of the posterior. (See R-Code 14.5.)

R-Code 14.5: JAGS sampler for the normal-normal model, with $n = 10$. (See Figure 14.4.)

```
set.seed( 4)
n <- 10
obs <- rnorm( n, 1, sqrt(1.1))    # generate artificial data
writeLines("model {
              for (i in 1:n) {               # define a likelihood for each
                 y[i] ~ dnorm( mu, 1/1.1)    # individual observation
              }
              mu ~ dnorm( 0, 1/0.8)
           }",       con="jags02.txt")
jagsModel <- jags.model( "jags02.txt", data=list('y'=obs, 'n'=n), quiet=T)
postSamples <- coda.samples( jagsModel, 'mu', n.iter=2000)

plot( postSamples, trace=FALSE, main="", auto.layout=FALSE,
     xlim=c(-.5, 3), ylim=c(0, 1.3))
rug( obs, col=3)
y <- seq(-.7, to=3.5, length=100)
lines( y, dnorm( y, mean(obs), sqrt(1.1/n)), col=3)   # likelihood
lines( y, dnorm( y, 0, sqrt(0.8)   ), col=4)          # prior
lines( y, dnorm( y, n/1.1*mean(obs)*(1.1*0.8/(n*0.8 + 1.1)),
                sqrt(1.1*0.8/(n*0.8 + 1.1)) ), col=2) # posterior
plot( postSamples, density=FALSE, main="", auto.layout=FALSE)
```

Example 14.7. We consider an extension of the previous example by including an unknown variance, respectively unknown precision. That means that we now specify two prior distributions and we have a priori no knowledge of the posterior and cannot compare the empirical posterior density with a true (bivariate) density (as we had the red densities in Figures 14.3 and 14.4).

R-Code 14.6 implements the following model in JAGS:

$$Y_i \mid \mu, \kappa \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1/\kappa), \qquad i = 1, \dots, n, \text{ with } n = 10, \tag{14.12}$$

$$\mu \sim \mathcal{N}(\eta, 1/\lambda), \qquad \text{with } \eta = 0,\ \lambda = 1.25, \tag{14.13}$$

$$\kappa \sim \mathcal{Gam}(\alpha, \beta), \qquad \text{with } \alpha = 1,\ \beta = 0.2. \tag{14.14}$$

For more flexibility with the code, we also pass the hyper-parameters $\eta, \lambda, \alpha, \beta$ to the JAGS MCMC engine.

Figure 14.5 gives the marginal empirical posterior densities of μ and κ, as well as the priors (based on (14.13) and (14.14)) and likelihood (based on (14.12)). The posterior is quite data driven and by the choice of the prior, slightly shrunk towards zero.

Note that the marginal likelihood for μ is $\mathcal{N}(\bar{y}, s^2/\sqrt{n})$, i.e., we have replaced the parameters in the model with their unbiased estimates. The marginal likelihood for κ is a gamma distribution based on parameters $n/2 + 1$ and $ns^2/2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/2$, see Problem 13.3.a. ♣

R-Code 14.6 JAGS sampler for priors on mean and precision parameter, with $n = 10$.

```r
eta <- 0         # start with defining the four hyperparameters
lambda <- 1.25   # corresponds to a variance 0.8, as in previous examples
alpha <- 1
beta <- 0.2
writeLines("model {                             # JAGS model as above with ...
             for (i in 1:n) {
                y[i] ~ dnorm( mu, kappa)
             }
             mu ~ dnorm( eta, lambda)
             kappa ~ dgamma(alpha, beta)        # ... one additional prior
          }",   con="jags03.txt")
jagsModel <- jags.model('jags03.txt', quiet=T, data=list('y'=obs, 'n'=n,
                'eta'=eta, 'lambda'=lambda, 'alpha'=alpha, 'beta'=beta))
postSamples <- coda.samples(jagsModel, c('mu','kappa'), n.iter=2000)
plot( postSamples[,"mu"], trace=FALSE, auto.layout=FALSE,
      xlim=c(-1,3), ylim=c(0, 1.3))
y <- seq( -2, to=5, length=100)
lines( y, dnorm(y, 0, sqrt(1.2)   ), col=4)             # likelihood
lines( y, dnorm(y, mean(obs), sd(obs)/sqrt(n)), col=3)   # prior
plot( postSamples[,"kappa"], trace=FALSE, auto.layout=FALSE, ylim=c(0, 1.3))
y <- seq( 0, to=5, length=100)
lines( y, dgamma( y, 1, .2), type='l', col=4)           # likelihood
lines( y, dgamma( y, n/2+1, (n-1)*var(obs)/2 ), col=3)   # prior
```

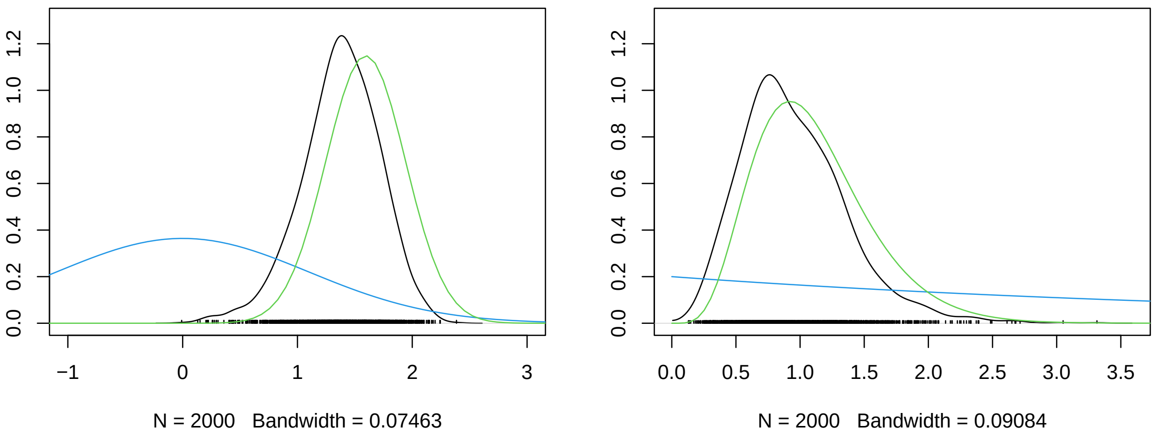


Figure 14.5: Empirical posterior densities of $\mu \mid y_1, \dots, y_n$ (left) and $\kappa = 1/\sigma^2 \mid y_1, \dots, y_n$ (right), MCMC based (black), prior (blue), likelihood (green). (See R-Code 14.6.)

This last example is another classical Bayesian example and with a very careful specification of the priors, we can construct a closed form posterior density. Problem 14.1 gives a hint towards this more advanced topic.

Remark 14.1. The marginal distribution of μ in the normal-normal-gamma model is a (shifted and scaled) t-distribution. ♣

Note that the function `jags.model()` writes some local files that may be cleaned after the analysis.

14.4 Bayesian Hierarchical Models

We conclude this chapter with a final rather flexible class of models, called *Bayesian hierarchical models*.

A hierarchical Bayesian model is a Bayesian model in which the prior distribution of some of the parameters depends on further parameters to which we assign priors too.

Suppose that we observe a linear relationship between two variables. The relationship may be different for different subjects. A very simple hierarchical Bayesian model is the following

$$Y_{ij} \mid \boldsymbol{\beta}_i, \kappa \overset{\text{indep}}{\sim} \mathcal{N}(\boldsymbol{x}_{ij}\boldsymbol{\beta}_i, 1/\kappa), \qquad i = 1, \dots, n, j = 1, \dots, n_i, \tag{14.15}$$

$$\boldsymbol{\beta}_i, \mid \boldsymbol{\eta}, \lambda \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\eta}, \mathbf{I}/\lambda), \tag{14.16}$$

$$\boldsymbol{\eta} \sim \mathcal{N}_p(\boldsymbol{\eta}_0, \mathbf{I}/\tau), \quad \lambda \sim \mathcal{G}am(\alpha_\lambda, \beta_\lambda), \quad \kappa \sim \mathcal{G}am(\alpha_\kappa, \beta_\kappa). \tag{14.17}$$

where τ, α_λ, β_λ, α_κ and β_κ are hyperparameters. The three levels (14.15) to (14.17) are often referred to as observation level, state or process level and prior level.

Example 14.8. Parasite infection can pose a large economic burden on livestock such as sheep, horses etc. Infected herds or animals receive an anthelmintic treatment that reduces the infection of parasitic worms. To assess the efficacy of the treatment, the number of parasitic eggs per gram feces are evaluated. We use the following Poisson model for the pre- and post-treatment counts, denoted as y_i^C and y_i^T:

$$Y_i^C \mid \mu_i^C \overset{\text{indep}}{\sim} \mathcal{P}\mathit{ois}(\mu_i^C), \qquad Y_i^T \mid \mu_i^T \overset{\text{indep}}{\sim} \mathcal{P}\mathit{ois}(\delta\mu_i^C), \qquad i = 1, \dots, n, \tag{14.18}$$

$$\mu_i^C \sim \mathcal{G}\mathit{am}(\kappa, \kappa/\mu). \tag{14.19}$$

$$\delta \sim \mathcal{U}(0,1), \quad \kappa \sim \mathcal{G}\mathit{am}(1, 0.001), \quad \mu \sim \mathcal{G}\mathit{am}(1, 0.7), \tag{14.20}$$

where we have used directly the numerical values for the hyperparameters (Wang *et al.*, 2017). The parameter δ represents the efficacy of the treatment. Notice that with the parameterization of the Gamma distribution for μ_i^C the mean thereof is μ.

The package `eggCounts` provides the dataset `epgs` containing 14 eggs per gram (epg) values in sheep before and after anthelmintic treatment of benzimidazole. The correction factor of the diagnostic technique was 50 (thus all numbers are multiples of 50, due to the measuring technique, see also Problem 14.4). R-Code 14.7 illustrates the implementation in JAGS. We reduce all sheep that did not have any parasites, followed by the setup of the JAGS file with `dpois()`, `dunif()` and `dgamma()` for the distributions of the different layers of the hierarchy. Although the arguments of the gamma distribution are named differently in JAGS we can pass the same parameters in the same order.

The sampler does not indicate any convergence issues (trace-plots and the empirical posterior densities behave well). The posterior median reduction factor is about 0.077, with a 95% HPD interval $[0.073, 0.0812]$, virtually identical to the quantile based credible interval $[0.0729, 0.0812]$, (`TeachingDemos::emp.hpd(postSamples[,"delta"][[1]], conf=0.95)` and `summary( postSamples)$quantiles["delta",c(1,5)]`). The posterior median epg is reduced from 1885.8 to 145.

As a sanity check, we can compare the posterior median (or mean values) with corresponding frequentist estimates. The average epgs before and after the treatment are 2094.4 and 161.1 (`colMeans(epgs2)`).

To compare the variances we can use the following ad-hoc approach. The (prior) variance of μ_i^C is μ^2/κ. Although the posterior distribution of μ_i^C is not gamma anymore, we use the same formula to estimate the variance (here we have very few observations and for a simple Poisson likelihood, a gamma prior is conjugate, see Problem 13.4.**b**). Based on the posterior medians, we have the following values 5.326×10^6 and 3.15×10^4, which are somewhat smaller than the frequentist values 9.007×10^6 and 6.861×10^4. We should not be too surprised about the difference but rather be assured that we have properly specified and interpreted the parameters of the gamma distribution. ♣

R-Code 14.7: JAGS sampler for *epgs* data. (See Figure 14.6.)

```r
require(eggCounts)
require(rjags)
data(epgs)
epgs2 <- epgs[rowSums(epgs[,c("before","after")])>0,c("before","after")]
```

```
n <- nrow(epgs2)
writeLines("model {
              for (i in 1:n) {               # define a likelihood for each
                 yC[i] ~ dpois( muC[i])        # pre-treatment
                 yT[i] ~ dpois( delta*muC[i])  # post-treatment
                 muC[i] ~ dgamma( kappa, kappa/mu)  # pre-treatment mean
              }
              delta ~ dunif( 0, 1)         # reduction
              mu ~ dgamma(1, 0.001)        # pre-treatment mean
              kappa ~ dgamma(1, 0.7)       #
           }",       con="jagsEggs.txt")
jagsModel <- jags.model( "jagsEggs.txt",   # write the model
                  data=list('yC'=epgs2[,1],'yT'=epgs2[,2], 'n'=n), quiet=T)
postSamples <- coda.samples( jagsModel,  # run sampler and monitor all param.
                             c('mu', 'kappa', 'delta'), n.iter=5000)

summary( postSamples)
##
## Iterations = 1001:6000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean       SD Naive SE Time-series SE
## delta    0.077 2.13e-03 3.01e-05       4.10e-05
## kappa    0.684 2.60e-01 3.68e-03       5.21e-03
## mu    1982.866 7.06e+02 9.99e+00       1.52e+01
##
## 2. Quantiles for each variable:
##
##           2.5%      25%      50%      75%    97.5%
## delta   0.0727 7.56e-02    0.077 7.85e-02 8.12e-02
## kappa   0.2892 4.96e-01    0.644 8.29e-01 1.30e+00
## mu    943.9016 1.49e+03 1864.182 2.35e+03 3.71e+03
par(mfcol=c(2,3), mai=c(.6,.6,.1,.1))
plot( postSamples[,"delta"], main="", auto.layout=FALSE, xlab=bquote(delta))
plot( postSamples[,"mu"], main="", auto.layout=FALSE, xlab=bquote(mu))
plot( postSamples[,"kappa"], main="", auto.layout=FALSE, xlab=bquote(kappa))
```

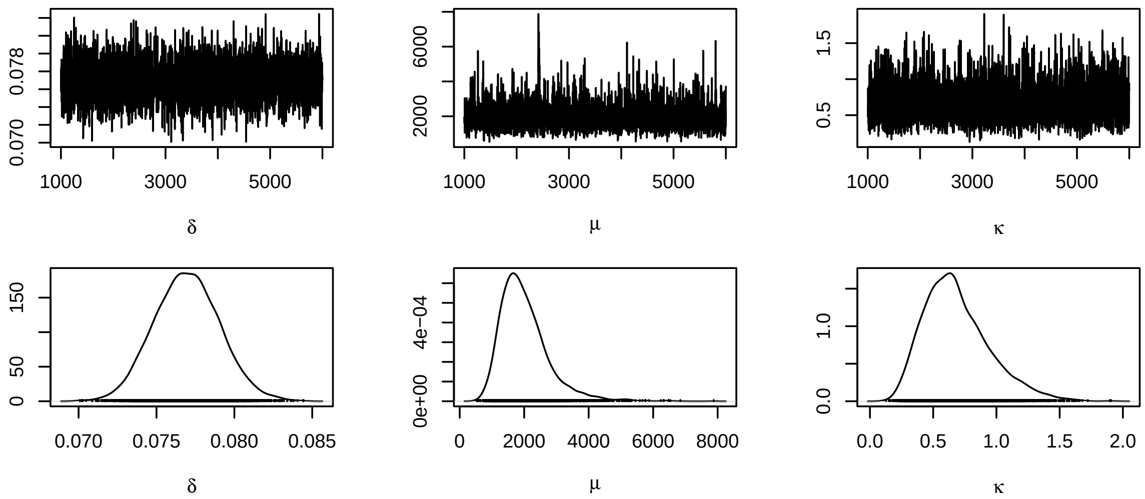


Figure 14.6: Top row: trace-plots of the parameters δ, μ and κ. Bottom row: empirical posterior densities of the parameters δ, μ and κ. (See R-Code 14.7.)

14.5 Bibliographic Remarks

There is ample literature about in-depth Bayesian methods and the computational implementation of these and we only give a few relevant links.

The list of textbooks discussing MCMC is long and extensive. Held and Sabanés Bové (2014) has some basic and accessible ideas. Accessible examples for actual implementations can be found in Kruschke (2015) (JAGS and STAN) and Kruschke (2010) (Bugs).

Further information about MCMC diagnostics is found in general Bayesian text books like Lunn *et al.* (2012). Specific and often used tests are published in Geweke (1992) Gelman and Rubin (1992), and Raftery and Lewis, 1992 and are implemented in the package *coda* with *geweke.plot()*, *gelman.diag()*, *raftery.diag()*.

An alternative to JAGS is BUGS (Bayesian inference Using Gibbs Sampling) which is distributed as two main versions: WinBUGS and OpenBUGS, see also Lunn *et al.* (2012). Additionally, there is the R-Interface package (*R2OpenBUGS*, Sturtz *et al.*, 2005). Other possibilities are the *Stan* or *INLA* engines with convenient user interfaces to R through *rstan* and *INLA* (Gelman *et al.*, 2015; Rue *et al.*, 2009; Lindgren and Rue, 2015).

14.6 Exercises and Problems

Problem 14.1 (Normal-normal-gamma model) Let $Y_1, Y_2, \dots, Y_n \mid \mu, \tau \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1/\kappa)$. Instead of independent priors on μ and κ, we propose a joint prior density that can be factorized by the density of κ and $\mu \mid \kappa$. We assume $\kappa \sim \mathcal{G}am(\alpha, \beta)$ and $\mu \mid \kappa \sim \mathcal{N}(\eta, 1/(\kappa\nu))$, for some hyper-parameters η, $\nu > 0$, $\alpha > 0$, and $\beta > 0$. This distribution is a so-called normal-gamma distribution, denoted by $\mathcal{N}\Gamma(\eta, \nu, \alpha, \beta)$.

a) Create an artificial dataset consisting for $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(1, 1)$, with $n = 20$.

b) Write a function called `dnormgamma()` that calculates the density at `mu`, `kappa` based on the parameters `eta`, `nu`, `alpha`, `beta`. Visualize the bivariate density based on $\eta = 1$, $\nu = 1.5$, $\alpha = 1$, and $\beta = 0.2$.

c) Setup a Gibbs sampler for the following values $\eta = 0$, $\nu = 1.5$, $\alpha = 1$, and $\beta = 0.2$. For a sample of length 2000 illustrate the (empirical) joint posterior density of $\mu, \kappa \mid y_1, \dots, y_n$.

Hint: Follow closely R-Code 13.6.

d) It can be shown that the posterior is again normal-gamma with parameters

$$\eta_{\text{post}} = \frac{1}{n+\nu}(n\bar{y} + \nu\eta) \qquad \nu_{\text{post}} = \nu + n \tag{14.21}$$

$$\alpha_{\text{post}} = \alpha + \frac{n}{2} \qquad \beta_{\text{post}} = \beta + \frac{1}{2}\Big((n-1)s^2 + \frac{n\nu(\eta - \bar{x})^2}{n+\nu}\Big) \tag{14.22}$$

where s^2 is the usual unbiased estimate of σ^2. Superimpose the true isolines of the normal-gamma prior and posterior density in the plot form the previous problem.

e) Compare the posterior mean of the normal-normal model with η_{post}.

Problem 14.2 (Monte Carlo integration) Estimate the volume of the unit ball in $d = 2, 3, \dots, 10$ dimensions and compare it to the exact value $\pi^{d/2}/\Gamma(d/2 + 1)$. What do you notice?

Problem 14.3 (Rejection sampling) A random variable X has a Laplace distribution with parameters $\mu \in \mathbb{R}$ and $\lambda > 0$ if its density is of the form

$$f_X(x) = \frac{1}{2\lambda}\exp\Big(-\frac{|x-\mu|}{\lambda}\Big).$$

a) Draw 100 realizations of a Laplace distribution with parameters $\mu = 1$ and $\lambda = 1$ with a rejection sampling approach.

b) Propose an intuitive alternative sampling approach based on `rexp()`.

Problem 14.4 (Anthelmintic model) The process of determining the parasitic load, a fecal sample is taken and is thoroughly mixed after dilution. We assume that the eggs are homogeneously distributed within each sample. A proportion of the diluted sample $p = 1/f$ is then counted. Denote the raw number of eggs in the diluted sample of the ith control animal as Y_i^{*C}, with $i = 1, 2, \dots, n$. Given the true number of eggs per gram of feces Y_i^C, the raw count Y_i^{*C} follows a binomial distribution $\mathcal{B}in(Y_i^C, p)$. This captures both the dilution and the counting variability. For the true epg counts Y_i^C we use the same models as in Example 14.8. Similar approach is used for the observations after the treatment.

a) Implement the model in JAGS and compare the results with those of the simple JAGS sampler of Example 14.8.

b) The package `eggCounts` provides samplers for the specific model discussed here as well as further extensions. Interpret the output of

```
model <- fecr_stanSimple(epgs2$before, epgs2$after)
```

and compare the result with those of **a**). (See also the vignette of the package *eggCounts*).

Epilogue

Are we done yet? No of course not!

An introduction is typically followed by an immersion. Here based on the lecture *STA121 Statistical Modeling.*

Throughout the book we have encountered several gaps that should be filled to feel proficient in statistics. To name a few: the extension from linear models to mixed models or to generalized linear models. In a similar fashion we can work with arbitrary many predictors and work with non-parametric regressions (e.g., guide the eye curves in the scatterplots) or neural nets. Dropping the iid Gaussian assumption and working with time series data, spatial data and extreme values. Real world examples are typically much larger and messier than what we have encountered here and thus methods to find low(er) dimensional structures in the data are often first steps in an analysis. Similarly, one often has to find clusters, i.e., grouping observations into similar groups, or construct classifiers, i.e., allocate new observations to known groups.

After all that, *are we done yet?* No of course not, the educational iteration requires a further refinement. Books at virtually arbitrary length could be written for all the topics above. However, I often feel that too many iterations is not advancing my capability to help statistically in a proportional manner. Being a good statistician does not only require having solid knowledge in statistics but also having domain specific knowledge, the skills to listen and to talk to experts and to have fun stepping outside the own comfort zone, rather into a foreign backyard (in the sense of John Tukey's quote "The best thing about being a statistician is that you get to play in everyone's backyard.").

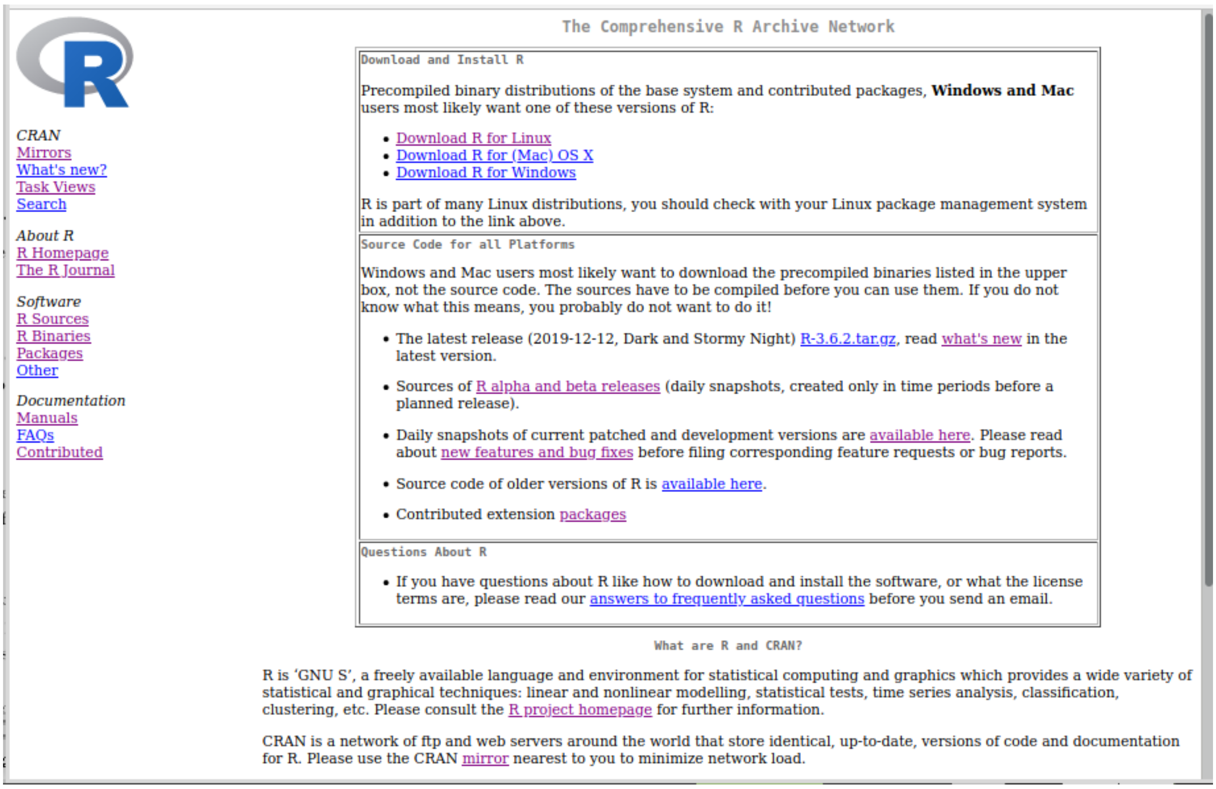
For the specific document, I am done here up to some standard appendices and indices.

Appendix A

Software Environment R

R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques. It compiles and runs on a wide varieties operating systems (Windows, Mac, and Linux), its central entry point is https://www.r-project.org.

The R software can be downloaded from CRAN (Comprehensive R Archive Network) https://cran.r-project.org, a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Figure A.1 shows a screenshot of the web page.



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-12-12, Dark and Stormy Night) R-3.6.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

What are R and CRAN?

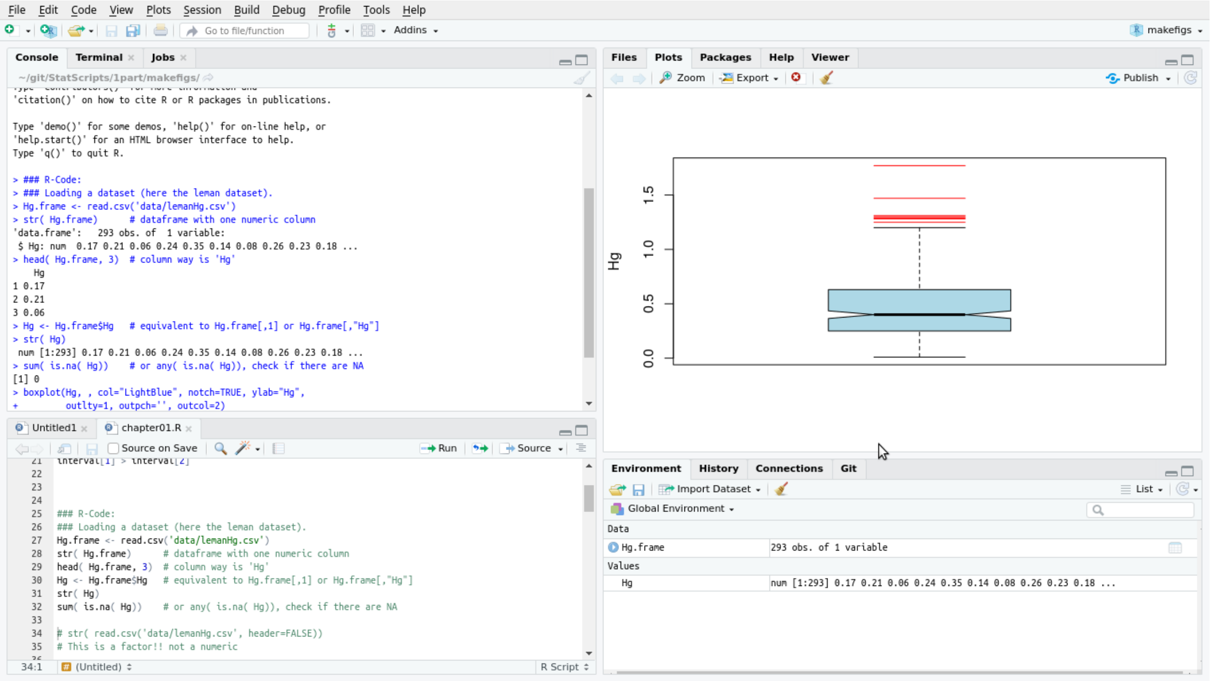
R is ‘GNU S’, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Figure A.1: Screenshot of the entry webpage of CRAN (Comprehensive R Archive Network, https://cran.r-project.org).

R is console based, that means that individual commands have to be *typed*. It is very important to save these commands to construct a reproducible workflow – which *the* big advantage over a "click-and-go" approach. We strongly recommend to use some graphical, integrated development environment (IDE) for R. The prime choice these days is RStudio. RStudio includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management, see Figure A.2.

RStudio is available in a desktop open source version for many different operating systems (Windows, Mac, and Linux) or in a browser connected to an RStudio Server. There are several providers of such servers, including https://rstudio.math.uzh.ch for the students of the STA120 lecture.



```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ### R-Code:
> ### Loading a dataset (here the leman dataset).
> Hg.frame <- read.csv('data/lemanHg.csv')
> str( Hg.frame)      # dataframe with one numeric column
'data.frame':   293 obs. of  1 variable:
 $ Hg: num  0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
> head( Hg.frame, 3)  # column way is 'Hg'
    Hg
1 0.17
2 0.21
3 0.06
> Hg <- Hg.frame$Hg   # equivalent to Hg.frame[,1] or Hg.frame[,"Hg"]
> str( Hg)
 num [1:293] 0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
> sum( is.na( Hg))    # or any( is.na( Hg)), check if there are NA
[1] 0
> boxplot(Hg, , col="LightBlue", notch=TRUE, ylab="Hg",
+         outlty=1, outpch='', outcol=2)
```

```
interval[1] > interval[2]



### R-Code:
### Loading a dataset (here the leman dataset).
Hg.frame <- read.csv('data/lemanHg.csv')
str( Hg.frame)      # dataframe with one numeric column
head( Hg.frame, 3)  # column way is 'Hg'
Hg <- Hg.frame$Hg   # equivalent to Hg.frame[,1] or Hg.frame[,"Hg"]
str( Hg)
sum( is.na( Hg))    # or any( is.na( Hg)), check if there are NA

# str( read.csv('data/lemanHg.csv', header=FALSE))
# This is a factor!! not a numeric
```

```
Data
Hg.frame        293 obs. of 1 variable
Values
  Hg            num [1:293] 0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
```

Figure A.2: RStudio screenshot. The four panels shown are (clock-wise starting top left): (i) console, (ii) plots, (iii) environment, (iv) script.

The installation of all software components are quite straightforward, but the look of the download page may change from time to time and the precise steps may vary a bit. Some examples are given by the attached videos. 4 min

The biggest advantage of using R is the support from and for a huge user community. Sheer endless packages provide almost seemingly every statistical task, often implemented by several authors. The packages are documented and by the upload to CRAN confined to a limited level of documentation, coding standards, (unit) testing etc. There are several forums (e.g., R mailing lists, Stack Overflow with tag "r") to get additional help, see https://www.r-project.org/help.html.

In this document we tried to keep the level of R quite low and rely on few packages only, examples were `MASS`, `vioplot`, `mvtnorm`, `ellipse` and some more. Due to complex dependencies, more than the actual loaded packages are used. The following R-Code output shows all packages (and their version number) used to compile this document (not including any packages required for the problems).

R-Code A.1: R-Session infomation of this document.

```
print( sessionInfo(), locale=FALSE)
## R Under development (unstable) (2023-01-31 r83741)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/R-devel/lib/R/lib/libRblas.so
## LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3;  LAPACK version 3.10.0
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] LearnBayes_2.15.1      BayesFactor_0.9.12-4.4 eggCounts_2.3-2
##  [4] Rcpp_1.0.10            arm_1.13-1             lme4_1.1-31
##  [7] Matrix_1.5-3           rjags_4-13             coda_0.19-4
## [10] TeachingDemos_2.12     daewr_1.2-7            pwr_1.3-0
## [13] car_3.1-1              carData_3.0-5          faraway_1.0.8
## [16] mvtnorm_1.1-3          ellipse_0.4.3          fields_14.1
## [19] viridis_0.6.2          viridisLite_0.4.1      spam_2.9-1
## [22] exactRankTests_0.8-35  coin_1.4-2             survival_3.5-0
## [25] palmerpenguins_0.1.1   vioplot_0.4.0          zoo_1.8-11
## [28] sm_2.2-5.7.1           MASS_7.3-58.2          knitr_1.42
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0     rootSolve_1.8.2.3    DoE.base_1.2-1
##  [4] libcoin_1.0-9        dplyr_1.1.0          loo_2.5.1
##  [7] combinat_0.0-8       TH.data_1.1-1        mathjaxr_1.6-0
## [10] numbers_0.8-5        digest_0.6.31        dotCall64_1.0-2
## [13] lifecycle_1.0.3      StanHeaders_2.21.0-7 processx_3.8.0
## [16] magrittr_2.0.3       compiler_4.3.0       rlang_1.0.6
## [19] tools_4.3.0          igraph_1.4.0         utf8_1.2.3
## [22] prettyunits_1.1.1    pkgbuild_1.4.0       scatterplot3d_0.3-42
## [25] multcomp_1.4-20      abind_1.4-5          partitions_1.10-7
## [28] grid_4.3.0           stats4_4.3.0         fansi_1.0.4
## [31] conf.design_2.0.0    inline_0.3.19        colorspace_2.1-0
## [34] ggplot2_3.4.0        scales_1.2.1         cli_3.6.0
## [37] crayon_1.5.2         generics_0.1.3       RcppParallel_5.1.6
## [40] FrF2_2.2-3           pbapply_1.7-0        minqa_1.2.5
## [43] polynom_1.4-1        stringr_1.5.0        rstan_2.21.8
## [46] modeltools_0.2-23    splines_4.3.0        maps_3.4.1
## [49] parallel_4.3.0       matrixStats_0.63.0   vctrs_0.5.2
## [52] boot_1.3-28.1        sandwich_3.0-2       callr_3.7.3
## [55] vcd_1.4-11           glue_1.6.2           nloptr_2.0.3
## [58] ps_1.7.2             codetools_0.2-18     stringi_1.7.12
## [61] gtable_0.3.1         sfsmisc_1.1-14       lmtest_0.9-40
## [64] gmp_0.7-1            munsell_0.5.0        tibble_3.1.8
## [67] pillar_1.8.1         R6_2.5.1             Rdpack_2.4
## [70] evaluate_0.16        lattice_0.20-45      highr_0.10
## [73] rbibutils_2.2.13     MatrixModels_0.5-1   rstantools_2.2.0
## [76] gridExtra_2.3        nlme_3.1-161         xfun_0.37
## [79] pkgconfig_2.0.3
```

Appendix B

Calculus

In this chapter we present some of the most important ideas and concepts of calculus. For example, we will not discuss sequences and series. It is impossible to give a formal, mathematically precise exposition. Further, we cannot present all rules, identities, guidelines or even tricks.

B.1 Functions

We start with one of the most basic concepts, a formal definition that describes a relation between two sets.

Definition B.1. A function f from a set D to a set W is a rule that assigns a unique value element $f(x) \in W$ to each element $x \in D$. We write

$$f : D \to W \tag{B.1}$$

$$x \mapsto f(x) \tag{B.2}$$

The set D is called the domain, the set W is called the range (or target set or codomain). The graph of a function f is the set $\big\{(x, f(x)) : x \in D\big\}$. ♢

The function will not necessarily map to every element in W, and there may be several elements in D with the same image in W. These functions are characterized as follows.

Definition B.2.

1. A function f is called injective, if the image of two different elements in D is different.
2. A function f is called surjective, if for every element y in W there is at least one element x in D such that $y = f(x)$.
3. A function f is called bijective if it is surjective and injective. Such a function is also called a *one-to-one* function. ♢

As an illustration, the first point can be 'translated' to $\forall x, z \in D, x \neq z \implies f(x) \neq f(z)$, which is equivalent to $\forall x, z \in D, f(x) = f(z) \implies x = z$.

By restricting the range, it is possible to render a function surjective. It is often possible to restrict the domain to obtain a locally bijective function.

In general, there is virtually no restriction on the domain and codomain. However, we often work with real functions, i.e., $D \subset \mathbb{R}$ and $W \subset \mathbb{R}$.

There are many different characterizations of functions. Some relevant one are as follows.

Definition B.3. A real function f is

1. periodic if there exists an $\omega > 0$ such that $f(x+\omega) = f(x)$ for all $x \in D$. The smallest value ω is called the period of f;
2. called increasing if $f(x) \leq f(x+h)$ for all $h \geq 0$. In case of strict inequalities, we call the function strictly increasing. Similar definitions hold when reversing the inequalities. ♢

The inverse $f^{-1}(y)$ of a bijective function $f : D \to W$ is defined as

$$\begin{aligned} f^{-1} : W &\to D \\ y &\mapsto f^{-1}(y), \text{ such that } y = f\big(f^{-1}(y)\big). \end{aligned} \tag{B.3}$$

Subsequently, we require the "inverse" of increasing functions by generalizing the previous definition. We call these function quantile functions.

To capture the behavior of a function locally, say at a point $x_0 \in D$, we use the concept of a *limit*.

Definition B.4. Let $f : D \to \mathbb{R}$ and $x_0 \in D$. The limit of f as x approaches x_0 is a, written as $\lim_{x\to x_0} f(x) = a$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D$ with $0 < |x - x_0| < \delta \implies |f(x) - a| < \epsilon$. ♢

The latter definition does not assume that the function is defined at x_0.

It is possible to define "directional" limits, in the sense that x approaches x_0 from above (from the right side) or from below (from the left side). These limits are denoted with

$$\lim_{x\to x_0^+} \quad \lim_{x\searrow x_0} \quad \text{for the former; or} \quad \lim_{x\to x_0^-} \quad \lim_{x\nearrow x_0} \quad \text{for the latter.} \tag{B.4}$$

We are used to interpret graphs and when we sketch an arbitrary function we often use a single, continuous line. This concept of not lifting the pen while sketching is formalized as follows and linked directly to limits, introduced above.

Definition B.5. A function f is continous in x_0 if the following limits exist

$$\lim_{h\nearrow 0} f(x_0+h) \qquad \lim_{h\searrow 0} f(x_0+h) \tag{B.5}$$

and are equal to $f(x_0)$. ♢

There are many other approaches to define coninuity, for example in terms of neighborhoods, in terms of limits of sequences.

Another very important (local) characterization of a function is the derivative, which quantifies the (infinitesimal) rate of change.

Definition B.6. The derivative of a function $f(x)$ with respect to the variable x at the point x_0 is defined by

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}, \tag{B.6}$$

provided the limit exists. We also write $\frac{df(x_0)}{dx} = f'(x_0)$.

If the derivative exists for all $x_0 \in D$, the function f is differentiable. ♢

Some of the most important properties in differential calculus are:

Property B.1. 1. *Differentability implies continuity.*

2. *(Mean value theorem) For a continuous function $f : [a, b] \to \mathbb{R}$, which is differentiable on (a, b) there exists a point $\xi \in (a, b)$ such that $f'(\xi) = \frac{f(b) - f(a)}{b - a}$.*

The *integral* of a (positive) function quantifies the area between the function and the x-axis. A mathematical definition is a bit more complicated.

Definition B.7. Let $f(x) : D \to \mathbb{R}$ a function and $[a, b] \in D$ a finite interval such that $|f(x)| < \infty$ for $x \in [a, b]$. For any n, let $t_0 = a < t_1 < \cdots < t_n = b$ a partition of $[a, b]$.

The integral of f from a to b is defined as

$$\int_a^b f(x)dx = \lim_{n \to \infty} \sum_{i=1}^{n} f(t_i)(t_i - t_{i-1}). \tag{B.7}$$

♢

For non-finite a and b, the definition of the integral can be extended via limits.

Property B.2. *(Fundamental theorem of calculus (I)). Let $f : [a, b] \to \mathbb{R}$ continuous. For all $x \in [a, b]$, let $F(x) = \int_a^x f(u)du$. Then F is continuous on $[a, b]$, differentiable on (a, b) and $F'(x) = f(x)$, for all $x \in (a, b)$.*

The function F is often called the antiderivative of f. There exists a second form of the previous theorem that does not assume continuity of f but only Riemann integrability, that means that an integral exists.

Property B.3. *(Fundamental theorem of calculus (II)). Let $f : [a, b] \to \mathbb{R}$. And let F such that $F'(x) = f(x)$, for all $x \in (a, b)$. If f is Riemann integrable then $\int_a^b f(u)du = F(b) - F(a)$.*

There are many 'rules' to calculate integrals. One of the most used ones is called integration by substitution and is as follows.

Property B.4. *Let I be an interval and $\varphi : [a, b] \to I$ be a differentiable function with integrable derivative. Let $f : I \to \mathbb{R}$ be a continuous function. Then*

$$\int_{\varphi(a)}^{\varphi(b)} f(u)\, du = \int_a^b f(\varphi(x))\varphi'(x)\, dx. \tag{B.8}$$

B.2 Functions in Higher Dimensions

We denote with $\mathbb{R}^m$ the vector space with elements $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$, called vectors, equipped with the standard operations. We will discuss *vectors* and *vector notation* in more details in the subsequent chapter.

A natural extension of a real function is as follows. The set D is subset of $\mathbb{R}^m$ and thus we write

$$\begin{aligned} f : D \subset \mathbb{R}^m &\to W \\ \boldsymbol{x} &\mapsto f(\boldsymbol{x}). \end{aligned} \tag{B.9}$$

Note that we keep $W \subset \mathbb{R}$.

The concept of limit and continuity translates one-to-one. Differentiability, however, is different and slightly more delicate.

Definition B.8. The partial derivative of $f : D \subset \mathbb{R} \to W$ with respect to x_j is defined by

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{j-1}, x_j + h, x_{j+1}, \ldots, x_m) - f(x_1, \ldots, x_n)}{h}, \tag{B.10}$$

(provided it exists). ♢

The derivative of f with respect to all components is thus a vector

$$\boldsymbol{f}'(\boldsymbol{x}) = \Big(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_m}\Big)^\top \tag{B.11}$$

Hence $\boldsymbol{f}'(\boldsymbol{x})$ is a vector valued function from D to $\mathbb{R}^m$ and is called the gradient of f at $\boldsymbol{x}$, also denoted with $\mathrm{grad}(f(x)) = \nabla f(x)$.

Remark B.1. The existence of partial derivatives is not sufficient for the differentiability of the function f. ♣

In a similar fashion, higher order derivatives can be calculated. For example, taking the derivative of each component of (B.11) with respect to all components is an matrix with components

$$\boldsymbol{f}''(\boldsymbol{x}) = \Big(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\Big), \tag{B.12}$$

called the Hessian matrix.

It is important to realize that the second derivative constitutes a set of derivatives of f: all possible double derivatives.

B.3 Approximating Functions

Quite often, we want to approximate functions.

Property B.5. *Let $f : D \to \mathbb{R}$ with continuous Then there exists $\xi \in [a, x]$ such that*

$$\begin{aligned} f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \dots \\ + \frac{1}{m!}f^{(m)}(a)(x-a)^m + \frac{1}{(m+1)!}f^{(m+1)}(\xi)(x-a)^m \end{aligned} \tag{B.13}$$

We call (B.13) Taylor's formula and the last term, often denoted by $R_n(x)$, as the reminder of order n. Taylor's formula is an extension of the mean value theorem.

If the function has bounded derivatives, the reminder $R_n(x)$ converges to zero as $x \to a$.

Hence, if the function is at least twice differentiable in a neighborhood of a then

$$f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 \tag{B.14}$$

is the best quadratic approximation in this neighborhood.

If all derivatives of f exist in an open interval I with $a \in I$, we have for all $x \in I$

$$f(x) = \sum_{r=0}^{\infty} \frac{1}{r!} f^{(r)}(a)(x-a)^r \tag{B.15}$$

Often the approximation is for $x = a + h$, h small.

Taylor's formula can be expressed for multivariate real functions. Without stating the precise assumptions we consider here the following example

$$f(\boldsymbol{a} + \boldsymbol{h}) = \sum_{r=0}^{\infty} \sum_{\boldsymbol{i}:i_1+\dots+i_n=r} \frac{1}{i_1!i_2!\dots i_n!} \frac{\partial^r f(\boldsymbol{a})}{\partial x_{i_1}\dots\partial x_{i_n}} h_1^{i_1} h_2^{i_2} \dots h_n^{i_n}, \tag{B.16}$$

extending (B.15) with $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{h}$.

Appendix C

Linear Algebra

In this chapter we cover the most important aspects of linear algebra, namely of notational nature.

C.1 Vectors, Matrices and Operations

A collection of p real numbers is called a vector, an array of $n \times m$ real numbers is called a matrix. We write

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \qquad \mathbf{A} = (a_{ij}) = \begin{pmatrix} a_{11} & \dots & \boldsymbol{a}_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix}. \tag{C.1}$$

Providing the dimensions are coherent, vector and matrix addition (and subtraction) is performed componentwise, as is scalar multiplication. That means, for example, that $\boldsymbol{x} \pm \boldsymbol{y}$ is a vector with elements $x_i \pm y_i$ and $c\mathbf{A}$ is a matrix with elements ca_{ij}.

The $n \times n$ identity matrix $\mathbf{I}$ is defined as the matrix with ones on the diagonal and zeros elsewhere. We denote the vector with solely one elements with $\mathbf{1}$ similarly, $\mathbf{0}$ is a vector with only zero elements. A matrix with entries $d_1, \dots, d_n$ on the diagonal and zero elsewhere is denoted with $\operatorname{diag}(d_1, \dots, d_n)$ or $\operatorname{diag}(d_i)$ for short and called a diagonal matrix. Hence, $\mathbf{I} = \operatorname{diag}(\mathbf{1})$.

To indicate the ith-jth element of $\mathbf{A}$, we use $(\mathbf{A})_{ij}$. The transpose of a vector or a matrix flips its dimension. When a matrix is transposed, i.e., when all rows of the matrix are turned into columns (and vice-versa), the elements a_{ij} and a_{ji} are exchanged. Thus $(\mathbf{A}^\top)_{ij} = (\mathbf{A})_{ji}$. The vector $\boldsymbol{x}^\top = (x_a, \dots, x_p)$ is termed a row vector. We work mainly with column vectors as shown in (C.1).

In the classical setting of real numbers, there is only one type of multiplication. As soon as we have several dimensions, several different types of multiplications exist, notably scalar multiplication, matrix multiplication and inner product (and actually more such as the vector product, outer product).

Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times p$ and $p \times m$ matrices. Matrix multiplication $\mathbf{AB}$ is defined as

$$\mathbf{AB} = \mathbf{C} \qquad \text{with} \qquad (\mathbf{C})_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}. \tag{C.2}$$

This last equation shows that the matrix $\mathbf{I}$ is the neutral element (or identity element) of the matrix multiplication.

Definition C.1. The inner product between two p-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as $\boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^p x_i y_i$. There are several different notations used: $\boldsymbol{x}^\top \boldsymbol{y} = \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{x} \cdot \boldsymbol{y}$.

If for an $n \times n$ matrix $\mathbf{A}$ there exists an $n \times n$ matrix $\mathbf{B}$ such that

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}, \tag{C.3}$$

then the matrix $\mathbf{B}$ is uniquely determined by $\mathbf{A}$ and is called the inverse of $\mathbf{A}$, denoted by $\mathbf{A}^{-1}$.

C.2 Linear Spaces and Basis

The following definition formalizes one of the main spaces we work in.

Definition C.2. A vector space over $\mathbb{R}$ is a set V with the following two operations:

1. $+ : V \times V \to V$ (vector addition)
2. $\cdot : \mathbb{R} \times V \to V$ (scalar multiplication). ♢

Typically, V is $\mathbb{R}^p$, $p \in \mathbb{N}$.

In the following we assume a fixed d and the usual operations on the vectors.

Definition C.3. 1. The vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are linearly dependent if there exists scalars $a_1, \dots, a_k$ (not all equal to zero), such that $a_1 \boldsymbol{v}_1 + \cdots + a_k \boldsymbol{v}_k = \mathbf{0}$.

2. The vectors $\boldsymbol{v}_1, \dots \boldsymbol{v}_k$ are linearly independent if $a_1 \boldsymbol{v}_1 + \cdots + a_k \boldsymbol{v}_k = \mathbf{0}$ cannot be satisfied by any scalars $a_1, \dots, a_k$ (not all equal to zero). ♢

In a set of linearly dependent vectors, each vector can be expressed as a linear combination of the others.

Definition C.4. The set of vectors $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_d\}$ is a basis of a vectors space V if the set is linearly independent and any other vector $\boldsymbol{v} \in V$ can be expressed by $\boldsymbol{v} = v_1 \boldsymbol{b}_1 + \cdots + v_d \boldsymbol{b}_d$. ♢

The following proposition summarizes some of the relevant properties of a basis.

Property C.1. *1. The decomposition of a vector $\boldsymbol{v} \in V$ in $\boldsymbol{v} = v_1 \boldsymbol{b}_1 + \cdots + v_d \boldsymbol{b}_d$ is unique.*

2. All basis of V have the same cardinality, which is called the dimension of V, $\dim(V)$*.*

3. If there are two basis $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_d\}$ and $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_d\}$ then there exists a $d \times d$ matrix $\mathbf{A}$ such that $\boldsymbol{e}_i = \mathbf{A} \boldsymbol{b}_i$, for all i.

Definition C.5. The standard basis, or canonical basis of $V = \mathbb{R}^d$ is $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_d\}$ with $\boldsymbol{e}_i = (0, \dots, 0, 1, 0, \dots)^\top$, i.e., the vector with a one at the ith position and zero elsewhere. ♢

Definition C.6. Let $\mathbf{A}$ be a $n \times m$ matrix. The column rank of the matrix is the dimension of the subspace that the m columns of $\mathbf{A}$ span and is denoted by $\operatorname{rank}(\mathbf{A})$. A matrix is said to have full rank if $\operatorname{rank}(A) = m$.

The row rank is the column rank of $\mathbf{A}^\top$. ♢

Some fundamental properties of the rank are as follows.

Property C.2. *Let* $\mathbf{A}$ *be a* $n \times m$ *matrix.*

1. *The column rank and row rank are identical.*
2. $\operatorname{rank}(\mathbf{A}^\top\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}^\top) = \operatorname{rank}(\mathbf{A})$.
3. $\operatorname{rank}(\mathbf{A}) \le \dim(V)$.
4. $\operatorname{rank}(\mathbf{A}) \le \min(m, n)$.
5. *For an appropriately sized matrix* $\mathbf{B}$ $\operatorname{rank}(\mathbf{A}+\mathbf{B}) \le \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B})$ *and* $\operatorname{rank}(\mathbf{AB}) \le \min\bigl(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\bigr)$.

C.3 Projections

We consider classical Euclidean vector spaces with elements $\boldsymbol{x} = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$ with Euclidean norm $\|\boldsymbol{x}\| = (\sum_i x_i^2)^{1/2}$.

To illustrate projections, consider the setup illustrated in Figure C.1, where $\boldsymbol{y}$ and $\boldsymbol{a}$ are two vectors in $\mathbb{R}^2$. The subspace spanned by $\boldsymbol{a}$ is

$$\{\lambda \boldsymbol{a}, \lambda \in \mathbb{R}\} = \{\lambda \boldsymbol{a}/\|\boldsymbol{a}\|, \lambda \in \mathbb{R}\} \tag{C.4}$$

where the second expression is based on a normalized vector $\boldsymbol{a}/\|\boldsymbol{a}\|$. By the (geometric) definition of the inner product (dot product),

$$< \boldsymbol{a}, \boldsymbol{b} > = \boldsymbol{a}^\top \boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos\theta \tag{C.5}$$

where θ is the angle between the vectors. Classical trigonometric properties state that the length of the projection is $\boldsymbol{a}/\|\boldsymbol{a}\| \cdot \|\boldsymbol{y}\| \cos(\theta)$. Hence, the projected vector is

$$\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \frac{\boldsymbol{a}^\top}{\|\boldsymbol{a}\|} \boldsymbol{y} = \boldsymbol{a}(\boldsymbol{a}^\top \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{y}. \tag{C.6}$$

In statistics we often encounter expressions like this last term. For example, ordinary least squares ("classical" multiple regression) is a projection of the vector $\boldsymbol{y}$ onto the column space spanned by $\mathbf{X}$, i.e., the space spanned by the columns of the matrix $\mathbf{X}$. The projection is $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$. Usually, the column space is in a lower dimension.

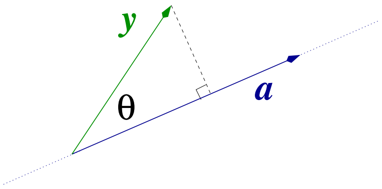


Figure C.1: Projection of the vector $\boldsymbol{y}$ onto the subspace spanned by $\boldsymbol{a}$.

Remark C.1. Projection matrices (like $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$) have many nice properties such as being symmetric, being idempotent, i.e., $\mathbf{H} = \mathbf{H}\mathbf{H}$, having eigenvalues within $[0, 1]$, (see next section), $\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X})$, etc. ♣

C.4 Matrix Decompositions

In this section we elaborate representations of a matrix as a product of two or three other matrices.

Let $\boldsymbol{x}$ be a non-zero n-vector (i.e., at least one element is not zero) and $\mathbf{A}$ an $n \times n$ matrix. We can interpret $\mathbf{A}(\boldsymbol{x})$ as a function that maps $\boldsymbol{x}$ to $\mathbf{A}\boldsymbol{x}$. We are interested in vectors that change by a scalar factor by such a mapping

$$\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}, \tag{C.7}$$

where λ is called an eigenvalue and $\boldsymbol{x}$ an eigenvector.

A matrix has n eigenvalues, $\{\lambda_1, \dots, \lambda_n\}$, albeit not necessarily different and not necessarily real. The set of eigenvalues and the associated eigenvectors denotes an eigendecomposition.

For all square matrices, the set of eigenvectors span an orthogonal basis, i.e., are constructed that way.

We often denote the set of eigenvectors with $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n$. Let $\boldsymbol{\Gamma}$ be the matrix with columns $\boldsymbol{\gamma}_i$, i.e., $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n)$. Then

$$\boldsymbol{\Gamma}^\top\mathbf{A}\boldsymbol{\Gamma} = \text{diag}(\lambda_1, \dots, \lambda_n), \tag{C.8}$$

due to the orthogonality property of the eigenvectors $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \mathbf{I}$. This last identity also implies that $\mathbf{A} = \boldsymbol{\Gamma}\,\text{diag}(\lambda_1, \dots, \lambda_n)\boldsymbol{\Gamma}^\top$.

In cases of non-square matrices, an eigendecomposition is not possible and a more general approach is required. The so-called singular value decomposition (SVD) works or any $n \times m$ matrix $\mathbf{B}$,

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \tag{C.9}$$

where $\mathbf{U}$ is an $n \times \min(n, m)$ orthogonal matrix (i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$), $\mathbf{D}$ is an diagonal matrix containing the so-called singular values and $\mathbf{V}$ is an $\min(n, m) \times m$ orthogonal matrix (i.e., $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_m$).

We say that the columns of $\mathbf{U}$ and $\mathbf{V}$ are the left-singular vectors and right-singular vectors, respectively.

Note however, that the dimensions of the corresponding matrices differ in the literature, some write $\mathbf{U}$ and $\mathbf{V}$ as square matrices and $\mathbf{V}$ as a rectangular matrix.

Remark C.2. Given an SVD of $\mathbf{B}$, the following two relations hold:

$$\mathbf{B}\mathbf{B}^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}\mathbf{U}^\top = \mathbf{U}\mathbf{D}\mathbf{D}\mathbf{U}^\top \tag{C.10}$$

$$\mathbf{B}^\top\mathbf{B} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{V}\mathbf{D}\mathbf{D}\mathbf{V}^\top \tag{C.11}$$

and hence the columns of $\mathbf{U}$ and $\mathbf{V}$ are eigenvectors of $\mathbf{B}\mathbf{B}^\top$ and $\mathbf{B}^\top\mathbf{B}$, respectively, and most importantly, the elements of $\mathbf{D}$ are the square roots of the (non-zero) eigenvalues of $\mathbf{B}\mathbf{B}^\top$ or $\mathbf{B}^\top\mathbf{B}$. ♣

Besides an SVD there are many other matrix factorization. We often use the so-called Cholesky factorization, as - to a certain degree - it generalizes the concept of a square root for matrices. Assume that all eigenvalues of $\mathbf{A}$ are strictly positive, then there exists a unique lower triangular matrix $\mathbf{L}$ with positive entries on the diagonal such that $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$. There exist very efficient algorithm to calculate $\mathbf{L}$ and solving large linear systems is often based on a Cholesky factorization.

The determinant of a square matrix essentially describes the change in "volume" that associated linear transformation induces. The formal definition is quite complex but it can be written as $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$ for matrices with real eigenvalues.

The trace of a matrix is the sum of its diagonal entries.

C.5 Positive Definite Matrices

Besides matrices containing covariates, we often work with variance-covariance matrices, which represent an important class of matrices as we see now.

Definition C.7. A $n \times n$ matrix $\mathbf{A}$ is positive definite (pd) if

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} > 0, \qquad \text{for all } \boldsymbol{x} \neq \mathbf{0}. \tag{C.12}$$

Further, if $\mathbf{A} = \mathbf{A}^\top$, the matrix is symmetric positive definite (spd). ♢

Relevant properties of spd matrices $\mathbf{A} = (a_{ij})$ are given as follows.

Property C.3. *1.* $\operatorname{rank}(\mathbf{A}) = n$

2. the determinant is positive, $\det(\mathbf{A}) > 0$

3. all eigenvalues are positive, $\lambda_i > 0$

4. all elements on the diagonal are positive, $a_{ii} > 0$

5. $a_{ii}a_{jj} - a_{ij}^2 > 0$*,* $i \neq j$

6. $a_{ii} + a_{jj} - 2|a_{ij}| > 0$*,* $i \neq j$

7. $\mathbf{A}^{-1}$ *is spd*

8. all principal sub-matrices of $\mathbf{A}$ *are spd.*

For a non-singular matrix $\mathbf{A}$, written as a 2×2 block matrix (with square matrices $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$), we have

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C} \\ -\mathbf{C}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{C} \end{pmatrix} \tag{C.13}$$

with $\mathbf{C} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$. Note that $\mathbf{A}_{11}$ and $\mathbf{C}$ need to be invertible.

It also holds that $\det(\mathbf{A}) = \det(\mathbf{A}_{11})\det(\mathbf{A}_{22})$.

Bibliography

Abraham, B. and Ledolter, J. (2006). *Introduction To Regression Modeling.* Duxbury applied series. Thomson Brooks/Cole.

Agresti, A. (2002). *Categorical data analysis.* A Wiley-Interscience publication. Wiley, New York, second edition.

Agresti, A. (2007). *An introduction to categorical data analysis.* Wiley, New York, second edition.

Ahrens, J. H. and Dieter, U. (1972). Computer methods for sampling from the exponential and normal distributions. *Communications of the ACM*, **15**, 873–882.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *2nd International Symposium on Information Theory*, 267–281. Akadémiai Kiadó.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley & Sons Inc., Chichester.

Bland, J. M. and Bland, D. G. (1994). Statistics notes: One and two sided tests of significance. *BMJ*, **309**, 248.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-building and Response Surfaces.* Wiley.

Brown, L. D., Cai, T. T., and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30**, 160–201.

Canal, L. (2005). A normal approximation for the chi-square distribution. *Computational Statistics & Data Analysis*, **48**, 803 – 808.

Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, New Jersey, U.S.A.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Routledge.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, **84**, 945–957.

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences.* Brooks/Cole, 8th edition.

Edgington, E. and Onghena, P. (2007). *Randomization Tests.* CRC Press, 4th edition.

Ellis, T. H. N., Hofer, J. M. I., Swain, M. T., and van Dijk, P. J. (2019). Mendel's pea crosses: varieties, traits and statistics. *Hereditas*, **156**, 33.

Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen.* Springer, 2 edition.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications.* Springer.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* CRC Press.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347–388.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications: Volume I.* Number Bd. 1 in Wiley series in probability and mathematical statistics. John Wiley & Sons.

Fisher, R. A. (1938). Presidential address. *Sankhyā: The Indian Journal of Statistics*, **4**, 14–17.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **C-23**, 881–890.

Friendly, M. and Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Web document, www.math.yorku.ca/SCS/Gallery/milestone/. Accessed: September 16, 2014.

Furrer, R. and Genton, M. G. (1999). Robust spatial data analysis of lake Geneva sediments with S+SpatialStats. *Systems Research and Information Science*, **8**, 257–272.

Galarraga, V. and Boffetta, P. (2016). Coffee drinking and risk of lung cancer–a meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 951–957.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavior Science*, **40**, 530–543.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gemeinde Staldenried (1994). Verwaltungsrechnung 1993 Voranschlag 1994.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, J. F. M., editors, *Bayesian Statistics 4*, 169–193. Oxford University Press, Oxford.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, **31**, 337–350.

Grinstead, C. M. and Snell, J. L. (2003). *Introduction to Probability.* AMS.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions.* Wiley New York.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes.* Springer, Heidelberg.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference.* Springer.

Hernán, M. A., Alonso, A., and Logroscino, G. (2008). Cigarette smoking and dementia: potential selection bias in the elderly. *Epidemiology*, **19**, 448–450.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods.* John Wiley & Sons.

Hombach, M., Ochoa, C., Maurer, F. P., Pfiffner, T., Böttger, E. C., and Furrer, R. (2016). Relative contribution of biological variation and technical variables to zone diameter variations of disc diffusion susceptibility testing. *Journal of Antimicrobial Chemotherapy*, **71**, 141–151.

Horst, A. M., Hill, A. P., and Gorman, K. B. (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data.* R package version 0.1.0.

Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons Inc.

Hüsler, J. and Zimmermann, H. (2010). *Statistische Prinzipien für medizinische Projekte.* Huber, 5 edition.

Jeffreys, H. (1983). *Theory of probability.* The Clarendon Press Oxford University Press, third edition.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions.* Wiley-Interscience, 3rd edition.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol. 1.* Wiley-Interscience, 2nd edition.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol. 2.* Wiley-Interscience, 2nd edition.

Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS.* Academic Press, first edition.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan.* Academic Press/Elsevier, second edition.

Kupper, T., De Alencastro, L., Gatsigazi, R., Furrer, R., Grandjean, D., and J., T. (2008). Concentrations and specific loads of brominated flame retardants in sewage sludge. *Chemosphere*, **71**, 1173–1180.

Landesman, R., Aguero, O., Wilson, K., LaRussa, R., Campbell, W., and Penaloza, O. (1965). The prophylactic use of chlorthalidone, a sulfonamide diuretic, in pregnancy. *J. Obstet. Gynaecol.*, **72**, 1004–1010.

LeBauer, D. S., Dietze, M. C., and Bolker, B. M. (2013). Translating Probability Density Functions: From R to BUGS and Back Again. *The R Journal*, **5**, 207–209.

Leemis, L. M. and McQueston, J. T. (2008). Univariate distribution relationships. *The American Statistician*, **62**, 45–53.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, second edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, 3 edition.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, **63**, i19.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Texts in Statistical Science. Chapman & Hall/CRC.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, **32**, 12–16.

Modigliani, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Social Research*, **33**, 160–217.

Moyé, L. A. and Tita, A. T. (2002). Defending the rationale for the two-tailed test in clinical research. *Circulation*, **105**, 3062–3065.

Needham, T. (1993). A visual explanation of jensen's inequality. *American Mathematical Monthly*, **100**, 768–771.

Olea, R. A. (1991). *Geostatistical Glossary and Multilingual Dictionary*. Oxford University Press.

Pachel, C. and Neilson, J. (2010). Comparison of feline water consumption between still and flowing water sources: A pilot study. *Journal of Veterinary Behavior*, **5**, 130–133.

Petersen, K. B. and Pedersen, M. S. (2008). The Matrix Cookbook. Version 2008-11-14, http://matrixcookbook.com.

Plagellat, C., Kupper, T., Furrer, R., de Alencastro, L. F., Grandjean, D., and Tarradellas, J. (2006). Concentrations and specific loads of UV filters in sewage sludge originating from a monitoring network in Switzerland. *Chemosphere*, **62**, 915–925.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna, Austria.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**, 493–497.

Redelmeier, D. A. and Singh, S. M. (2001). Survival in academy award–winning actors and actresses. *Annals of Internal Medicine*, **134**, 955–962.

Rice, J. A. (2006). *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition.

Rosenthal, R. and Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, **8**, 183–189.

Ross, S. M. (2010). *A First Course in Probability.* Pearson Prentice Hall, 8th edition.

Ruchti, S., Kratzer, G., Furrer, R., Hartnack, S., Würbel, H., and Gebhardt-Henrich, S. G. (2019). Progression and risk factors of pododermatitis in part-time group housed rabbit does in switzerland. *Preventive Veterinary Medicine*, **166**, 56–64.

Ruchti, S., Meier, A. R., Würbel, H., Kratzer, G., Gebhardt-Henrich, S. G., and Hartnack, S. (2018). Pododermatitis in group housed rabbit does in switzerland prevalence, severity and risk factors. *Preventive Veterinary Medicine*, **158**, 114–121.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, **71**, 319–392.

Siegel, S. and Castellan Jr, N. J. (1988). *Nonparametric Statistics for The Behavioral Sciences.* McGraw-Hill, 2nd edition.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, **28**, 9–12.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677–680.

Sturtz, S., Ligges, U., and Gelman, A. (2005). `R2WinBUGS`: A package for running `WinBUGS` from `R`. *Journal of Statistical Software*, **12**, 1–16.

Swayne, D. F., Temple Lang, D., Buja, A., and Cook, D. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, **43**, 423–444.

SWISS Magazine (2011). Key environmental figures, 10/2011–01/2012, p. 107. http://www.swiss.com/web/EN/fly_swiss/on_board/Pages/swiss_magazine.aspx.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Graphics Press.

Tufte, E. R. (1990). *Envisioning Information.* Graphics Press.

Tufte, E. R. (1997a). *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*. Graphics Press.

Tufte, E. R. (1997b). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wang, C., Torgerson, P. R., Höglund, J., and Furrer, R. (2017). Zero-inflated hierarchical models for faecal egg counts to assess anthelmintic efficacy. *Veterinary Parasitology*, **235**, 20–28.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, **70**, 129–133.

Glossary

Throughout the document we tried to be consistent with standard mathematical notation. We write random variables as uppercase letters (X, Y, $\dots$), realizations as lower case letters (x, y, $\dots$), matrices as bold uppercase letters ($\boldsymbol{\Sigma}$, $\mathbf{X}$, $\dots$), and vectors as bold italics lowercase letters ($\boldsymbol{x}$, $\boldsymbol{\beta}$, $\dots$). (The only slight confusion arises with random vectors and matrices.)

The following glossary contains a non-exhaustive list of the most important notation. Standard operators or products are not repeatedly explained.

$:=$ Define the left hand side by the expression on the other side.

$\clubsuit$, $\diamondsuit$ $\square$ End of example, end of definition end of remark.

$\int$, $\sum$, $\prod$ Integration, summation and product symbol. If there is no ambiguity, we omit the domain in inline formulas.

$[a, b]$, $[c, d[$ Closed interval $\{x \in \mathbb{R} \mid a \leq x \leq b\}$, and half open interval $\{x \in \mathbb{R} \mid c \leq x < d\}$.

$\cup$, $\cap$ Union, intersection of sets or events.

$\varnothing$ Empty set.

A^c Complement of the set A.

$B \backslash A$ Relative complement of A in B.
All elements of the set B that are not in the set A: $\{x \in B \mid x \notin A\}$.

$\widehat{\theta}$ Estimator or estimate of the parameter θ.

$\bar{x}$ Sample mean: $\sum_{i=1}^{n} x_i/n$.

$|x|$ Absolute value of the scalar x.

$\|\boldsymbol{x}\|$ Norm of the vector $\boldsymbol{x}$.

$\mathbf{X}^\top$ Transpose of an matrix $\mathbf{X}$.

$x_{(i)}$ Order statistics of the sample $x_1, \dots, x_n$.

$\mathbf{0}$, $\mathbf{1}$ Vector or matrix with components 0 respectively 1.

$\mathrm{Cov}(X, Y)$ Covariance between two random variables X and Y.

$\mathrm{Corr}(X, Y)$ Correlation between two random variables X and Y.

$\frac{d}{dx}$, $'$, $\frac{\partial}{\partial x}$ Derivative and partial derivative with respect to x.

$\mathrm{diag}(\mathbf{A})$ Diagonal entries of an $(n \times n)$-matrix $\mathbf{A}$.

ε, ε_i Random variable or process, usually measurement error.

$\mathrm{E}(X)$ Expectation of the random variable X.

e, $\exp(\cdot)$	Transcendental number $\mathrm{e} = 2.71828\,18284$, the exponential function.
$n!$	Factorial of a positive integer n: $n! = n(n-1)(n-2)\cdots 1$, with $0! = 1$.
$\binom{n}{k}$	Binomial coefficient defined as $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.
$\mathbf{I}_n = \mathbf{I}$	Identity matrix, $\mathbf{I} = (\delta_{ij})$.
$I_{\{A\}}$	Indicator function, talking the value one if A is true and zero otherwise.
$\lim_{x\to a}$, $\lim_{x\nearrow a}$, $\lim_{x\searrow a}$	Limits as x approaches a (two sided), one sided limits where x approaches a from the left and from the right.
$\log(\cdot)$	Logarithmic function to the base e.
$\max\{A\}$, $\min\{A\}$	Maximum, minimum of the set A.
$\mathrm{med}(x_i)$	Median of the sample $x_1, \dots, x_n$: $\begin{cases} x_{(n/2+1/2)}, & \text{if } n \text{ odd,} \\ \frac{1}{2}(x_{(n/2} + x_{(n/2+1)}), & \text{if } n \text{ odd,} \end{cases}$
$\mathbb{N}$, $\mathbb{N}^d$	Space of natural numbers, of d-vectors with natural elements.
$\varphi(x)$	Gaussian probability densitiy function $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.
$\Phi(x)$	Gaussian cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} \varphi(z)\,\mathsf{d}z$.
π	Transzendental number $\pi = 3.14159\,26535$.
$\mathrm{P}(A)$	Probability of the event A.
$\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$	Space of real numbers, real n-vectors and real $(n \times m)$-matrices.
$\mathrm{rank}(\mathbf{A})$	The rank of a matrix $\mathbf{A}$ is defined as the number of linearly independent rows (or columns) of $\mathbf{A}$.
s	Sample standard deviation: $s = \sqrt{s^2}$.
s^2	Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
$\mathrm{tr}(\mathbf{A})$	Trace of an matrix $\mathbf{A}$ defined by the sum of its diagonal elements.
$\mathrm{Var}(X)$	Variance of the random variable X.
$\mathbb{Z}$, $\mathbb{Z}^d$	Space of integers, of d-vectors with integer elements.

The following table contains the abbreviations of the statistical distributions (dof denotes degrees of freedom).

$\mathcal{N}(0,1)$, z_p	Standard standard normal distribution, p-quantile thereof.
$\mathcal{N}(\mu, \sigma^2)$	Gaussian or normal distribution with parameters μ and σ^2 (being mean and variance), $\sigma^2 > 0$.
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal p dimensional distribution with mean vector $\boldsymbol{\mu}$ and symmetric positive definite (co-)variance matrix $\boldsymbol{\Sigma}$.
$\mathcal{B}in(n,p)$, $b_{n,p,1-\alpha}$	Binomial distribution with n trials and success probability p, $1-\alpha$-quantile thereof, $0 < p < 1$.
$\mathcal{P}ois(\lambda)$	Poisson distribution with parameter λ, $\lambda > 0$.
$\mathcal{E}xp(\lambda)$	Exponential distribution with rate parameter λ, $\lambda > 0$.

$\mathcal{U}(a,b)$	Uniform distribution over the support $[a,\, b]$, $-\infty < a < b < \infty$.
$\mathcal{X}^2_\nu$, $\chi^2_{\nu,p}$	Chi-squared distribution with ν dof, p-quantile thereof.
T_n, $t_{n,p}$	Student's t-distribution with n dof, p-quantile thereof.
$F_{m,n}$, $f_{m,n,p}$	F-distribution with m and n dof, p-quantile thereof.
$U_{\mathrm{crit}}(n_x,n_y;1-\alpha)$	$1-\alpha$-quantile of the distribution of the Wilcoxon rank sum statistic.
$W_{\mathrm{crit}}(n_\star;1-\alpha)$	$1-\alpha$-quantile of the distribution of the Wilcoxon signed rank statistic.

The following table contains the abbreviations of the statistical methods, properties and quality measures.

EDA	Exploratory data analysis.
DoE	Design of experiment.
DF, dof	Degrees of freedom.
MAD	Median absolute deviation.
MAE	Mean absolute error.
ML	Maximum likelihood (ML estimator or ML estimation).
MM	Method of moments.
MSE	Mean squared error.
OLS, LS	Ordinary least squares.
RMSE	Root mean squared error.
SS	Sums of squares.
WLS	Weighted least squares.

Index of Statistical Tests and Confidence Intervals

Video Index

The following index gives a short description of the available videos, including a link to the referenced page. The videos are uploaded to https://tube.switch.ch/.

Index of Terms