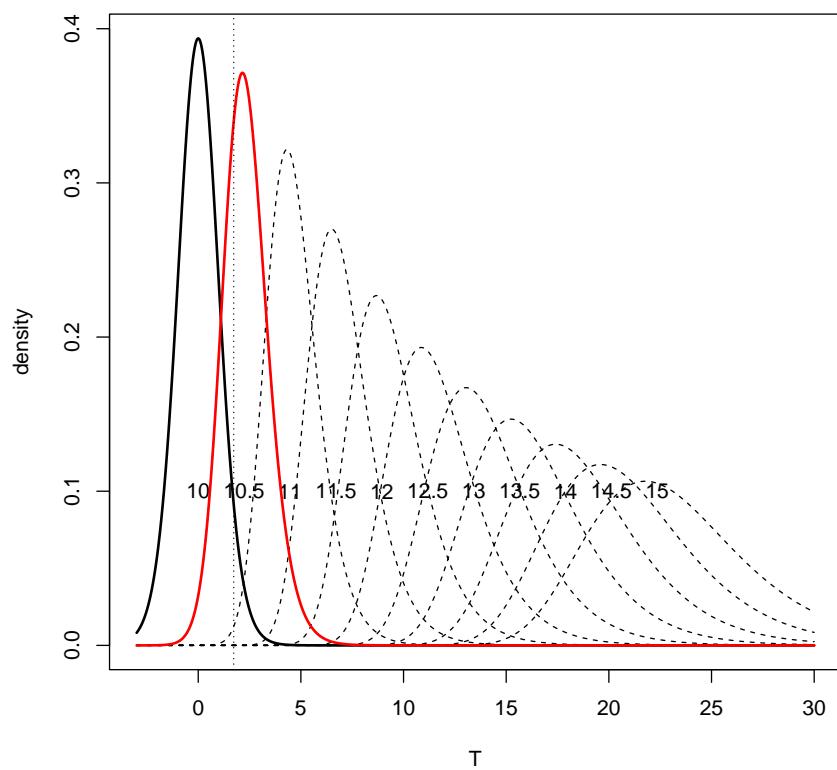


Quantitative Methoden

Wahrscheinlichkeitstheorie, Grundlagen Statistik und Lineare Modelle

André Meichtry

2025



Open-source software for typesetting and for statistical computing was used, that is L^AT_EX for typesetting, R for statistical computing and the **knitr**-package for literate programming.

Inhaltsverzeichnis

I Einführung in R, Algebra	2
1 Start with R	4
2 Algebra review	16
2.1 Wichtige Aspekte	16
2.2 Formelsammlung	20
2.2.1 Termumformungen	20
2.2.2 Brüche	20
2.2.3 Potenzen und Wurzeln	21
2.2.4 Logarithmen	22
II Wahrscheinlichkeitstheorie	23
3 Wahrscheinlichkeit	25
3.1 Grundbegriffe	25
3.2 Laplace-Wahrscheinlichkeit, objektive und subjektive Wahrscheinlichkeit .	30
3.3 Bedingte Wahrscheinlichkeit	32
3.4 Unabhängigkeit	36
3.5 Der Satz von Bayes	38
3.6 Bayesianische Statistik*	44
4 Zufallsvariablen und ihre Verteilung	45
4.1 Zufallsvariable	45
4.2 Verteilungen	48
4.3 Wichtige Verteilungen	51
4.3.1 Binomialverteilung	52
4.3.2 Poisson-Verteilung	53
4.3.3 Normalverteilung	54
4.4 Andere Verteilungen	58
4.5 Verteilungen mit R	61
4.6 Erwartungswerte	63
4.6.1 Erwartungswert	63
4.6.2 Varianz und Standardabweichung	64

4.6.3	Erwartungswerte von wichtigen parametrischen Verteilungen.	66
4.7	Gemeinsame Verteilungen	67
4.8	Kovarianz und Korrelation	69
4.8.1	Rechenregeln	70
III	Statistik	72
5	Beschreibende Statistik	74
5.1	Statistische Einheiten und Variablen	74
5.2	Typen von Merkmalen	74
5.2.1	Skalenniveau	75
5.2.2	Quantitative versus qualitative Merkmale	76
5.2.3	Stetige versus diskrete Merkmale	76
5.3	Stichprobe	77
5.4	Empirische Verteilungen	77
5.4.1	Häufigkeiten	78
5.4.2	Kumulierte Häufigkeiten	83
5.4.3	Quantile	84
5.5	Kennwerte von Verteilungen	85
5.5.1	Lagemasse	85
5.5.2	Streuungsmasse	90
5.6	Standardisierung	93
5.7	Deskriptive Zusammenfassungen in R	97
6	Multivariate Beschreibung	109
6.1	Empirische Kovarianz und Korrelation	109
6.2	Lineare Regression	115
6.3	Multiple Regression*	125
6.4	Ziele von Regressionsanalysen	127
7	Von der Stichprobe zur Population	128
7.1	Punktschätzung	128
7.1.1	Schätzstatistik und Schätzwert	129
7.1.2	Schätzen von μ	129
7.1.3	Standardfehler	130
7.1.4	Zusammenfassung	131
7.2	Verteilungsaussagen	132
7.2.1	Zentraler Grenzwertsatz	132
7.2.2	Exakte Aussagen bei Normalverteilung	133
7.3	Konstruktion von Schätzern*	134
7.3.1	Methode der kleinsten Quadrate	134
7.3.2	Maximum-Likelihood-Methode	135

7.4	Intervallschätzung	137
7.4.1	Konfidenzintervall	137
7.4.2	Frequentistische Wahrscheinlichkeit*	141
7.5	Schätzen von anderen Größen	143
7.5.1	Schätzen bei kontinuierlichen Daten und 2 Gruppen	143
7.5.2	Schätzen bei dichotomen Outomes und zwei Gruppen	147
7.6	Dualität mit Hypothesentests	152
8	Hypothesentests	154
8.1	Prinzipien von statistischen Tests	154
8.2	z -Test	156
8.3	t -Test	158
8.4	Zwei-Stichproben Problem. t -Test für unabhängige Stichproben	162
8.5	t -Test für gepaarte Daten	168
8.6	Teststärke	172
8.7	Das Testen vom Strohmann*	177
8.8	Äquivalenztests*	182
9	Verteilungsfreie Testverfahren	185
9.1	Wilcoxon-Vorzeichen-Rang-Test	185
9.2	Wilcoxon-Rangsummentest	188
10	Verfahren für Häufigkeitsdaten	195
10.1	Eine kategoriale Variable	195
10.2	Zwei kategoriale Variablen	198
IV	Lineare Modelle	207
11	Lineare Modelle	209
11.1	Lineare Funktionen	209
11.2	Lineare Regression	211
12	Kurze Matrixalgebra	221
13	Das Allgemeine Lineare Modell (LM)	231
13.1	Das Modell	232
13.2	Spezialfälle	234
13.3	Die Methode der kleinsten Quadrate	237
13.4	Annahmen	238
13.5	Die Geometrie der Regression	238
13.6	Erwartungswerte, Varianz und Verteilung der Schätzer	243
13.6.1	Erwartungswerte und Varianz der Schätzer	243
13.6.2	Verteilung der Schätzer bei Normalverteilung	244

13.7 Tests und Konfidenzbereiche	244
13.7.1 <i>t</i> -Test	244
13.7.2 Modellvergleich und <i>F</i> -Test	245
13.8 Beispiel: Lineares Modell für Fertilität	247
13.9 Testen von Hypothesen als Modellvergleiche	253
13.10 Residuenanalyse und Modellannahmen	256
14 Overfitting und Modellwahl	262
14.1 Überanpassung	262
14.2 Modellwahl	265
14.3 Schrittweise Prozeduren*	266
15 Kategoriale Eingangsgrößen	268
15.1 Eine kategoriale Eingangsgröße	268
15.1.1 Modell	268
15.1.2 Beispiel	270
15.2 Zwei kategoriale Eingangsgrößen	277
15.2.1 Modell	277
15.2.2 Beispiel	278
15.3 Eine kategoriale und eine kontinuierliche Eingangsgröße	288
15.3.1 Modell	288
15.3.2 Beispiel	289
16 Generalisierte Lineare Modelle (GLM)*	302
16.1 Allgemeiner Fall	302
16.2 Spezialfälle	303
16.2.1 Lineares Modell	303
16.2.2 Logistische Regression	303
16.2.3 Poisson Regression	304
17 Gemischte Modelle (LMM, GLMM)*	305
17.1 Korrelierte Daten	305
17.2 Notation für wiederholte Messungen	305
17.3 Gemischte Modelle	307
A Notation	309
B Testübersicht und Tabellen	311
Literaturverzeichnis	318

Inhaltsverzeichnis

Abschnitte mit (*) können weggelassen werden ohne Verlust der Kontinuität.

Teil I

Einführung in R, Algebra

To understand computations in R, two slogans are helpful: Everything that exists is an object. Everything that happens is a function call. John Chambers.

Kapitel 1

Start with R

Purpose. The purpose of this chapter is to help get you set up and started with R, the software we will use at the School of Health Professions at the ZHAW. Even if you have previously used R this will be a second view/refresher and may introduce you to a few new ideas. Do not worry if you can't get everything to work on the first try! If you have already installed R and RStudio, you may switch directly to [1](#).

R vs. RStudio. R is an open-source statistical software that was first publicly introduced in 1993. RStudio is a user-friendly interface for R that was founded in 2008. We will use R through RStudio. I will refer to both R and RStudio interchangeably and make a distinction between the two only when necessary. The current version of R is R version 4.5.1 (2025-06-13) [\[40\]](#).

Installation of local Desktop versions. To access R through RStudio:

1. First install R
 - Go to <http://www.r-project.org/>.
 - In the menu to the left, under “Download,” click on the link “CRAN.”
 - Select a mirror site (i.e. <https://stat.ethz.ch/CRAN/>).
 - In the first boxed section titled “Download and Install R,” click on the operating system for your computer.
 - *Windows.* Click on the link “base”. Then download the latest version of R. Wait for download and then follow instructions that appear.
 - *Macintosh.* Click on the link to download the latest version of R. Wait for download. Click “Continue” or “Agree” at every step. No need to customize.
2. Then install RStudio

Once you have installed R on your computer, it is ready for you to use (by double-clicking on the R icon). However, there is a more user-friendly environment for using R, called RStudio, which we will use in this class. RStudio is an integrated

development environment (IDE) for R. It requires that you already have R installed on your computer, which you did in the previous step.

To download RStudio, go to <http://rstudio.org/> and click on the button “Download RStudio.”

Note: The remainder of this handout assumes you’re using RStudio.

Basics.

- The different windows in RStudio.
 - **Console window.** This is where you type commands. BUT: Most of the time, you should edit your code in a **script file** in the Text editor window for reproducibility!
 - **Text editor window.** This is where you can **edit your script files**.
 - **Environment window.** This shows all **objects that you have created**. (More on “objects” later). The workspace is your current R working environment and includes any user-defined objects (vectors, matrices, data frames, lists, functions). At the end of an R session, the user can save an image of the current workspace that is automatically reloaded the next time R is started.
 - **History window.** This shows a history of commands you’ve typed directly into the Console window.
 - **Files window.** File management window similar to Windows Explorer or Mac Finder.
 - **Plots window.** This is where plots that you create will appear.
 - **Packages window.** List of R packages. More on ‘packages’ later.
 - **Help window.** Self-explanatory.
 - R works in a question-and-answer style: in general, you enter a command at the command prompt (>) and press Enter, and R returns an answer: either the result you requested, an error/warning message, or a + prompt which is a signal for you to complete your input, or a > prompt which means it did what you asked but didn’t need to return you anything.
- Keep in mind:** R will do whatever you ask it to do, as long as you ask it correctly (i.e. correct syntax). But R may not give you what you expect, but most often that is because you asked it the wrong question in the first place. So you still need to first *think* about what you want and then translate it (correctly) into R.
- *Comments.* Anything preceded by the hash sign (#) is considered a comment, and will be ignored by R. This is useful for documenting your code.
 - *Recalling previous commands.* To recall previous commands, hit the up-arrow key.

- *Case-sensitive.* R is case-sensitive so that `x` is different from `X`.
- To understand computations in R, two slogans are helpful:
 - Everything that **exists** is an **object**.
 - Everything that **happens** is a **function call**.

R as a calculator. The simplest task of R is to act as a calculator. What happens when you type the following? Is this what you expected?

- `3+2`
- `3*2`
- `3^2`
- `3/2`
- `(3+2)*4`
- `sqrt(4)`
 - Note: `sqrt()` is an example of a “function” and 4 is its “argument”.
- `sqrT(4)`
 - Why did you get the response (an error message) you did? *Hint:* Compare this input with the one above and remember that R is case-sensitive.
- Hit the up-arrow key once. What happens?
- Hit the up-arrow key twice. What happens?
- Keep hitting the up-arrow key...
- `log(64)`
 - Note: `log()` is another example of a "function" and 64 is its argument"
- `#log(64)`
- `sqrt(-1)`
 - Note: `NaN` stands for “Not a Number.” Note that R will evaluate this expression for you, but also give you a warning message.
- `5+`
 - Note: This is an example of an *incomplete* input. When this happens, R gives you the `+` prompt, telling you that it is waiting for you to complete the input. In this case, type something that will complete the input.

Projects and script files. Let’s pause and talk about script files. An R script file is simply a **plain text file** containing a series of commands to be executed by R. A good script file will generally also include comments that document the purpose of the

commands.

One advantage of R is that it is command-based, which in turn allows us to keep a script file as a “record” of our work. Let’s create a script file for this lab session:

1. Let’s first create a new “project” by choosing **File -> New Project**. When the pop-up window appears, select “New Directory” and save it in a location that you’ll remember, such as a folder dedicated to this class. Now, every file you create for this lab will “live” inside this project.
2. Now create a new script by choosing **File -> New -> R Script**
3. Save the file with a .R extension
4. Make a header at the top of your script that includes: the purpose of your script, the author, the date the script was created, the date the script was last modified, the location where your script is saved, any data files that are used, and any other notes. (Note: a header is not required for your code to run, but it should be included as **good practice**). Here is an *example* header in an R Script file (note the use of # for comments):

```
#####
## Autor: Hans Muster
## Erstellt: 1.9.2021
## Zuletzt geändert: 2.9.2021
## Verzeichnis: ~/Rwork
## Ziel: Intro R
## Notizen: ...
#####
```

5. Now you’re ready to add code (with comments!) to your script file. Type commands in the script file (which you can think of as a blank notepad).
6. To actually execute the command, you need to submit it to the Console - place your cursor on the line with the command to be executed and type
 - **Command-Enter** (on a Mac)
 - **Control-Enter** (on a PC).

Assignments and Naming Variables. Almost always you’ll want to save the objects (e.g. a number, a dataframe) you’re using under some name (technically you’re saving the objects to your workspace). To do this, use the assignment operator, **->**. For example, suppose for some reason we want to save the value 5 into the object **x**.

```
x <- 5
```

Notice that nothing really exciting happens on our screen, but in the background, R has now stored the value of 5 into the object called **x** in your workspace. Now type the following and record what happens:

```
x
x+10
y<-x*4
y
```

Note on Naming Conventions. Here we chose `x` to be the name of our variable. However, we could have chosen **almost any name** that's a combination of letters, numbers, and the period (.). There are a few conditions on the name: no spaces, cannot start with a number nor a period followed by a number, and are case-sensitive (e.g. `X` is not the same as `x`). There are also some other “reserved” names such as `c`, `q`, `t`, `C`, `D`, `F`, `I`, `T` and others. Some of these letters may look familiar to you; that should clue you in on why they are reserved names and should not be names that you use for your variables. Do not choose characters such as π as names, you will overwrite it. Type the following and see what happens:

```
pi
pi <- 4
pi
```

Let us correct this by removing the created object from the workspace

```
rm(pi)
pi

## [1] 3.142
```

Vectors. In statistics, we often don't deal with single numbers at a time. Rather, we deal with many numbers at a time. The collection of these scores can be represented as a *vector* of numbers. Type the following into R and record what happens:

```
1:10
c(1,2,3,4,5,6,7,8,9,10)
c(2,4,5,2,12,4,2,3,1)
seq(1,25,by=2)
seq(5,-5,by=-1)
seq(1,10,length=2)
```

The function `c()` combines values or elements into a vector. Create a vector named `myvector` with the sequence of values $5, 6, 7, \dots, 15$:

```
myvector <- c(5:15)
myvector
```

Each number in the vector is called an “element” of the vector. Sometimes we are only interested in a *subset* of the elements. We can view this subset using the `[]` (subset) command. Type each of the following and record what happens:

```
myvector[1]
myvector[-1]
myvector[c(1,4)]
myvector[2:5]
```

Create a vector named `myvector2`

```
myvector2<-c(222,9,1,88,55)
```

We can do calculations (and compute our first statistics) with the data in the vector. Type each of the following and record what happens:

```
sum(myvector)
mean(myvector)
sd(myvector)
summary(myvector)
length(myvector)
sort(myvector2)
rank(myvector2)
```

Other Data Structures. We've seen how to work with single scalar numbers and vectors. You might be asking yourself, What about matrices and other data structures? Now is a good time to pause and consider this question. **R stores everything as “objects.”** Each object is of a particular structure, e.g. vector, matrix, factor, list, and data frame:

Vectors. A *vector* is a string of either **numeric**, **character** or **logical** elements. A single scalar number is a vector of length 1. Examples of vectors:

```
1:9 ## integer
c(1.2,2.4,3.6,6) #numeric
c("Math", "Economics", "Biology", "Psychology") #character
c(1==1,2==3,sqrt(9)==2,log(1)==0) #logical vector (In 1==1, "==" is a relational operator with value TRUE or FALSE)
c(24,"hours")
4 #vector with one element
charvector <-c("Math", "Economics", "Biology", "Psychology")
charvector
str(charvector) ##structure of the object, we see this is a character vector
mode(charvector) ## type of data
```

Matrix. A $n \times m$ *matrix* is an array of either numeric or character or logical elements. We usually work with matrices of numeric elements. Examples:

```
matrix(1:9,nrow=3)
matrix(1:9,nrow=3,byrow=TRUE)
```

By default, R works with **column-major-order**, see [link](#).

```
matrix(c("A", "B", "C", "D", "E", "F"), nrow = 3) ## Default: column-major
matrix(c("A", "B", "C", "D", "E", "F"), nrow = 3, byrow=TRUE) ## row-major
A<-matrix(c(3,4,5,6,7,8,9,10,11),nrow=3)
A
str(A) ## Structure of the object
mode(A) ## Type of data
```

We can extract elements, rows and columns from a matrix by using the [row, column] (subset) command.

```
A[,1] ##choose first column
A[1,] ##choose first row
A[1:2,] ##row 1 and 2
A[2,2] ##row 2 column 2
```

Factor. A *factor* will be important when we work with **categorical data**. Type each of the following and record what happens:

```
sex <- factor(c("male", "female", "female", "male", "male", "male", "female"))
sex
```

R will assign the integer 1 to the level “female” and the integer 2 to the level “male” (because “f” comes before “m”, even though the first element in this vector is “male”)! You can check this by using the function `levels()`, and check the number of levels using `levels()`:

```
levels(sex)
nlevels(sex)
str(sex) ## structure of the object
mode(sex) ## type of data: In R's memory, factors are represented by integer numbers (1, 2, 3).
```

We can count the number of occurrences of each level with `table()`.

```
table(sex)

## sex
## female    male
##      3      4
```

Sometimes, we have to declare integer vectors as factors, and set a character vector for the level attributes.

```
sex2<-c(1,2,1,2,1,2,1,2,1)
sex2<-as.factor(sex2)
sex2
levels(sex2)<-c("man", "woman")
sex2
```

Data Frame. A *data frame* is a **matrix-like structure** in which the columns can be of different types (e.g. numerical, character, logical and categorical/factors). **We will use data frames very often** to represent data with variables in columns and observations in rows.

Let us create four variables X_1, X_2, X_3 and X_4 with 3 observations each.

```
X1<-c(1,2,3)
X2<-c("A","B","C")
X3<as.factor(c("small","moderate","large"))
X4<-c(2==2,3==3,4==5)
str(X1) ##to check the structure of the object
mode(X1)##to check the type of the object
str(X2)
mode(X2)
str(X3)
mode(X3) ## factors are stored as integers
str(X4)
mode(X4)
```

From these variables, we create a data.frame.

```
mydf1<-data.frame(X1,X2,X3,X4)
mydf1
str(mydf1)
mode(mydf1) ##Data frames are a special case of a list, see below
summary(mydf1)
```

List. A *list* is an *ordered collection* of objects. Lists can contain objects of different type and different length. Therefore, **vectors and data frames are special cases of lists**.

```
mylist<-list(height=170,age=c(23,16,14),color=c("green","blue","red"))
mylist
as.data.frame(mylist) ## the value 3 is recycled!
mylist2<-list(height=170,age=c(23,16,14,34),color=c("green","blue","red"))
mylist2
##as.data.frame(mylist2) ## comment out and check error message, not matrix-like.
```

Subsetting in lists: `[]` can be used to select a single element, dropping names, whereas `[` keeps them.

```
mylist[1] ## Element of a list
mylist[[1]] ## dropping names
mylist[2]
mylist[[2]]
```

Exploring data. R comes with many datasets ready for your use. One of these datasets records time between eruptions of the Old Faithful geyser in Yellowstone National Park, https://de.wikipedia.org/wiki/Old_Faithful. It's called `faithful` and to access it, just type `faithful` at an R prompt:

```
faithful
```

You may have noticed that this just prints a bunch of data onto your screen, which isn't particularly useful for exploring the data. It is good practice to "save" the data into an object with a name of your choice, e.g. "d.faith":

```
d.faith <- faithful
```

Notice that R appears to not have done anything. In fact, it *did* do something. In particular, it saved the `faithful` data in your workspace under the name `d.faith`. Now we can work with this object to explore its contents.

As a first step to working with data, you should understand which variables are recorded, how many observations there are, etc. There are many ways to do this in R. Type the following commands and record what happens:

```
d.faith  
str(d.faith)  
head(d.faith)  
names(d.faith)  
nrow(d.faith)  
ncol(d.faith)  
dim(d.faith)
```

The command `names(d.faith)` tells you the names of the variables in this dataset. Here, we have `eruptions` and `waiting`. What do these variables represent? It's not very intuitive (as a lesson to you, when you have the chance to name variables, choose names that are as descriptive as possible yet that don't have spaces or punctuations), but fortunately since this dataset came pre-installed with R, there is a help file on it.

```
?faithful
```

In the "Format" section of the help file, you can see a brief description of the 2 variables. We find that `eruptions` gives the eruption times in minutes, and `waiting` gives the waiting time to next eruption (in minutes). Each observation (row) in the dataset corresponds to a particular eruption of the Old Faithful geyser in Yellowstone National Park.

What is the *average eruption time* of Old Faithful? The answer to this question is an example of a *summary statistic*. We can request summaries of particular variables in the dataset by using the `$` syntax. In general, we use the form *Name of your data object\$Name of your variable*. For example, record what happens when you type:

```
mean(d.faith$eruptions)
```

An alternative would be

```
mean(d.faith[,1])  
mean(d.faith[,"eruptions"])
```

Often, we will use the `summary()`-function to describe variables in a data frame object.

```
summary(d.faith)
```

We can also look the distribution of eruption times. A histogram helps us explore this:

```
hist(d.faith$eruptions)
```

We can also look at the bivariate distribution of eruption times and waiting times.

```
plot(d.faith,col="blue")
```

This should open a pop-up window with a histogram. More on this later, but just for fun – do you notice anything interesting in the distribution of eruption times and waiting times? One advantage of R is that it can produce very nice-looking graphs. The default graph create here isn't so exciting, but with a little more experience with R, we can change that.

The graph you made here used a command from base R.

Saving graphs. One (preferred) way to save graphs is by enclosing your commands for plotting with the “pdf” command. This is an automated approach to plotting, and as such is good for reproducible research.

```
pdf(file = "myfirstplot.pdf")  
hist(d.faith$eruptions)  
dev.off()
```

```
pdf(file = "mysecondplot.pdf")  
plot(d.faith) # OR plot(d.faith$eruptions,d.faith$waiting)  
dev.off()
```

These files will be saved in your project (and thus in the same directory as your project). Another, less preferred, way to save a graph uses a more point & click approach (and as such is not automated nor reproducible):

1. In the Plots window, use the Left/Right arrow icons to browse through the plots until the graph you wish to save appears.
2. Click on Export and select the option you wish to use.
 - If you're preparing a report in Word, you may want to “copy to clipboard” and then paste the graph into your Word file.

- If you're preparing a report in L^AT_EX, then you may want to choose "Save as PDF".
- If you wish to keep a file of your graph, then choose either "Save as PDF" or "Save as image."

R Notebooks*. In this lab, you typed all of your code in a "script" file (.R extension); everything you typed had to be code, unless you commented it out. Soon you will instead use R Notebooks (.Rmd extension), which are files that take both R code for doing data analysis with text of a paper that presents results of the analysis. As such, R Notebooks facilitate reproducible research. You will be using these to prepare all of your homework and projects for this class.

We'll introduce R Notebooks soon, but in the meantime, to get a very early flavor of R Notebooks, check out these resources:

- https://rmarkdown.rstudio.com/r_notebooks
- http://rmarkdown.rstudio.com/authoring_basics.html

If you're adventurous, you can even start exploring R Notebooks yourself by opening a new file and selecting "R Notebooks" in RStudio.

Explore! This is only the tip of the R iceberg. Don't be afraid to explore and be creative; build upon these basics. Read the help files or other resources listed below to find out more about particular commands. This way, you'll grow in confidence and comfort with R.

Help! Where can I get help for R?

- To obtain help on any of the commands, you can either access R-Help from the menu or by typing at the command prompt:

```
> ?hist
```

This will bring up a separate help window on the command you typed.

- If you do not know the exact name of the command, you can do this instead:

```
> help.search("histogram")
```
- Google your question. There are a *lot* of R resources on the web. *Hint:* Include the letter "R" in your key word searches.
- RStudio Support pages: <https://support.rstudio.com/hc/en-us>
- The World Wide Web
- *An Introduction to R*, available at <http://cran.r-project.org/doc/manuals/R-intro.html>.

- There are other manuals (some on more advanced topics) located at: <http://www.r-project.org/>. Click on the **Manuals** link on the left menu bar, for a list of R manuals.
- <http://cran.r-project.org/doc/contrib/usingR.pdf>

Background* Some S facts

- S is a language and system for organizing, visualizing, and analyzing data.
- S has been a project of statistics research at Bell Labs since 1976.
- The language has evolved through several major versions to become the most widely used environment for research in data analysis and statistics.
- In 1998, S became the first statistical system to receive the Software System Award, the top software award from the ACM.

Products and projects based on S

- The S-Plus language is based on the S software from Bell Labs. S-Plus products are distributed by the Insightful Corporation, which has an exclusive license to distribute software based on Bell Labs' S.
- The R language is an open-source system distributed under the GPL license. It is a separate project based on the S language, with a number of differences to Splus.
R is what we will be using in this class.

Some R facts

- R is an environment for data analysis and visualization.
- R is an open source implementation of the S language (S-Plus is a commercial implementation of the S language).
- The current version of R is R version 4.5.1 (2025-06-13)

The R software

- R is mainly written in C.
- R is available for many platforms:
 - Unix of many flavors, including Linux, Solaris, FreeBSD, AIX.
 - Windows 95 and later.
 - MacOS X.
- Binaries and source code are available from www.r-project.org.
- R “talks” to data bases, programming languages, and other statistical packages.
- R should be source code compatible with most of the Splus code written.

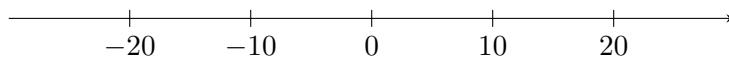
Kapitel 2

Algebra review

2.1 Wichtige Aspekte

Mengen. Eine Menge stellen wir mit geschweiften Klammern dar: $\{\dots\}$. Für eine Menge von x , die die Bedingung y erfüllen, schreiben wir $\{x \mid y\}$. Die Notation $x \in y$ bedeutet: x ist Element von y . Wichtige Zahlenmengen sind nun:

- \mathbb{N} : Menge der natürlichen Zahlen $\{1, 2, 3, \dots\}$
- \mathbb{Z} : Menge der ganzen Zahlen $\{\dots, -2, -1, 0, 1, 2, \dots\}$
- \mathbb{Q} : Menge der rationalen Zahlen, Brüche: $\{\frac{a}{b} \mid a \in \mathbb{Z}, b \in \mathbb{N}, b \neq 0\}$
- \mathbb{R} : Menge der reellen Zahlen. Zusätzlich zu den rationalen Zahlen kommen die irrationalen Zahlen wie z.B.
 - $\pi = 3.14159265358979$
 - $\sqrt{2} = 1.4142135623731$
 - $e = 2.71828182845905$

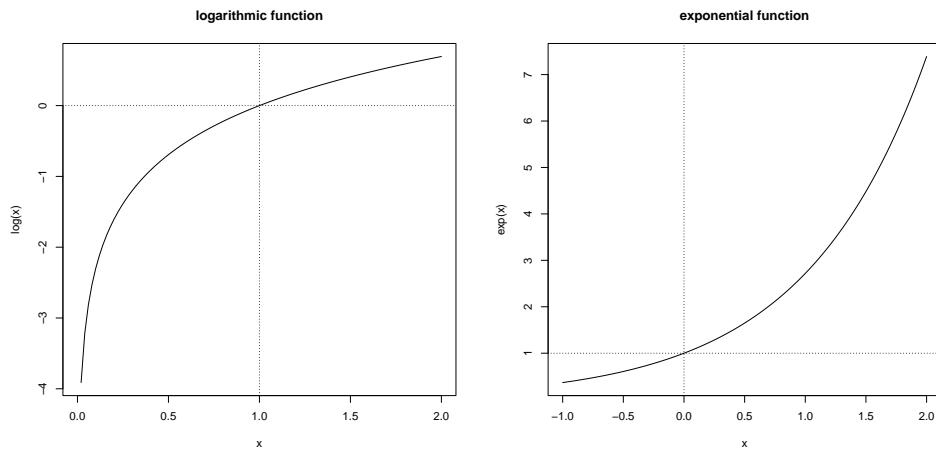


Algebra. Wichtige Aspekte sind:

- Symbole als Variablen verstehen, und für was sie stehen.
- Rechnen mit Klammern
- Ausklammern
- Negation ausklammern
- Brüche kürzen und erweitern
- Addition, Subtraktion, Multiplikation and Division von Brüchen
- Doppelbrüche
- Potenzieren

- Potenzieren mit ganzen Zahlen und mit Brüchen im Exponenten: $a^{1/n} = \sqrt[n]{a}$
- Potenzieren mit Summe im Exponenten: $a^{m+n} = a^m \times a^n$
- Potenzieren mit Differenz im Exponenten : $a^{m-n} = a^m/a^n$
- Potenzieren mit negativem Exponenten: $a^{-m} = a^{0-m} = \frac{a^0}{a^m} = 1/a^m$.
- Wenn $a = e = 2.718$, schreiben wir häufig $\exp(x)$ für e^x
- Spezielle Exponenten: $a^0 = 1$ und $a^1 = a$
- n -te Wurzeln
 - $\sqrt[n]{a} = a^{1/n}$
 - $\sqrt[n]{\frac{1}{a}} = (\frac{1}{a})^{1/n} = (a^{-1})^{1/n} = a^{-\frac{1}{n}}$
- Logarithmus
 - log zur Basis b von x , $y = \log_b x$, ist die Zahl y so dass $b^y = x$
 - $\log_b b = 1$
 - $\log_b b^x = x \log_b b = x$
 - $\log_b a^x = x \log_b a$
 - $\log_b a^{-x} = -x \log_b a$
 - $\log_b(xy) = \log_b x + \log_b y$
 - $\log_b \frac{x}{y} = \log_b x - \log_b y$
 - Basiswechsel: $\log_a b = \frac{\log_c b}{\log_c a}$
 - Wenn $b = e = 2.718$, ist die Basis der natürliche Logarithmus. $\log_e x$ schreiben wir kurz $\log x$ oder $\ln x$
 - $\log e = 1$
- Die Umkehrfunktion vom log zur Basis b von x ist b^x
- Logarithmus $\log(\cdot)$ und Exponentialfunktion $\exp(\cdot)$ Funktion sind zentral in der Wissenschaft und in der Statistik.
 - $\log(x) : \mathbb{R}^+ \rightarrow \mathbb{R} : y = \log(x)$
 - $\exp(x) : \mathbb{R} \rightarrow \mathbb{R}^+ : y = \exp(x)$

```
curve(log(x),from=0,to=2,main="logarithmic function")
abline(v=1,h=0,lty=3)
curve(exp(x),from=-1,to=2,main="exponential function")
abline(v=0,h=1,lty=3)
```

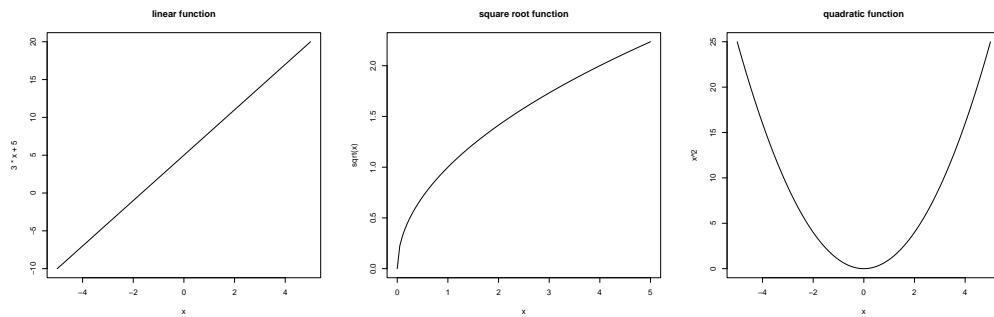


- Funktionen verstehen wie $y = a + bx$
- Indikatorvariablen: $I_{x=3}$: wahr, wenn $x = 3$, sonst falsch (oder 1 wenn $x = 3$, 0 sonst),

$$I_{x=3} = \begin{cases} 1 & x = 3 \\ 0 & x \neq 3. \end{cases}$$

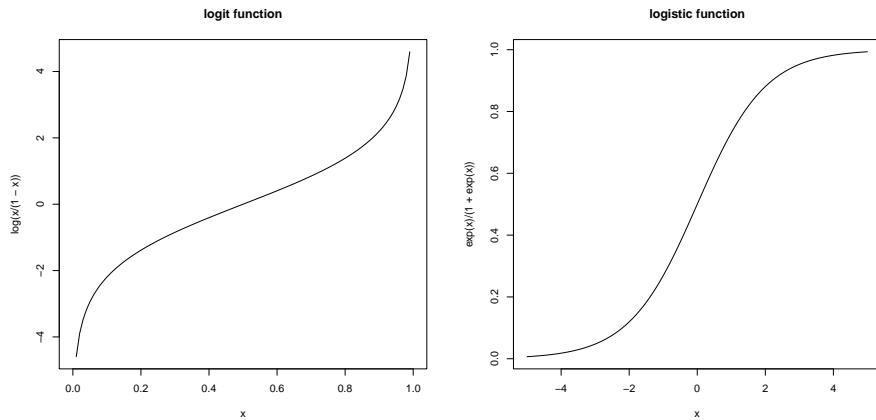
- Lineare Gleichungen $ax + b = 0$
- Quadratische Gleichungen $ax^2 + bx + c = 0$
- Funktionen darstellen. (Mit R wird das einfach sein.)

```
curve(3*x+5,from=-5,to=5,main="linear function")
curve(sqrt(x),from=0,to=5,main="square root function")
curve(x^2,from=-5,to=5,main="quadratic function")
```



- Die Logistische ($\text{logistic}(\cdot)$) und Logit-Funktion ($\text{logit}(\cdot)$) werden wir häufig antreffen. Eine ist die Umkehrfunktion der anderen.
 - $\text{logit}(x) : [0, 1] \rightarrow \mathbb{R}, \quad x \mapsto \log \frac{x}{1-x}.$
 - $\text{logistic}(x) : \mathbb{R} \rightarrow [0, 1], \quad x \mapsto \frac{\exp(x)}{\exp(x)+1}.$

```
curve(log(x/(1-x)),main="logit function")
curve(exp(x)/(1+exp(x)),from=-5,to=5,main="logistic function")
```



- Wenn wir zum linearen Modell kommen (am Ende des Moduls und im nächsten Modul), wird Notation mit Vektor/Linearer Algebra wichtig sein:

– Sei \mathbf{x} ein (Kolumnen)vektor mit den Grössen x_1, x_2, \dots, x_p , also $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$

(z.B. die Werte von p Variablen einer Person wie Alter, Blutdruck, usw.)

– Sei $\boldsymbol{\beta}$ ein (Kolumnen)vektor mit den Grössen $\beta_1, \beta_2, \dots, \beta_p$, also $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

(Gewichte/Regressionskoeffizienten)

– \mathbf{x}^T ist dann ein Zeilenvektor (T : “transponiert”) und

– $\mathbf{x}^T \boldsymbol{\beta} = (x_1, x_2, \dots, x_p) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ ist dann ein

Skalarprodukt (eine Zahl) und heisst linearer Prädiktor, die mit den Gewichten $\beta_1, \beta_2, \dots, \beta_p$ gewichtete Summe der x_1, x_2, \dots, x_p .

2.2 Formelsammlung

2.2.1 Termumformungen

Kommutativgesetz	Assoziativgesetz
$a + b = b + a$	$a + (b + c) = (a + b) + c$
$ab = ba$	$a(bc) = (ab)c$
Distributivgesetz	Ausklammern
$a(b + c) = ab + ac$	$ab + ac = a(b + c)$
Binomische Formeln	
$(x + y)^2 = x^2 + 2xy + y^2$	
$(x - y)^2 = x^2 - 2xy + y^2$	
$x^2 - y^2 = (x + y)(x - y)$	

Tabelle 2.1: Termumformungen

2.2.2 Brüche

Kürzen	Erweitern
$\frac{ax}{ay} = \frac{x}{y}$	$\frac{x}{y} = \frac{ax}{ay}$
Addieren und Subtrahieren	
$\frac{a}{b} \pm \frac{c}{d} = \frac{ad \pm cb}{bd}$	
Multiplizieren	
$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$	$\frac{a}{b} : \frac{c}{d} = \frac{a}{b} \cdot \frac{d}{c} = \frac{ad}{bc}$
Dividieren	
Doppelbrüche	
$\frac{a/b}{c/d} = \frac{a}{b} : \frac{c}{d} = \frac{ad}{bc}$	

Tabelle 2.2: Brüche

2.2.3 Potenzen und Wurzeln

$\underbrace{a \cdot a \cdots a}_{n \text{ mal}} = a^n$	$\sqrt[n]{a} = a^{\frac{1}{n}}$
Rechenregeln	
$a^m a^n = a^{m+n}$	$\sqrt[m]{a} \sqrt[n]{a} = a^{\frac{m+n}{mn}}$
$\frac{a^m}{a^n} = a^{m-n}$	$\frac{\sqrt[n]{a}}{\sqrt[m]{b}} = \sqrt[m]{\frac{a}{b}}$
$(a^m)^n = a^{mn}$	$(\sqrt[n]{a})^m = \sqrt[m]{a^m}$
$a^n b^n = (ab)^n$	$\sqrt[m]{\sqrt[n]{a}} = \sqrt[mn]{a}$
$\frac{a^n}{b^n} = (\frac{a}{b})^n$	$a^{p/q} = \sqrt[q]{a^p}$
$a^{-n} = (\frac{1}{a})^n$	$a^{-1/n} = \sqrt[n]{\frac{1}{a}}$
Spezielle Exponenten	
$a^0 = 1$	$a^1 = a$

Tabelle 2.3: Potenzen und Wurzeln

```
9^(1/2) #square root
(27)^(1/3) #3th root
```

2.2.4 Logarithmen

$x = \log_a(b) \Leftrightarrow a^x = b$
Rechenregeln
$\log(ab) = \log(a) + \log(b)$
$\log(a/b) = \log(a) - \log(b)$
$\log(a^x) = x \log(a)$
Identitäten
$a^{\log_a(b)} = b$
$\log_a(a^b) = b$
Spezielle Basen
$\log_{10}(x) = \lg(x)$
$\log_e(x) = \log(x) = \ln(x)$
Basiswechsel
$\log_a(b) = \frac{\log_c(b)}{\log_c(a)}$
$\ln(b) = \log(b) = \frac{\log_{10}(b)}{\log_{10}(e)}$

Tabelle 2.4: Logarithmen

```
a<-10
b<-30
log(a*b)
log(a)+log(b)
log(a/b)
log(a)-log(b)
log(a^5)
5*log(a)
```

```
e<-exp(1) #Eulersche Zahl e
e
log(20) #natürlicher Log mit Basis e
log10(20)/log10(e) #Basiswechsel
```

Teil II

Wahrscheinlichkeitstheorie

Ziel. Einführung von Grundbegriffen, Vermittlung von stochastischem Denken.

Kapitel 3

Wahrscheinlichkeit

In diesem Kapitel führen wir den Begriff der *Wahrscheinlichkeit* ein. Die *Wahrscheinlichkeitstheorie* befasst sich mit Aspekten, in denen der Zufall eine Rolle spielt. Wahrscheinlichkeitsrechnung ist ein mächtiges Werkzeug, das man in der Statistik, in der Epistemologie und Wissenschaftsphilosophie braucht.

Was ist Wahrscheinlichkeit? Alltagssätze haben die Struktur “Die Wahrscheinlichkeit von A ist p ”, mit A als einem *Ereignis* oder einer *Aussage* und p als einer Quantität des *Grades der Überzeugung* in A . Das Verständnis der mathematischen Definition von Wahrscheinlichkeit ist äusserst wichtig in der Statistik.

Am Ende dieses Abschnittes werden wir sehen, dass Wahrscheinlichkeiten numerische positive Quantitäten sind – definiert auf einer *Menge von Ergebnissen* – die additiv sind über sich gegenseitig ausschliessende Ergebnisse und sich auf eins summieren über alle möglichen sich gegenseitig ausschliessenden Ergebnisse.

Grundidee. *Zufallsexperimente* sind Experimente, deren Ergebnisse nicht immer exakt vorausgesagt werden können. Wir möchten dafür ein mathematisches Modell.

3.1 Grundbegriffe

Notation. $\Pr(A)$ steht für die Wahrscheinlichkeit des Ereignisses A . Der Begriff vom Ereignis wird unten definiert. Eine Menge stellen wir mit geschweiften Klammern dar: $\{\dots\}$. Für eine Menge von x , die die Bedingung y erfüllen, schreiben wir $\{x \mid y\}$. \mathbb{N} steht für die Menge der natürlichen Zahlen, \mathbb{Z} für die Menge der ganzen Zahlen, \mathbb{Q} für die Menge der rationalen Zahlen, \mathbb{R} für die Menge der reellen Zahlen und \mathbb{R}_+ für die Menge der nicht-negativen reellen Zahlen. A^C : Komplement von A . \in : Element von, \subseteq, \subset : Teilmenge von, \emptyset : leere Menge, \cup : ODER, \cap : UND, $|A|$: Anzahl Elemente in Menge A , \sum : Summe, \prod : Produkt.

Definition. Der *Ergebnisraum* oder *Stichprobenraum* Ω ist die Menge aller möglichen *Ergebnisse* des betrachteten Zufallsexperimentes. Die Elemente $\omega \in \Omega$ heissen *Elementarereignisse*.

1. Beim Werfen eines Würfels ist $\Omega = \{1, 2, \dots, 6\}$
2. Für die Heilungszeit eines Patienten ist $\Omega = \{t \mid t \geq 0\} = \mathbb{R}_+$
3. Für die (ganzzahligen) Noten bei einer Prüfung ist $\Omega = \{1, 2, 3, \dots, 6\}$
4. Für eine Kraftmessung K ist $\Omega = \{k \mid k \geq 0\} = \mathbb{R}_+$
5. Heilung eines Patienten: $\Omega = \{+, -\}$
6. Heilung von zwei Patienten: $\Omega = \{++, +-, -+, --\}$

Definition. Ein *Ereignis* ist eine Teilmenge $A \subseteq \Omega$, also eine Kollektion von Elementarereignissen. Die Klasse aller *beobachtbaren Ereignisse* ist der *Ereignisraum* \mathcal{F} . Dieser ist eine Teilmenge der *Potenzmenge* $2^\Omega = \mathcal{P}(\Omega)$ von Ω . Die Potenzmenge ist die *Menge aller Teilmengen* von Ω .

Ist Ω endlich oder abzählbar (Beispiele 1, 3, 5 und 6), so wählt man oft als Ereignisraum \mathcal{F} die Potenzmenge 2^Ω . Ist Ω überabzählbar (wie in Beispiel 2 und 4), so muss man als \mathcal{F} eine echte Teilklasse von 2^Ω nehmen (Wir gehen hier nicht auf die Details ein, weil sie für die Praxis unwesentlich sind). In jedem Fall muss der Ereignisraum \mathcal{F} folgende Bedingungen erfüllen:

1. $\Omega \in \mathcal{F}$.
2. Wenn $A_1, A_2, \dots \in \mathcal{F}$, dann $\bigcup A_i \in \mathcal{F}$. (“abgeschlossen unter Vereinigung”)
3. Wenn $A \in \mathcal{F}$, dann $A^C \in \mathcal{F}$. (“abgeschlossen unter Komplementbildung”)

Dasselbe Experiment kann man in der Regel durch verschiedene (Ω, \mathcal{F}) beschreiben:

Beispiel. Jemand macht eine Prüfung, teilt aber nur mit, ob das Ereignis bestanden oder nicht bestanden eingetreten ist.

- Erste Beschreibung: $\Omega_1 = \{\text{pass}, \text{fail}\}$ und $\mathcal{F}_1 = \{\emptyset, \Omega_1, \{\text{pass}\}, \{\text{fail}\}\}$.
- Zweite Beschreibung: $\Omega_2 = \{1, 2, 3, \dots, 6\}$; dann darf aber nicht $\mathcal{F}_2 = 2^{\Omega_2}$, weil wir ja die exakte Note nicht beobachten können. Hier wäre dann $\mathcal{F}_2 = \{\emptyset, \Omega_2, \{1, 2, 3\}, \{4, 5, 6\}\}$.

Man beachte, dass \mathcal{F}_1 und \mathcal{F}_2 dieselbe Anzahl von Mengen (beobachtbare Ereignisse) enthalten.

Wir stellen uns allgemein vor, dass wir bei unserem Zufallsexperiment genau ein Elementarereignis ω erhalten. Wir sagen, dass das Ereignis A *eintritt*, falls das realisierte Elementarereignis ω in A liegt, d.h. $\omega \in A$. Mit Hilfe von Mengenoperationen können wir dann aus $A, B \in \mathcal{F}$ neue Ereignisse bilden. Abbildungen 3.1 und 3.2 stellen Ereignisse als *Mengen* dar.

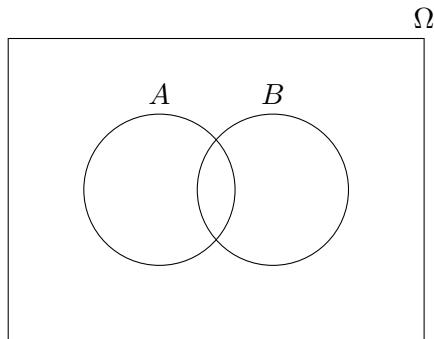


Abbildung 3.1: Venn diagram: Stichprobenraum Ω , Ereignisse A und B .

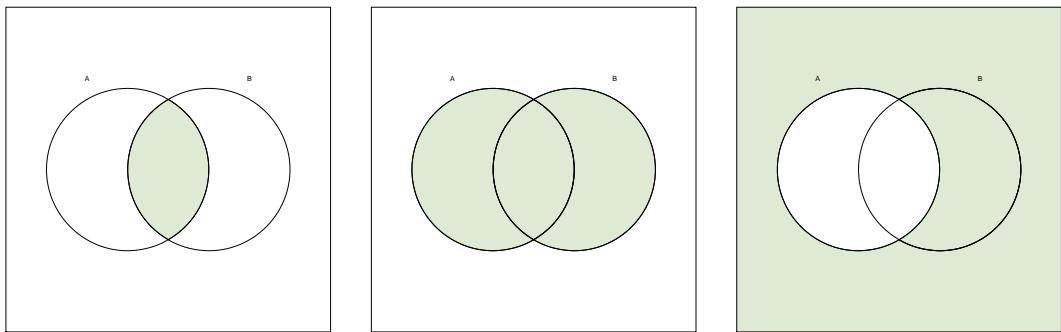


Abbildung 3.2: Schnittmenge $A \cap B$, Vereinigungsmenge $A \cup B$ und Komplement A^C

- Die *Schnittmenge* von zwei Ereignissen A und B besteht aus den Elementen, die sowohl zu A *und* zu B gehören. Wir notieren diese Menge $A \cap B$. Das ist das Ereignis, dass A *und* B eintreten.
- Die *Vereinigungsmenge* von A und B , bestehend aus den Elementen, die zu A *oder* B gehören, notieren wir mit $A \cup B$. Das ist das Ereignis, dass A *oder* B (oder beide) eintreten.
- Das Komplement von A , notiert mit A^C ist das Ereignis, dass A *nicht* eintritt.

Dank der Definition von \mathcal{F} liegen all diese Mengen auch wieder in \mathcal{F} .

Konstruktion von einem Ereignisraum*. Abbildung 3.3 illustriert den kleinsten Ereignisraum (das kleinste \mathcal{F}), der die zwei Ereignisse A und B beinhaltet. Es gibt $2^4 = 16$ Elemente in diesem Raum, *alle Vereinigungen* der Mengen in der ersten Reihe, $A^C \cap B^C, B \cap A^C, A \cap B^C, A \cap B$, die eine *Partition* von Ω darstellen. Unten rechts wäre die leere Menge.

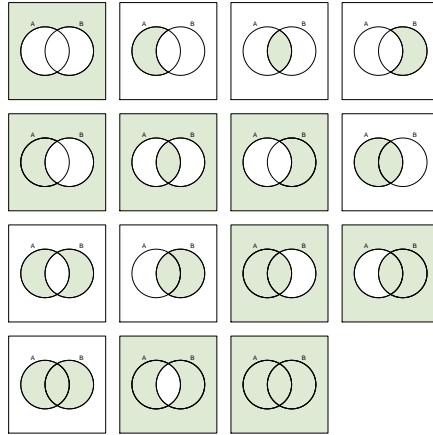


Abbildung 3.3: Ereignisraum \mathcal{F} aufbauend auf den Ereignissen A und B , die erste Zeile stellt eine Partition von Ω dar, die anderen Mengen sind Vereinigungen der Elemente der Partition.

Definition. Ein *Wahrscheinlichkeitsmass* ist eine Abbildung $\mathcal{F} \rightarrow [0, 1]$. Jedem Ereignis im Ereignisraum \mathcal{F} wird eine Zahl zwischen 0 und 1 zugeordnet. Für $A \in \mathcal{F}$ nennen wir $\Pr(A) \in [0, 1]$ die Wahrscheinlichkeit, dass A eintritt.

Bezüglich der Wahrscheinlichkeitsrechnung hat Kolmogoroff (1933) folgende *Axiome* aufgestellt. Diese Axiome gelten für alle Interpretationen von Wahrscheinlichkeit¹:

- Axiom 1: Die Wahrscheinlichkeit ist nicht negativ

$$\Pr(A) \geq 0 \quad (3.1.1)$$

- Axiom 2: Die Wahrscheinlichkeit des *sicheren* Ereignisses Ω ist 1

$$\Pr(\Omega) = 1 \quad (3.1.2)$$

- Axiom 3: Die Wahrscheinlichkeit der Vereinigung von *disjunkten* (sich gegenseitig ausschliessenden) Ereignissen A und B ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse. $A \cap B = \emptyset$, so

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad (3.1.3)$$

- Axiom 3b: Allgemein: Für jede Sequenz von paarweise disjunkten Ereignissen A_1, A_2, \dots (i.e., $A_i \cap A_j = \emptyset$, für $i \neq j$) gilt

$$\Pr\left(\bigcup A_i\right) = \sum \Pr(A_i) \quad (3.1.4)$$

¹objektive oder subjektive Wahrscheinlichkeit, dazu mehr später

Additionsregel. Aus den drei angeführten Axiomen folgen weitere wichtige Sätze der Wahrscheinlichkeitstheorie. Verwandt mit dem 3. Axiom (3.1.3) ist das sogenannte *Additionstheorem*. Es geht dabei um die allgemeine *Additionsregel* für beliebige A und B (die also nicht disjunkt sein müssen).

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \quad (3.1.5)$$

Wir können (3.1.5) mit Hilfe der Abbildung 3.1 nachvollziehen. In der Sprache der Venn-Diagramme ist $\Pr(A) = \frac{|A|}{|\Omega|}$. Die Notation $|\cdot|$ steht für die *Mächtigkeit* der Menge, der Anzahl Elemente in der Menge. Die Additionsregel kann man nun auch allgemein für mehrere Ereignisse verallgemeinern. Hat man zum Beispiel drei Ereignisse A , B und C gegeben, so gilt

$$\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C). \quad (3.1.6)$$

Die allgemeine Formel für n Ereignisse sieht dann aber auch nicht mehr so nett aus, deswegen lassen wir sie mal weg. Auf weitere Folgerungen gehen wir in 3.3 und 3.5 ein.

Am Ende dieses Abschnitts führen wir noch den Begriff der *Chance* oder *Odds* ein.

Definition. Eine Wahrscheinlichkeit p kann auch als *Chance* (oder englisch *Odds*) dargestellt werden. Chancen oder Odds von einem Ereignis sind definiert als das Verhältnis von der Wahrscheinlichkeit zur Gegenwahrscheinlichkeit von einem Ereignis,

$$Odds = \frac{p}{1-p}. \quad (3.1.7)$$

Umgekehrt gilt²

$$p = \frac{Odds}{1 + Odds}. \quad (3.1.8)$$

Odds können nicht-negative reelle Zahlen (Zahlen in \mathbb{R}_+) annehmen. Der Logarithmus der Odds heisst *logit*-Funktion,

$$y = \log \frac{p}{1-p} = \text{logit}(p). \quad (3.1.9)$$

Logits können Werte auf der ganzen reellen Zahlenachse \mathbb{R} annehmen. Die Umkehrfunktion von 3.1.9 nennt man die *logistische* Funktion,³

$$p = \frac{\exp(y)}{1 + \exp(y)}. \quad (3.1.10)$$

Diese zwei Funktionen werden wir später antreffen, wenn wir z.B. Wahrscheinlichkeiten (Zahlen im Intervall $[0, 1]$) auf den reellen Zahlenbereich $(-\infty, \infty)$ abbilden wollen und

² $O = p/(1-p) \Rightarrow O - Op = p \Rightarrow p + Op = O \Rightarrow p(O+1) = O \Rightarrow p = O/(1+O)$

³ $y = \log \frac{p}{1-p} \Rightarrow \exp(y) = \frac{p}{1-p} \Rightarrow p = \frac{\exp(y)}{1+\exp(y)}$

umgekehrt. Diese beiden Funktionen sind in der Abbildung 3.4 dargestellt.

```
curve(log(x/(1 - x)), from = 0, to = 1, xlab = "p", ylab = "logit", main = "logit-Funktion")
curve(exp(x)/(1 + exp(x)), from = -4, to = 4, xlab = "logit", ylab = "p", main = "logistische Funktion")
```

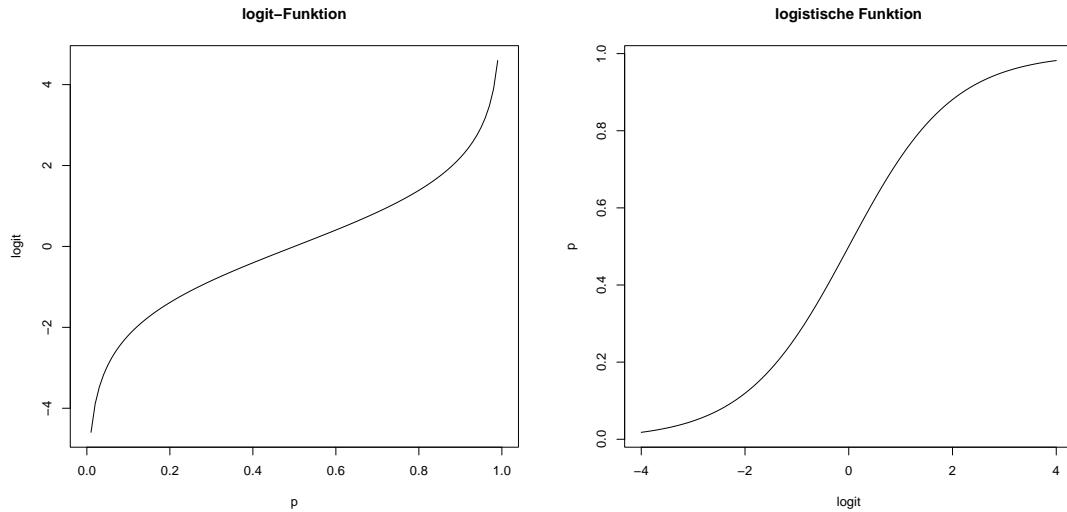


Abbildung 3.4: Die logit-Funktion (links) und die logistische Funktion (rechts).

3.2 Laplace-Wahrscheinlichkeit, objektive und subjektive Wahrscheinlichkeit

Was ist nun die *Wahrscheinlichkeit* $\Pr(A)$ von einem Ereignis A ? Wenn man davon ausgeht, dass alle möglichen Ergebnisse w_i *gleichwahrscheinlich* sind, dann stellt $\Pr(A)$ das Verhältnis der Anzahl Elemente in Teilmenge A relativ zur Anzahl der Elemente in der Menge Ω dar. Dies entspricht dem Verhältnis der *günstigen* zu den *möglichen* Ereignissen. Diese Wahrscheinlichkeit kennen wir aus der Schule, man nennt sie auch *Laplace-Wahrscheinlichkeit*. Sie ist in dem Sinne subjektiv, als man *a priori*, bevor man Daten sammelt, die Elementarereignisse als gleichwahrscheinlich annimmt.

Wenn wir z.B. eine Münze werfen, dann glauben wir *vor* dem Experiment an diese Gleichwahrscheinlichkeit von ‘Kopf’ und ‘Zahl’. Dasselbe gilt beim Würfeln: Glauben wir an einen fairen Würfel, so ist die Wahrscheinlichkeit jedes Elementarereignisses $\{w_i\}$ aus $\Omega = \{1, 2, 3, 4, 5, 6\}$ *a priori* $\Pr(w_i) = 1/6$.

Laplace-Wahrscheinlichkeit:

$$\Pr(A) = \frac{\text{Anzahl günstiger Ergebnisse für } A}{\text{Anzahl möglicher Ergebnisse}}. \quad (3.2.1)$$

Die Laplace-Wahrscheinlichkeit geht also von der a priori postulierten Gleichwahrscheinlichkeit der Elementarereignisse aus. Sind nun aber die Elementarereignisse nicht a priori gleichwahrscheinlich, wie im Beispiel oben mit den Schulnoten, wird der *frequentistische* Wahrscheinlichkeitsbegriff gebraucht. Das ist z.B. der Fall, wenn wir nicht an eine *faire* Münze beim Münzwurf glauben.

Dann brauchen wir eine *gemessene* Wahrscheinlichkeit, eine Wahrscheinlichkeit, die wir anhand von Daten quantifizieren. Die Wahrscheinlichkeit eines Ereignisses A wird dann anhand der *relativen Häufigkeit* vom Eintreten von A bei *häufigem Wiederholen* eines *Zufallsexperiments* beschrieben. Die Wahrscheinlichkeit ist der *Grenzwert* dieser *relativen Häufigkeit* bei sehr häufiger Wiederholung.

Auch wenn man nicht unendlich viele Wiederholungen eines Zufallsexperiments machen und so den Grenzwert wirklich bestimmen kann, so ist es eine empirische Tatsache, dass sich relative Häufigkeiten bei oftmaligem Wiederholen um eine Grösse stabilisieren; diese Grösse nennen wir dann Wahrscheinlichkeit.

Die klassische Statistik baut auf dieser *frequentistischen Interpretation* von Wahrscheinlichkeit oder auf dem sogenannten *objektiven* Wahrscheinlichkeitsbegriff auf. Der frequentistische Wahrscheinlichkeitsbegriff ist der vorherrschende Wahrscheinlichkeitsbegriff in den Wissenschaften. Die Wahrscheinlichkeit $\Pr(A)$ für ein Ereignis A wird hier also durch die *relative Häufigkeit* $f_n(A)$ des Eintreffens von A bei n Wiederholungen quantifiziert, wobei diese Schätzung um so genauer ausfällt, je grösser die Anzahl an Wiederholungen n ist:

$$f_n(A) = \frac{\text{Anzahl der Ergebnisse } A \text{ in } n \text{ Messungen}}{n}. \quad (3.2.2)$$

Objektive Wahrscheinlichkeit:

$$\Pr(A) = \lim_{n \rightarrow \infty} f_n(A) \quad (3.2.3)$$

Neben der Laplace-Wahrscheinlichkeit und der objektiven Wahrscheinlichkeit gibt es noch einen *subjektiven* Wahrscheinlichkeitsbegriff. Dieser sieht Wahrscheinlichkeit als eine persönliche, subjektive Überzeugung des Betrachters. Diese Wahrscheinlichkeit kann mit *Wetten* quantifiziert werden, dabei ist der *Wettquotient* gerade die subjektive Wahrscheinlichkeit für das Eintreten von A . Die subjektive Wahrscheinlichkeit spielt vor allem in der “radikaleren” Variante der Bayes-Statistik eine Rolle, siehe Abschnitt 3.5. Auch der subjektive Wahrscheinlichkeitsbegriff erfüllt die Axiome von Kolmogoroff (siehe 3.1).

Subjektive Wahrscheinlichkeit: Die subjektive Wahrscheinlichkeit $\Pr(A)$ ist ein Mass für die persönliche Überzeugung und entspricht dem Wettquotienten für das Eintreten von A .

3.3 Bedingte Wahrscheinlichkeit

Definition. Seien A und B Ereignisse und $\Pr(A) > 0$. Die *bedingte Wahrscheinlichkeit* von B unter der Bedingung, dass A eintritt (*gegeben A*) wird definiert durch

$$\boxed{\Pr(B | A) = \frac{\Pr(B \cap A)}{\Pr(A)}}. \quad (3.3.1)$$

(3.3.1) kann man mit der Abbildung 3.1 nachvollziehen. $\Pr(B \cap A)$ stellt das Verhältnis der Anzahl Elemente in $B \cap A$ zur Anzahl Elemente in Ω dar. Wir sehen im der Abbildung, dass $\Pr(B | A) = \frac{|B \cap A|}{|A|}$, da $\Pr(B \cap A) = \frac{|B \cap A|}{|\Omega|}$ und $\Pr(A) = \frac{|A|}{|\Omega|}$.

Beispiel. Beim Würfeln seien die Ereignisse

- $A = \{\text{gerade Augenzahl}\} = \{2, 4, 6\}$,
- $B = \{\text{Augenzahl} > 3\} = \{4, 5, 6\}$.

Meine Kollegin würfelt und sagt mir, dass A eingetreten sei. Mit dieser zusätzlichen Information berechne ich dann $\Pr(B | A)$, die bedingte Wahrscheinlichkeit, dass auch B eingetreten ist. Dann ist $A \cap B = \{4, 6\}$ und $\Pr(B | A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{2/6}{3/6} = \frac{2}{3}$.

In der Regel ist $\Pr(B | A) \neq \Pr(B)$, so ist im obigen Beispiel $\Pr(B) = 1/2$. Bedingte Wahrscheinlichkeiten $\Pr(\cdot | A)$ können als Wahrscheinlichkeiten auf einem neuen Ergebnisraum $\Omega^* = A$ aufgefasst werden.

Multiplikationsregel. Durch Umformen von 3.3.1 erhalten wir die sogenannte *Multiplikationsregel*. Diese behandelt die Wahrscheinlichkeit von Schnittmengen oder der *gemeinsamen* Wahrscheinlichkeit von Ereignissen.

$$\Pr(A \cap B) = \Pr(A | B) \cdot \Pr(B) \quad (3.3.2)$$

oder

$$\Pr(A \cap B) = \Pr(B | A) \cdot \Pr(A). \quad (3.3.3)$$

Mit Hilfe der Multiplikationsregel kann man oft Wahrscheinlichkeiten auf einfache Weise in mehreren Schritten berechnen. Die gemeinsame Wahrscheinlichkeit für mehrere Ereignisse $A_i, i = 1, \dots, n$ ist dann

$$\Pr(A_1 \cap A_2 \cap A_3 \dots \cap A_n) = \Pr(A_1) \Pr(A_2 | A_1) \Pr(A_3 | A_1, A_2) \dots \Pr(A_n | A_1, \dots, A_{n-1}) \quad (3.3.4)$$

Die allgemeine Formal ist dann (für Interessierte*)

$$\Pr\left(\bigcap_{k=1}^n A_k\right) = \prod_{k=1}^n \Pr\left(A_k | \bigcap_{j=1}^{k-1} A_j\right). \quad (3.3.5)$$

Schauen wir uns einige Folgerungen von (3.3.1) an:

- Wenn A und B disjunkt sind, dann $\Pr(A | B) = 0 / \Pr(B) = 0$.
- Wenn $A \subset B$ (A ist eine Untermenge von B), dann $\Pr(A | B) = \Pr(A) / \Pr(B) < 1$, d.h., “ B ist eine nötige, aber nicht hinreichende Bedingung für A ” ($B \not\Rightarrow A$), siehe Abbildung 3.5. Beispiel: A : “Kraft > 100 Newton”, B : “Kraft > 50 Newton”.

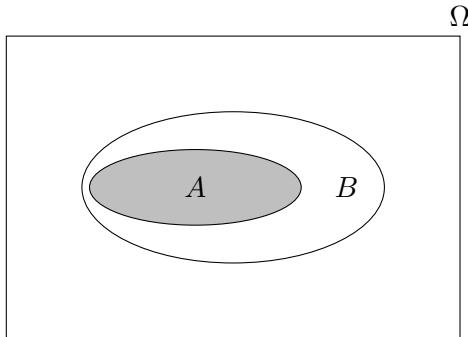


Abbildung 3.5: Ereignis $A \subset B$, $\Pr(A | B) < 1$, B impliziert nicht A .

- Wenn $B \subset A$, dann $\Pr(A | B) = \Pr(B) / \Pr(B) = 1$, d.h., A wird impliziert von B (B ist hinreichend für A) ($B \Rightarrow A$), siehe Abbildung 3.6. Beispiel: A : “ROM Knieflexion > 30 Grad”, B : “ROM Knieflexion > 90 Grad”.

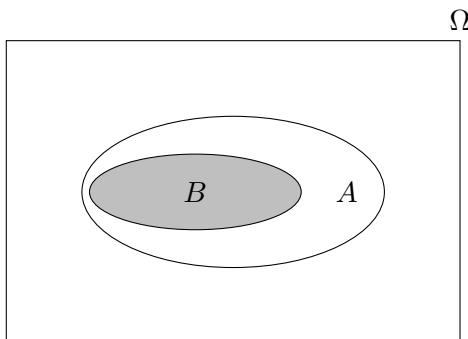


Abbildung 3.6: Ereignis $B \subset A$, $\Pr(A | B) = 1$, aus B folgt A .

- Die gemeinsame Wahrscheinlichkeit vom Ereignis $A \cap B$ kann nicht grösser sein als die Wahrscheinlichkeit von einem Ereignis A oder B . Bezogen auf Aussagen: *Eine Aussage kann nicht wahrscheinlicher sein als eine ihrer Folgerungen*. Beispiel: $A \cap B$: “Peter ist Sportler und seine Kraft ist grösser als 100 Newton”, A : “Peter ist Sportler”, B : “Kraft grösser als 100 Newton”.

Totale Wahrscheinlichkeit. Sei C_1, \dots, C_n eine Zerlegung (Partition) von Ω (in paarweise disjunkte Ereignisse, siehe Abbildung 3.7). Für beliebige Ereignisse A gilt dann

$$\Pr(A) = \sum_{i=1}^n \Pr(A \cap C_i) = \sum_{i=1}^n \Pr(A | C_i) \Pr(C_i). \quad (3.3.6)$$

Der Nutzen von 3.3.6 ist, dass manchmal die Berechnung von $\Pr(C_i)$ und $\Pr(A | C_i)$ einfacher ist als die direkte Berechnung von $\Pr(A)$.

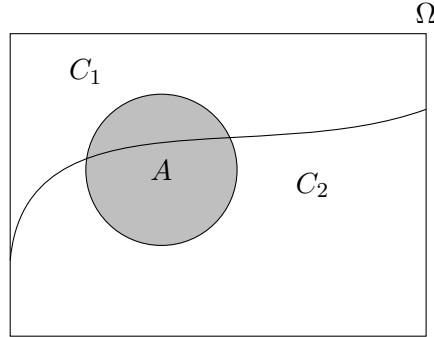


Abbildung 3.7: $\Pr(A) = \Pr(A \cap C_1) + \Pr(A \cap C_2) = \Pr(A | C_1) \Pr(C_1) + \Pr(A | C_2) \Pr(C_2)$

Beispiel. Bei einer Krankheitsdiagnose sind die folgenden Angaben bekannt:

- In der Bevölkerung sind 0.1% krank (*Prävalenz*).
- Von den kranken Personen werden 90% durch die Untersuchung entdeckt (*Sensitivität* des Tests).
- Von den gesunden Personen werden 99% durch die Untersuchung als gesund eingestuft (*Spezifität* des Tests).

Nun wird eine Person aus der Bevölkerung herausgegriffen, untersucht und als krank eingestuft. Wie wahrscheinlich ist es, dass das stimmt?

Sei A das Ereignis, dass eine zufällig gewählte Person krank ist, und B das Ereignis, dass die Untersuchung einer zufällig ausgewählten Person die Diagnose “krank” ergibt. Aus den Angaben haben wir

$$\begin{aligned} \Pr(A) &= 0.001, & \text{also } \Pr(A^C) &= 0.999 \\ \Pr(B | A) &= 0.9, & \text{also } \Pr(B^C | A) &= 0.1 \\ \Pr(B^C | A^C) &= 0.99, & \text{also } \Pr(B | A^C) &= 0.01 \end{aligned}$$

Die Wahrscheinlichkeiten $\Pr(B)$ und $\Pr(B^C)$ kennen wir (noch) nicht. Gesucht ist nun $\Pr(A | B)$, also die bedingte Wahrscheinlichkeit, dass unsere Person krank ist, gegeben,

dass sie als krank eingestuft wurde. Nach der Multiplikationsregel (3.3.2) gilt

$$\Pr(A \cap B) = \Pr(B | A) \Pr(A) = 0.9 \times 0.001 = 0.0009$$

und nach dem Satz der totalen Wahrscheinlichkeit (3.3.6)

$$\begin{aligned}\Pr(B) &= \Pr(B | A) \Pr(A) + \Pr(B | A^C) \Pr(A^C) \\ &= 0.9 \times 0.001 + 0.01 \times 0.999 \\ &= 0.01089\end{aligned}$$

Damit ist

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0.0009}{0.01089} = 0.0826.$$

also sind 8.3% aller als krank diagnostizierten Personen tatsächlich krank. Eine Diagnose ist bei weitem nicht unfehlbar!

Verbessert man im obigen Beispiel die Diagnostik zu $\Pr(B | A^C) = 0.001$, d.h. zu weniger Falsch-Positiven Befunden, so ergibt sich analog $\Pr(A | B) = 0.4739$, d.h. die Vertrauenswürdigkeit einer Krank-Diagnose steigt von 8.3% auf 47.4%. Dieses Beispiel illustriert die Problematik der Fehldiagnosen bei eher seltenen Krankheiten.

Im obigen Beispiel haben wir bereits den Satz von *Bayes* benutzt, auf den wir bald zurückkommen (3.5). Das Beispiel ist in der Abbildung als *Baumdiagramm* 3.8 dargestellt.

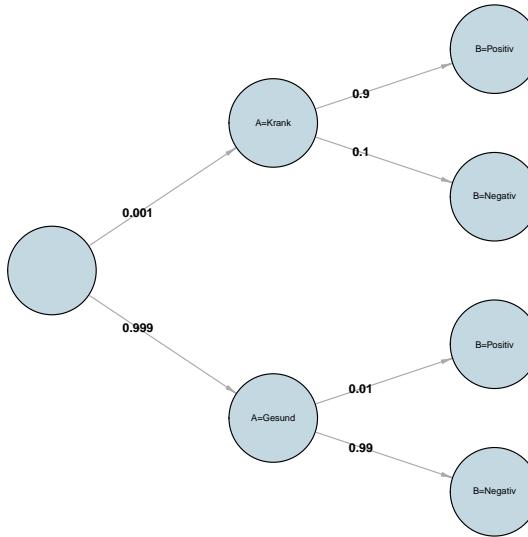


Abbildung 3.8: Baumdiagramm zum Beispiel der Diagnostik

3.4 Unabhängigkeit

Definition. Zwei Ereignisse A und B heißen (stochastisch) unabhängig, falls

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B). \quad (3.4.1)$$

Bei Unabhängigkeit gilt

- $\Pr(A | B) = \Pr(A)$ (wenn $\Pr(B) \neq 0$)
- $\Pr(B | A) = \Pr(B)$ (wenn $\Pr(A) \neq 0$).

Unabhängigkeit bedeutet, dass das Eintreten eines Ereignisses keinen Einfluss hat auf die Wahrscheinlichkeit vom zweiten Ereignis. Bei Unabhängigkeit ist die Wahrscheinlichkeit für das gemeinsame Auftreten von A und B also gleich dem Produkt ihrer Einzelwahrscheinlichkeiten.

Beispiel. Abbildung 3.9 zeigt ein Beispiel für Unabhängigkeit. Sei A wieder das Ereignis, dass eine zufällig gewählte Person krank ist, und B das Ereignis, dass die

Untersuchung einer zufällig ausgewählten Person die Diagnose “krank” ergibt.

- $\Pr(A) = 0.3$
- $\Pr(B) = 0.9 \times 0.3 + 0.9 \times 0.7 = 0.9$
- $\Pr(A \cap B) = 0.3 \times 0.9 = \Pr(A) \cdot \Pr(B) = 0.3 \times 0.9$

Daraus folgt, dass $\Pr(A | B) = \Pr(A)$ und $\Pr(B | A) = \Pr(B)$, d.h. der Test bringt keine neue Information. Das Testresultat ist unabhängig vom Krankheitsstatus und der Krankheitsstatus ist unabhängig vom Testresultat!

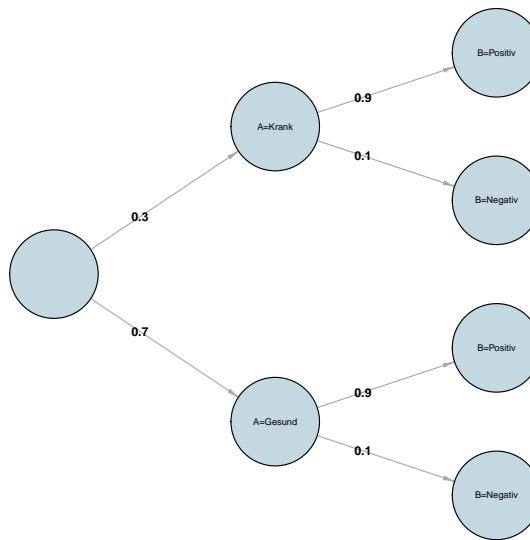


Abbildung 3.9: Unabhängigkeit von Krankheit und Testresultat

Definition. Die Ereignisse A_1, \dots, A_n heissen (stochastisch) unabhängig, wenn für jede endliche Teilfamilie die Produktformel gilt,

$$\Pr\left(\bigcap A_k\right) = \prod \Pr(A_k).$$

Unabhängige versus disjunkte Ereignisse. Wir haben oben gesehen: Wenn A und B disjunkt sind, dann $\Pr(A | B) = 0 / \Pr(B) = 0 \neq \Pr(A)$. Daraus folgt, dass disjunkte Ereignisse *nicht* unabhängig sind.

3.5 Der Satz von Bayes

In der sogenannten Bayes-Statistik braucht man z.T. einen subjektiven Wahrscheinlichkeitsbegriff, zumindest in der radikalen Variante des Bayesianismus. Wir wollen in diesem Kapitel nur die Grundlagen für die Bayes-Statistik betrachten.

Die bedingte Wahrscheinlichkeit $\Pr(A | B)$ haben wir schon kennengelernt. Diese Wahrscheinlichkeit ist die *bedingte* Wahrscheinlichkeit des Ereignisses A gegeben Ereignis B .

Betrachten wir nun ein bekanntes Gebiet mit bedingten Wahrscheinlichkeiten, nämlich das Gebiet der *Diagnostik*.

In der Diagnostik reden wir z.B. von der Wahrscheinlichkeit einer *Hypothese* H (z.B. Gegenwart von Krankheit) gegeben *Daten* E oder gegeben *Test* T . Dazu schauen wir uns das *Problem des Umkehrschlusses* an. Unter diesem Problem versteht man das irrtümliche Gleichsetzen der beiden bedingten Wahrscheinlichkeiten $\Pr(H | E)$ und $\Pr(E | H)$. Diese Wahrscheinlichkeiten sind aber i.A. nicht identisch, können aber durch das *Theorem von Bayes* verbunden werden. Dieses folgt direkt aus (3.3.2) und (3.3.3) und lautet

$$\boxed{\Pr(H | E) = \frac{\Pr(E | H) \Pr(H)}{\Pr(E)}}. \quad (3.5.1)$$

Das Theorem ist benannt nach dem englischen Mathematiker und Pfarrer THOMAS BAYES, der diesen Satz erstmals 1763 in einer posthum veröffentlichten Arbeit beschrieb [4]. In der wichtigen Gleichung (3.5.1) stellt $\Pr(H | E)$ die *a posteriori*-Wahrscheinlichkeit von H dar, $\Pr(H)$ stellt die *a priori*-Wahrscheinlichkeit (z.T. subjektive Überzeugung) von H dar (die Wahrscheinlichkeit der Hypothese, *bevor* man im Besitz von Daten ist). $\Pr(H)$ ist im besten Fall eine gute Schätzung der *Prävalenz* einer Krankheit, wenn z.B. H für “krank” steht. H^C bedeutet das Komplement von H , “gesund”.

Aufgrund der totalen Wahrscheinlichkeit (3.3.6) können wir (3.5.1) schreiben als

$$\boxed{\Pr(H | E) = \frac{\Pr(E | H) \Pr(H)}{\Pr(E | H) \Pr(H) + \Pr(E | H^C) \Pr(H^C)}}. \quad (3.5.2)$$

Definition. Die *Likelihood* einer Hypothese, $L(H)$, ist eine Funktion der Hypothese und stellt die Wahrscheinlichkeit der beobachteten Daten unter der Hypothese dar,

$$L(H) = \Pr(E | H). \quad (3.5.3)$$

Die Likelihood ist also ein Funktion des zweiten Arguments in $\Pr(E | H)$. Die Daten wurden beobachtet und die Likelihood der Hypothese ist gleich der Wahrscheinlichkeit der beobachteten Daten unter dieser Hypothese.

Bei zwei komplementären Hypothesen H und H^C ist

- die Likelihood von H : $L(H) = \Pr(E | H)$
- die Likelihood von H^C : $L(H^C) = \Pr(E | H^C)$

Teilen wir die a posteriori-Wahrscheinlichkeit von H in (3.5.1) durch diejenige von H^C . Dann kann man (3.5.1) auch in der *Odds-Form* schreiben.

$$\underbrace{\frac{\Pr(H | E)}{\Pr(H^C | E)}}_{\text{Posterior Odds}} = \underbrace{\frac{\Pr(E | H)}{\Pr(E | H^C)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{\Pr(H)}{\Pr(H^C)}}_{\text{Prior Odds}}. \quad (3.5.4)$$

Wir können das kurz schreiben, mit LR als der *Likelihood Ratio*,

$$Odds(H | E) = LR \times Odds(H). \quad (3.5.5)$$

Die a posteriori-Chance ist also das Produkt aus den a priori-Odds und dem durch die Daten bestimmten LR . Oder: Die LR transformiert die a priori-Chance von H zur a posteriori-Chance von H .

Wenn wir noch auf beiden Seiten den Logarithmus nehmen, dann können wir (3.5.5) additiv schreiben,

$$\boxed{\log Odds(H | E) = \log LR + \log Odds(H)}. \quad (3.5.6)$$

Die **philosophische Konsequenz** ist:

Erkenntnis in der Wissenschaft: A Posteriori = A Priori + empirische Daten

Betrachten wir als Spezialfall für die empirischen Daten E einen Test T ($T = 1$ für positives und $T = 0$ für ein negatives Resultat) sowie einer Hypothese der Krankheit ($H = 1$ für Krankheit versus $H = 0$ für Gesundheit). Dann sind folgende aus der Diagnostik bekannten Größen wichtig:

- $\Pr(T = 1 | H = 1)$: Sn des Tests (Richtig-Positiv-Rate, TPR)
- $\Pr(T = 0 | H = 0)$: Sp des Tests (Richtig-Negativ-Rate, TNR)
- $\Pr(T = 0 | H = 1)$: $1-Sn$ (Falsch-Negativ-Rate, FNR)
- $\Pr(T = 1 | H = 0)$: $1-Sp$ (Falsch-Positiv-Rate, FPR)
- $\Pr(H = 1 | T = 1)$: PPV des Tests
- $\Pr(H = 0 | T = 0)$: NPV des Tests
- $LR+ = \frac{L(H=1)}{L(H=0)} = \frac{\Pr(T=1|H=1)}{\Pr(T=1|H=0)} = \frac{TPR}{FPR} = \frac{Sn}{1-Sp}$: LR bei positivem Test
- $LR- = \frac{L(H=0)}{L(H=1)} = \frac{\Pr(T=0|H=1)}{\Pr(T=0|H=0)} = \frac{FNR}{TNR} = \frac{1-Sn}{Sp}$: LR bei negativem Test

Abkürzungen: Sn : Sensitivität, Sp : Spezifität, PPV : Positiv prädiktiver Wert, NPV : Negativ prädiktiver Wert, LR : Likelihood ratio.

Das Problem des Umkehrschlusses ist vor allem bei *seltenen Krankheiten* relevant. Ist die Prävalenz $\Pr(H = 1)$ einer Krankheit klein, dann kann $\Pr(H = 1 | T = 1)$ z.T. (viel) kleiner werden als $\Pr(T = 1 | H = 1)$.

Beispiel 1. Die Tabelle 3.2 zeigt die Resultate einer Studie, die die Sensitivität Sn und die Spezifität Sp , also die Gütekriterien bestimmen wollte für einen neuen Test für die Diagnose einer Krankheit. Mit den Daten dieser Studie können die Gütekriterien, prädiktiven Werte und Likelihood Ratios geschätzt werden: $Sn = 0.9$, $Sp = 0.9$.

		Wahrheit		
		$H = 1$	$H = 0$	Total
Test	$T = 1$	90	100	190
	$T = 0$	10	900	910
Total		100	1000	1100

Tabelle 3.1: Diagnostik bei einer seltenen Krankheit, $H = 1$: Krank, $H = 0$: Gesund, $T = 1$: Test positiv, $T = 0$: Test negativ. Die Berechnung der Gütekriterien ergibt $Sn = 0.9$ und $Sp = 0.9$.

Brauchen wir als (a priori) Prävalenz die beobachtete Prävalenz ($100/1100 = 0.091$), dann ist der $PPV = 0.474$ und der $NPV = 0.989$. Die Sensitivität Sn ist also viel grösser als der prädiktive Wert! Für den Patienten ist in diesem Fall ein positives Testresultat weniger dramatisch als man nach Betrachtung der hohen Sn und Sp meinen könnte.

Dieser Test eignet sich besser für den Ausschluss (weil NPV gross) und weniger für den Einschluss von Krankheit (PPV nicht gross). Für die LR 's ergibt sich $LR+ = 0.9/0.1 = 9$ und $LR- = 0.1/0.9 = 1/9$, die Wahrscheinlichkeit eines positiven Tests ist bei einem Kranken 9-mal grösser als bei einem Gesunden, und die Wahrscheinlichkeit eines negativen Tests ist bei einem Kranken 9-mal kleiner als bei einem Gesunden. (Allgemein ist natürlich $LR+$ nicht der Kehrwert von $LR-$).

Abbildung 3.10 zeigt für die Sn und Sp in diesem Beispiel, wie sich die Prävalenz auf den PPV und NPV auswirkt. Gestrichelt ist die beobachtete Prävalenz aus den erhobenen Daten.

```
p0obs <- 100/1100 #observed prevalence
p0 <- seq(0, by = 0.01, 1) #prevalences
Sn <- 0.9
Sp <- 0.9 #Sn and Sp
PPV <- (p0 * Sn)/(p0 * Sn + (1 - p0) * (1 - Sp)) #pos.pred. value
NPV <- (1 - p0) * Sp/((1 - p0) * Sp + p0 * (1 - Sn)) #neg. pred. value
plot(p0, PPV, xlab = "prevalence", ylab = "PPV", type = "l", ylim = c(0, 1))
abline(v = p0obs, h = (p0obs * Sn)/(p0obs * Sn + (1 - p0obs) * (1 - Sp)), lty = 2)
plot(p0, NPV, xlab = "prevalence", ylab = "NPV", type = "l", ylim = c(0, 1))
abline(v = p0obs, h = (1 - p0obs) * Sp/((1 - p0obs) * Sp + p0obs * (1 - Sn)), lty = 2)
```

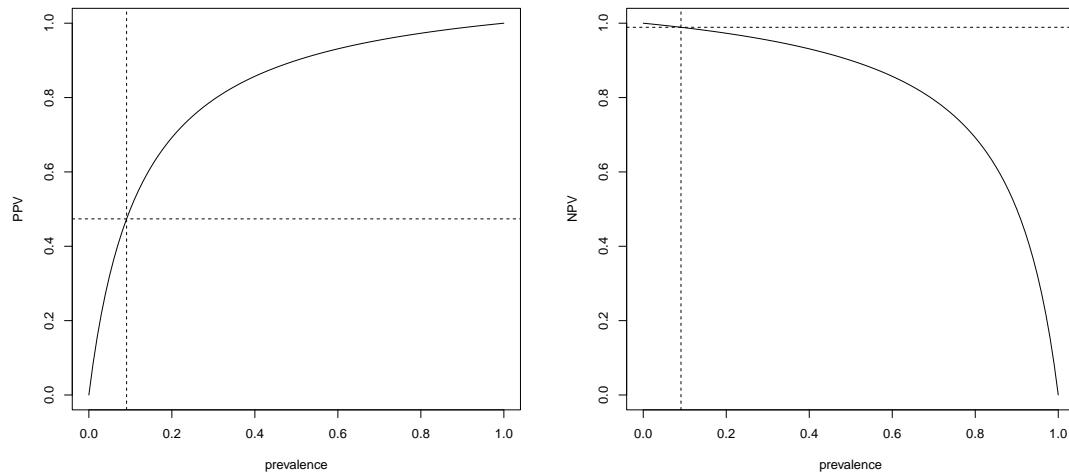


Abbildung 3.10: PPV und NPV als Funktion der Prävalenz, bei Test mit $Sn=0.9$, $Sp=0.9$. Gestrichelt: Beobachtete Prävalenz.

Bei niederprävalenten Merkmalen muss die Spezifität praktisch maximal sein (wenig Falsch-Positive), damit der positiv prädiktive Wert nicht absinkt. Denn schon wenige Falsch-Positive können den positiv prädiktiven Wert stark vermindern.

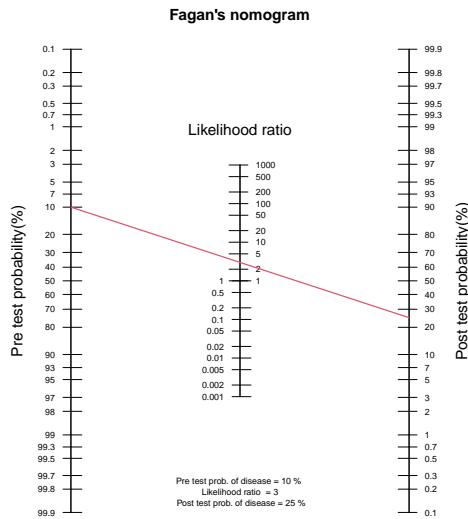
Will man also mit einem Test eine Krankheit (oder ein Merkmal wie ‘‘Blockade des Iliosakralgelenks’’) einschliessen, muss man spezifisch sein, was gleichbedeutend ist mit wenig Falsch-Positiven und einem grossen positiven prädiktiven Wert. Ist dieses Merkmal zudem noch *selten*, muss die Spezifität fast maximal sein. Will man hingegen Krankheit ausschliessen, muss man sensitiv sein (wenig Falsch-Negative), ist zudem die Krankheit *häufig*, muss die Sensitivität fast maximal sein.

Beispiel 2. Im folgenden Beispiel veranschaulichen wir Bayes mit einem *Rechenschieber*, wie man ihn auch früher physisch zum Rechnen brauchte. Dieser basiert auf der additiven Bayes-Formel 3.5.6.

Wir nehmen eine Vortestwahrscheinlichkeit von Krankheit $\Pr(H = 1) = 0.1$ an, dies entspricht also einer prior Odds von 0.111. Zudem haben wir einen diagnostischen Test mit $Sn = \Pr(T = 1 | H = 1) = 0.3$ und $Sp = \Pr(T = 0 | H = 0) = 0.9$.

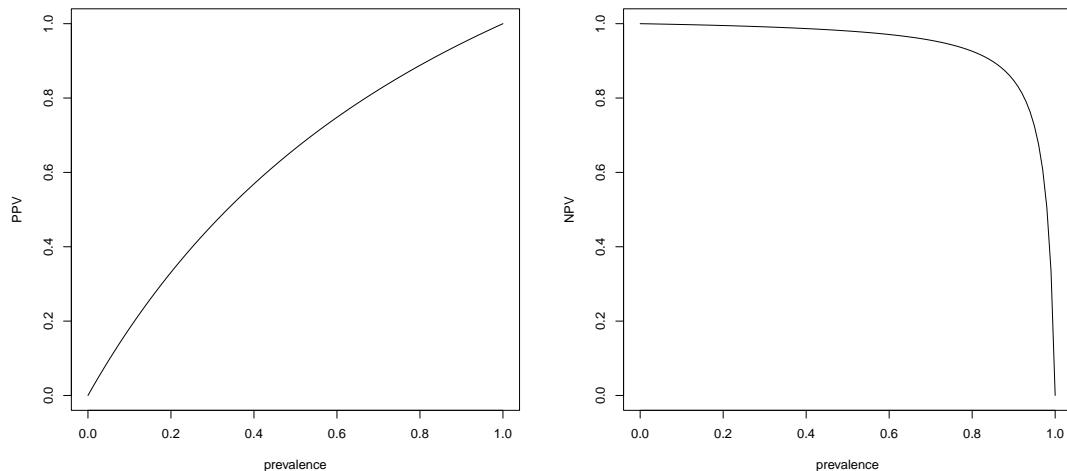
Die Likelihood ratio ($H = 1$ versus $H = 0$) nach Beobachtung eines positiven Tests (definiert als das Verhältnis der Wahrscheinlichkeit von einem positiven Test unter $H = 1$ zur Wahrscheinlichkeit eines positiven Tests unter $H = 0$) ist $LR+ = Sn/(1 - Sp) = 3$. Die resultierende Nachtestwahrscheinlichkeit ist $\Pr(H = 1 | T = 1) = 0.25$. Der Leser versuche, dies anhand (3.5.4) nachzurechnen.

```
## install.packages('TeachingDemos') ## auskommentieren für Installation der package, dann Zeile
## löschen!
library(TeachingDemos)    ## package laden
p0 <- 0.1
sn <- 0.3
sp <- 0.9
fagan.plot(probs.pre.test = p0, LR = sn/(1 - sp))
```



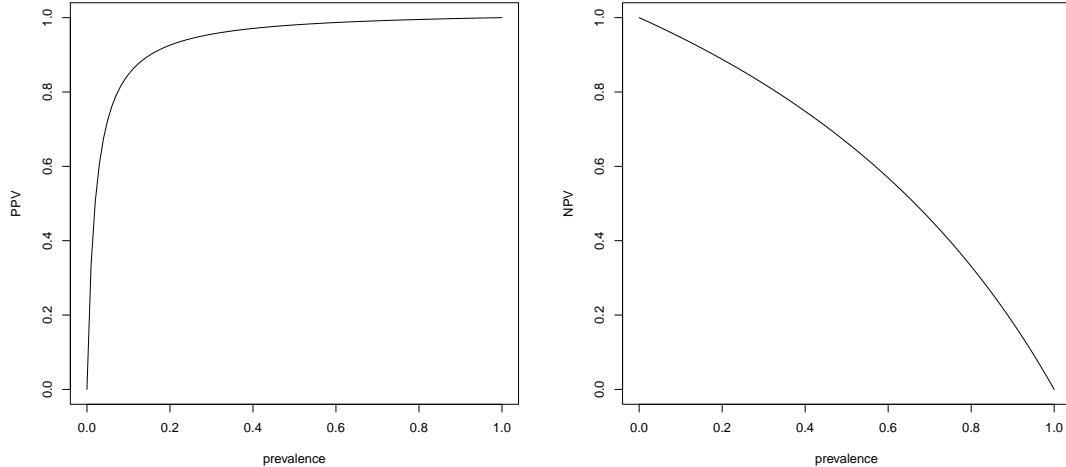
Mit `plotFagan2()` aus derselben package können die Vortestwahrscheinlichkeit, Sensitivität und Spezifität auch interaktiv eingegeben werden. Probiere es aus!

Visualisierung der Abhängigkeit von der Prävalenz bei unterschiedlichen Sn und Sp .
 Beispiel 1: Prädiktive Werte bei hochsensitivem Test mit $Sn = 0.99$ und $Sp = 0.5$:



Test ist geeignet für das *Ausschliessen* von Krankheit, dazu brauchen wir eher einen sensitiven Test: “*SnNout*”, bei hoher Sn , wenn neg., dann ausschliessen

Beispiel 2: Prädiktive Werte bei hochspezifischem Test mit $Sn = 0.5$ und $Sp = 0.99$:



Test ist geeignet für das *Einschliessen* von Krankheit, dazu brauchen wir eher einen spezifischen Test: “*SpPin*”, bei hoher Sp , wenn pos., dann einschliessen.

Spezialfall: Gleichheit von a priori und a posteriori-Wahrscheinlichkeit. Wann sind diese beiden Wahrscheinlichkeiten gleich?

Aus (3.5.1) folgt: Damit $\Pr(H | T) = \Pr(H)$ gilt, muss $\Pr(T | H) = \Pr(T)$ gelten, d.h. die Wahrscheinlichkeit von T ist *unabhängig* von H . Das ist gleichbedeutend mit $LR = 1$ oder $\log LR = 0$:

- Positiver Test: $LR+ = \frac{\Pr(T=1|H=1)}{\Pr(T=1|H=0)} = \frac{TPR}{FPR} = 1 \Leftrightarrow TPR = FPR$.
- Negativer Test: $LR- = \frac{\Pr(T=0|H=1)}{\Pr(T=0|H=0)} = \frac{FNR}{TNR} = 1 \Leftrightarrow FNR = TNR$.

Spezialfall: Gleichheit von bedingten Wahrscheinlichkeiten. Wir haben gesehen, dass im Allgemeinen

$$\Pr(H | T) \neq \Pr(T | H).$$

Aus (3.5.1) folgt: Damit wir T und H in den bedingten Wahrscheinlichkeiten vertauschen dürfen, muss

$$\Pr(T) = \Pr(H)$$

gelten, wir haben dann *Symmetrie* bezüglich T und H . Das ist bei der Schätzung dann der Fall, wenn die Häufigkeiten symmetrisch sind, d.h. wenn wir gleich viele Falsch-Positive und Falsch-Negative Ereignisse haben. Das wäre der Fall bei der folgenden symmetrischen Vierfeldertafel (symmetrisch bezüglich der Hauptdiagonalen):

		Wahrheit		Total
		$H = 1$	$H = 0$	
Empirie	$T = 1$	a	b	a+b
	$T = 0$	b	d	b+d
Total		a+b	b+d	a+2b+d

Tabelle 3.2: Symmetrie von T und H : $\Pr(H | T) = \Pr(T | H)$.

3.6 Bayesianische Statistik*

Der Gebrauch vom Bayes-Theorem in der Diagnostik ist etabliert als *formalisierte klinische Argumentation*. Kontroverser ist die *Bayesianische Statistik*.

Wir werden später sehen, dass statistisches Schätzen und Testen eine Analogie hat zur Diagnostik.

Bei diagnostischen Tests ist die Hypothese wie oben z.B. “*Patient ist krank*”. In der schliessenden Statistik werden Hypothesen allgemeinere Aussagen sein, wie etwa Aussagen über einen unbekannten Parameter θ in einer Population, so wie “ $\theta \leq 177$ cm” für die unbekannte mittlere Körpergrösse in einer Population. Wir werden uns beim Schätzen (Kapitel 7) und Testen von Hypothesen (Kapitel 8) aber zunächst auf die klassische Statistik beschränken, auf die Quantifizierung der Wahrscheinlichkeit von empirischen Daten, gegeben eine Hypothese, also eigentlich auf obige Likelihood (3.5.3).

Die umgekehrte Wahrscheinlichkeit, die a posteriori-Wahrscheinlichkeit, ist nur über das Bayes-Theorem zu berechnen. Und damit wir die Wahrscheinlichkeit einer Hypothese berechnen können, brauchen wir grundsätzlich ein a priori. Die sogenannte Bayesianische Statistik braucht den Satz von Bayes in der statistischen Analyse, wo ein Parameter θ eine unbekannte Quantität ist (z.B. wie das obige θ). Die a priori Verteilung $p(\theta)$ muss dann spezifiziert werden. Dieser Schritt kann angesehen werden als eine natürliche Erweiterung auf die subjektive Interpretation von Wahrscheinlichkeit, siehe auch [32].

Nach diesem Einblick in das Bayes-Theorem wollen wir daher mit folgendem für jeden Praktiker heilsamen Satz schliessen:

Bayes: Ohne Subjektivität gibt es – strenggenommen – keine Erkenntnis.

Kapitel 4

Zufallsvariablen und ihre Verteilung

Meistens lässt sich ein *Zufallsexperiment* beschreiben durch eine Abbildung auf einem Grundraum Ω und durch die Wahrscheinlichkeiten, mit denen diese Abbildung die Werte im Wertebereich annimmt. Eine solche Abbildung nennt man *Zufallsvariable*, und die Wahrscheinlichkeiten werden durch die *Verteilung* der Zufallsvariable beschrieben.

4.1 Zufallsvariable

Definition. Eine *Zufallsvariable* ist eine Abbildung vom Ergebnisraum Ω auf die reellen Zahlen \mathbb{R} ,

$$X : \Omega \rightarrow \mathbb{R} \quad (4.1.1)$$

Eine Zufallsvariable stellt eine *unbekannte* Quantität dar, die einen Wert in einer Menge von Werten (Zahlen) einnehmen kann. Wir schreiben Zufallsvariablen mit grossen lateinischen Buchstaben, X, Y, \dots , vor deren Beobachtung. Nach Beobachtung brauchen wir kleine Buchstaben x, y, \dots für den spezifischen beobachteten Wert.

Kumulative Verteilungsfunktion. Die *kumulative Verteilungsfunktion* von X ist die Abbildung

$$P : \mathbb{R} \rightarrow [0, 1] \quad P(x) = \Pr(X \leq x). \quad (4.1.2)$$

Dabei ist $P(x) = \Pr(X \leq x) = \Pr(\{\omega \mid X(\omega) \leq x\})$, mit

- $\Pr(\{\omega \mid X(\omega) \leq x\})$: “Die Wahrscheinlichkeit des Ereignisses $\{\omega \mid X(\omega) \leq x\}$ ”.
- $\{\omega \mid X(\omega) \leq x\}$: “Die Menge aller Ergebnisse, für die die Zufallsvariable einen Wert kleiner gleich x annimmt”.

Bemerkung.* Bedingung an eine Zufallsvariable: Die Menge $\{X \leq x\} = \{\omega \mid X(\omega) \leq x\}$ muss für jedes x ein beobachtbares Ereignis, also in \mathcal{F} sein.

Beispiel 1. Wir betrachten die “Heilung” nach Behandlung von drei Patienten. Der Ergebnisraum ist $\Omega = \{GGG, GGK, GKK, GKG, KKK, KGG, KKG, KGK\}$. Ein

Beispiel für eine Zufallsvariable wäre die *Anzahl von Heilungen*.

ω	GGG	GGK	GKK	GKG	KKK	KGG	KKG	KGK
$X(\omega)$	3	2	1	2	0	2	1	1

Beispiel 2. Wir behandeln einen Patienten solange, bis er “geheilt” ist. Der Grundraum wäre dann $\Omega = \{G, KG, KKG, KKKG, KKKKG, \dots\}$. Die *Anzahl von Therapien bis zur Heilung* wäre dann eine mögliche Zufallsvariable:

ω	G	KG	KKG	KKKG	KKKG	KKKKKG
$X(\omega)$	1	2	3	4	5	6

In beiden Beispielen haben wir eine *diskrete* Zufallsvariable. Diskrete Zufallsvariable haben eine *abzählbare* Anzahl von Werten.

Wahrscheinlichkeitsmasse. Sei X eine diskrete Zufallsvariable¹. Dann ist die *Wahrscheinlichkeitsmasse* oder *Gewichtsfunktion*

$$p : \mathbb{R} \rightarrow [0, 1] \quad p(x) = \Pr(X = x). \quad (4.1.3)$$

Sie ordnet jedem Wert von X eine Wahrscheinlichkeit $p(x)$ zu. Im Gegensatz zur obigen Verteilungsfunktion geht es hier um die Wahrscheinlichkeit vom Ereignis $X = x$ und nicht vom Ereignis $X \leq x$.

Wahrscheinlichkeitsdichte. Wenn eine Zufallsvariable X mit beliebiger Präzision gemessen werden kann, hat sie in jedem noch so kleinen Intervall beliebig viele Werte, eine solche Variable nennen wir *kontinuierlich* oder *stetig*. Wir können dann nicht mehr von der Wahrscheinlichkeit eines ganz bestimmten Wertes reden. Die Wahrscheinlichkeit, dass z.B. eine Körpergrösse *genau* 170.002019394932... cm ist, ist dann 0, also $\Pr(X = x) = 0$.

Wir können aber die Wahrscheinlichkeit angeben, mit der X in ein infinitesimal² kleines Intervall dx fällt,

$$\Pr(X \in [x, x + dx]) = p(x)dx. \quad (4.1.4)$$

Wahrscheinlichkeitsmasse und Wahrscheinlichkeitsdichte verhalten sich analog zu Masse und Dichte in der Physik. Also steht $p(x)dx$ für eine Wahrscheinlichkeitsmasse (“Dichte mal Volumen”) und die *Wahrscheinlichkeitsdichte* $p(x)$ ist dann

$$p : \mathbb{R} \rightarrow [0, \infty[\quad p(x) = \frac{\Pr(X \in [x, x + dx])}{dx} \quad (4.1.5)$$

¹Mit *abzählbarer* Anzahl von Werten.

²“ins unendlich Kleine gehend”

Dichten sind nicht mehr Zahlen zwischen 0 und 1, sondern Zahlen zwischen 0 und ∞ . Die Dichtefunktion ist also *fast* das Gegenstück zur Gewichtsfunktion für kontinuierliche Variablen. Wir werden **dieselbe Notation** $p(x)$ brauchen für Gewichtsfunktionen von diskreten Zufallsvariable und Dichtefunktionen von kontinuierlichen Zufallsvariable. Für die kumulative Verteilungsfunktion schreiben wir $P(x)$ für diskrete und kontinuierliche Zufallsvariable.

Abbildung 4.1 zeigt verschiedene Typen von Zufallsvariablen und ihre Verteilungen. Drei Verteilungen stellen die Gewichtsfunktion einer *diskreten* Zufallsvariablen dar, die unten rechts stellt die Dichtefunktion einer *kontinuierlichen* Zufallsvariablen dar.

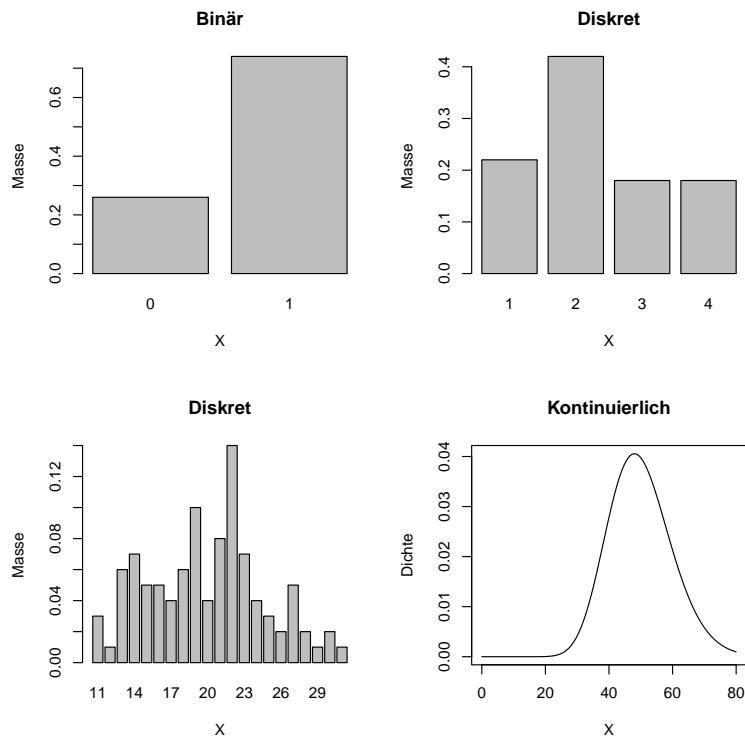


Abbildung 4.1: Typen von Verteilungen

4.2 Verteilungen

Diskrete Zufallsvariable. Abbildung 4.2 zeigt die Gewichtsfunktion $p(x)$ und die Verteilungsfunktion $P(x)$ für zwei diskrete Zufallsvariablen aus Abbildung 4.1. $P(x)$ ist eine Art Treppenfunktion, die an den Stellen x_i um den Wert p_i nach oben springt.

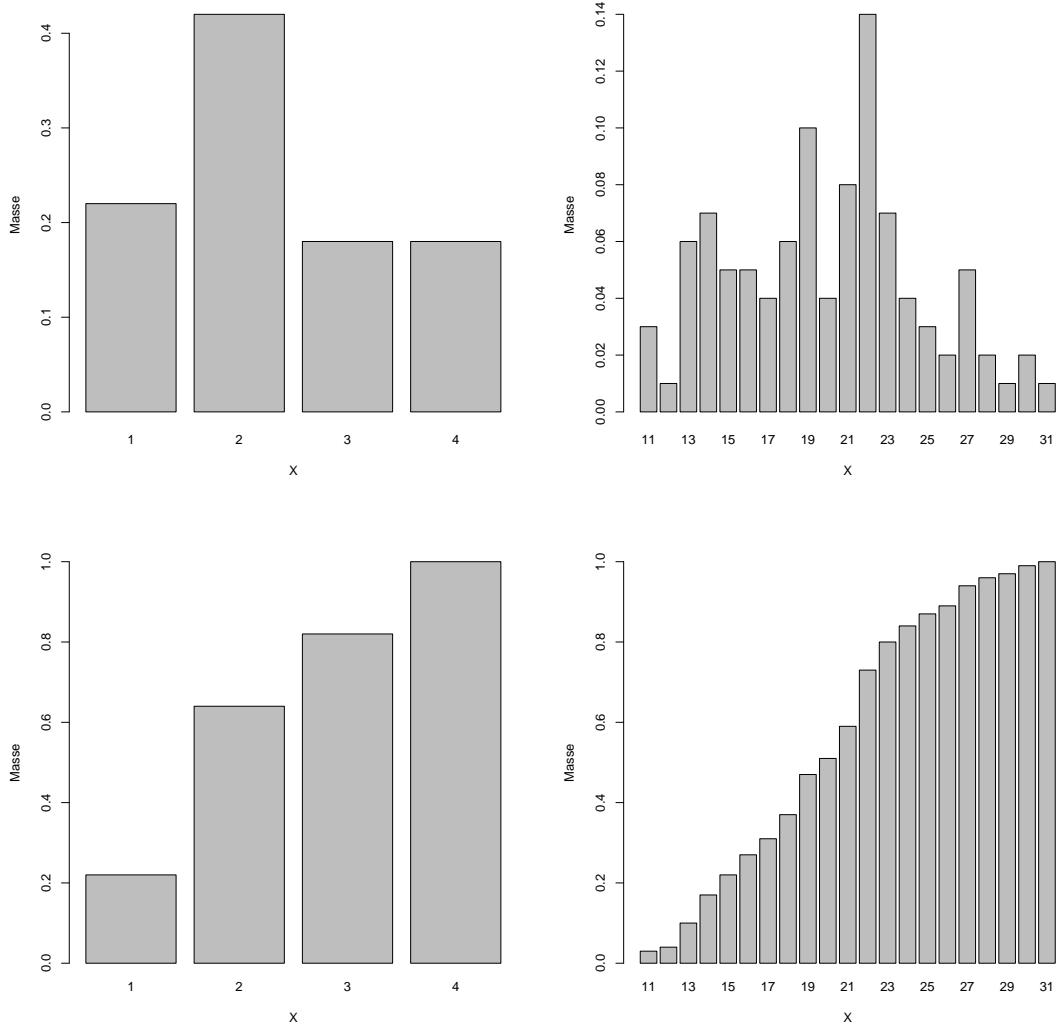


Abbildung 4.2: Diskrete Zufallsvariable. Gewichtsfunktion (oben) und kumulative Verteilungsfunktion (unten)

Stetige Zufallsvariable. Wenn wir eine Zufallsvariable X mit beliebiger Präzision messen können, ist Differential- und Integralrechnung nötig. Im Falle von stetigen Zufallsvariablen ist die Wahrscheinlichkeit für das Ereignis, dass X ein durch a und b begrenztes Intervall trifft, gegeben durch

$$\Pr(a \leq X \leq b) = \int_a^b p(x)dx = P(b) - P(a), \quad (4.2.1)$$

mit $p(x)$ als der oben eingeführten Wahrscheinlichkeitsdichte von X . Die Dichtefunktion wird zwischen den Grenzen a und b “aufsummiert” oder *integriert*. Abbildung 4.3 zeigt eine Dichtefunktion $p(x)$ zusammen mit der Wahrscheinlichkeit für das Ereignis, dass sich X in einem Intervall $[a, b]$ befindet. Wenn das Intervall $[a, b]$ den ganzen Wertebereich von X darstellt, dann ist die blaue Fläche=1. Es sei noch gesagt, dass bei kontinuierlichen Zufallsvariablen $a < X < b$ äquivalent ist mit $a \leq X \leq b$. Die Verteilungsfunktion bei kontinuierlicher Zufallsvariable ist

$$P(x) = \Pr(X < x) = \int_{-\infty}^x p(t)dt. \quad (4.2.2)$$

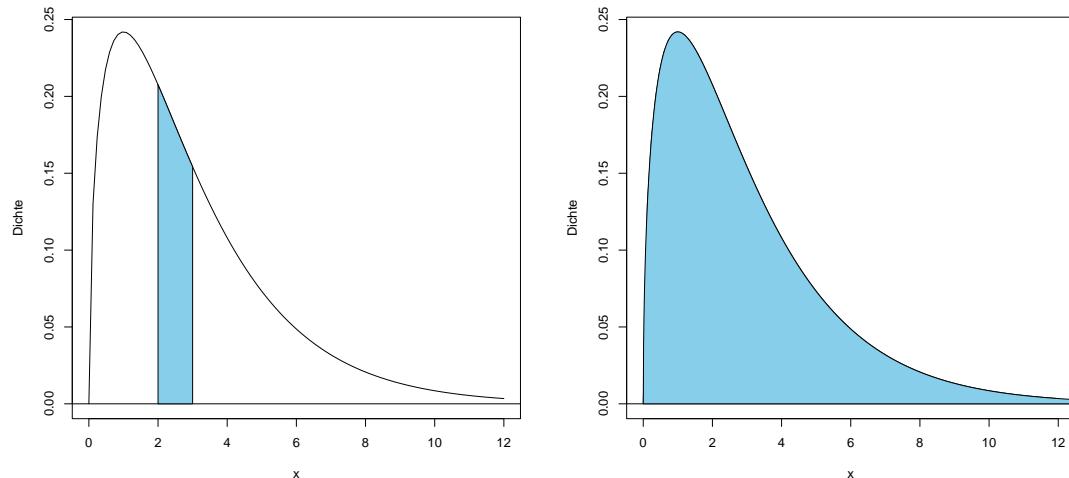


Abbildung 4.3: Kontinuierliche Zufallsvariable. Wahrscheinlichkeit für das Ereignis, dass $a < X < b$

Abbildung 4.4 zeigt die Dichtefunktion $p(x)$ und die Verteilungsfunktion $P(x)$ für eine kontinuierliche Zufallsvariable.

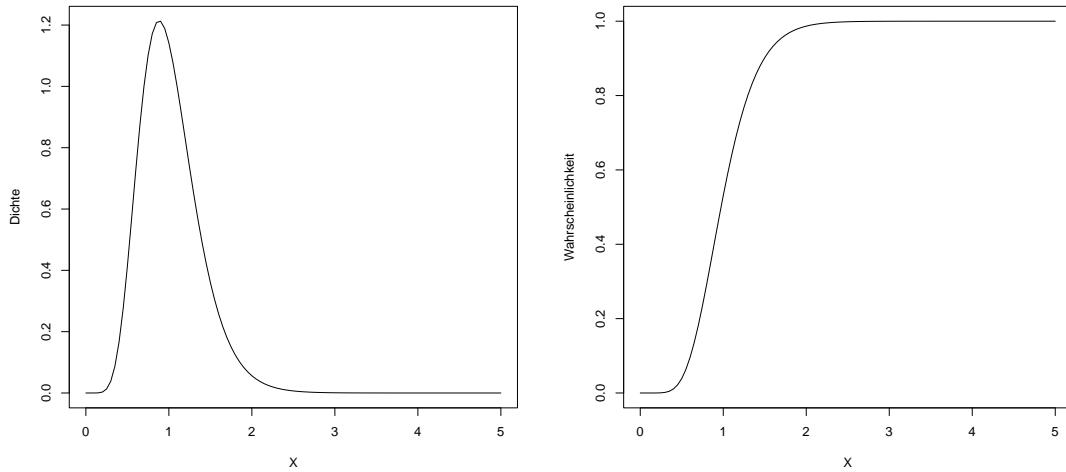


Abbildung 4.4: Kontinuierliche Zufallsvariable. Links: Dichtefunktion $p(x)$. Rechts: Kumulative Verteilungsfunktion $P(x)$

Abbildung 4.5 zeigt den Zusammenhang zwischen Dichtefunktion und kumulativer Verteilungsfunktion (4.2.2) für $X = 5$. Die Fläche unter der Dichtefunktion ist 0.828. Das können wir direkt in der Verteilungsfunktion ablesen.

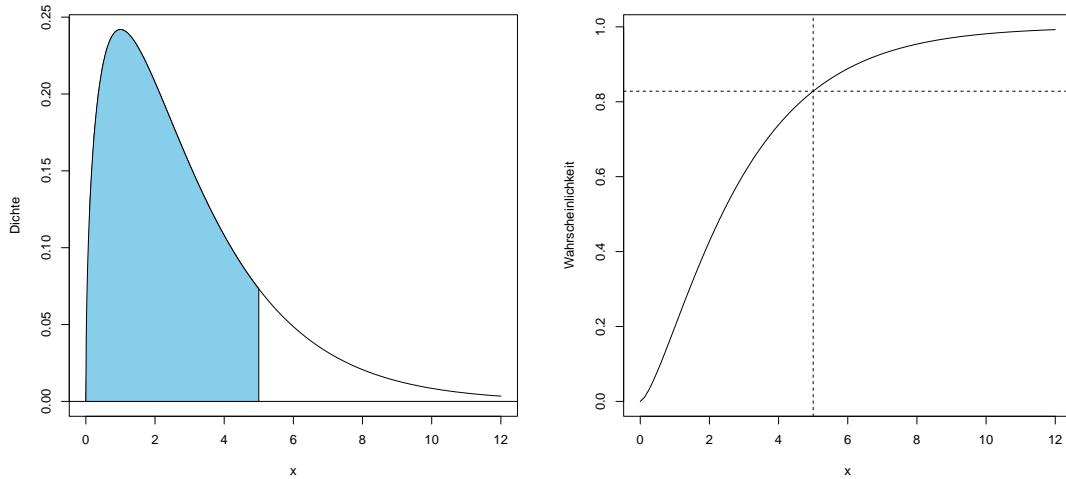


Abbildung 4.5: Kontinuierliche Zufallsvariable. Links: Integrierte Dichte bei $X = 5$. Rechts: Kumulative Verteilungsfunktion bei $X = 5$

Quantilfunktion. Die Umkehrfunktion der Verteilungsfunktion heisst *Quantilfunktion*, Sei $P(x) = \Pr(X \leq x)$ die kumulierte Wahrscheinlichkeit bei $X = x$. Wenn die Verteilung kontinuierlich ist und monoton steigend, dann ist das p -Quantil eindeutig.

$$Q_p = x \in \mathbb{R} \mid \Pr(X \leq x) = p \quad (4.2.3)$$

Besondere Quantile: $Q_{0.25}$ nennt man das erste *Quartil*, $Q_{0.5}$ den *Median* und $Q_{0.75}$ das dritte Quartil. $Q_{0.4}$ nennt man das vierte *Desil*. $Q_{0.06}$ ist das sechste *Perzentil*. Abbildung 4.6 zeigt die Verteilungsfunktion und deren Umkehrfunktion (Quantilfunktion) bei einer kontinuierlichen Zufallsvariable X . Der Wert 0.524 entspricht z.B. der 70. Perzentile.

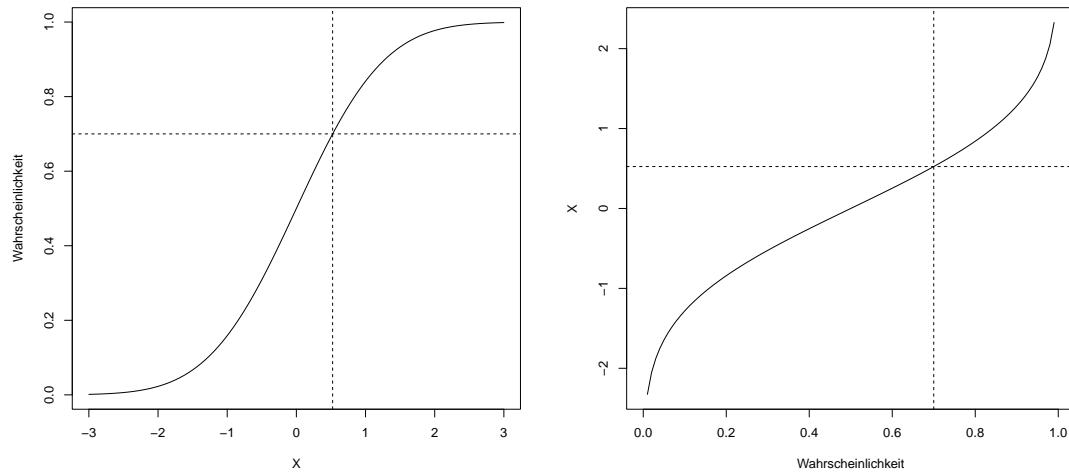


Abbildung 4.6: Links: Verteilungsfunktion. Rechts: Quantilfunktion

Bei diskreten Verteilungen wird es ein bisschen komplizierter³.

4.3 Wichtige Verteilungen

Es gibt – neben den sogenannten nichtparametrischen – viele parametrische Wahrscheinlichkeitsverteilungen. Wir lernen nun im Folgenden drei der wichtigsten Wahrscheinlichkeitsfunktionen kennen, die diskrete *Binomialverteilung*, die diskrete *Poisson-Verteilung* und die stetige *Normalverteilung*.

³Für die Interessierten*. Die Menge aller p -Quantile ist gegeben durch ein untere Grenze $\inf\{x \in \mathbb{R} \mid P(x) \geq p\}$ und einer oberen Grenze $\sup\{x \in \mathbb{R} \mid P(x) \leq p\}$ (*untere Schranke* von X mit kumulierter relativer Häufigkeit grösser gleich p respektive *obere Schranke* von X mit kumulierter relativer Häufigkeit kleiner gleich p). Beispiel Median: $p = 0.5$. Der *Untermedian* ist $\inf\{x \in \mathbb{R} \mid P(x) \geq 0.5\}$. Der *Obermedian* ist $\sup\{x \in \mathbb{R} \mid P(x) \leq 0.5\}$.

4.3.1 Binomialverteilung

Die *Binomialverteilung* ist eine diskrete Verteilung, die aufgrund wahrscheinlichkeitstheoretischer Überlegungen hergeleitet wird. Sie beruht auf dem einfachsten Zufallsexperiment. Wir betrachten n wiederholte Experimente mit dichotomem Ausgang. Zum Beispiel das Ereignis “Therapieerfolg beim i -ten Patienten”. π sei dann die – meist unbekannte – Eintretenswahrscheinlichkeit.

Wir nehmen dabei an, dass die Eintretenswahrscheinlichkeit π für jede Beobachtung gleich ist und dass die Beobachtungen voneinander *unabhängig* sind. (Das entspricht einem “Ziehen mit Zurücklegen”). Eine Zufallsvariable X heisst nun binomialverteilt mit den Parametern π und n , kurz $X \sim \text{Bin}(\pi, n)$, wenn die Gewichtsfunktion gegeben ist durch

$$p(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n. \quad (4.3.1)$$

Dabei ist $\binom{n}{x} = \frac{n!}{(n-x)!x!}$ der *Binomialkoeffizient*. ($k!$ ist die *Fakultät* $k! = 1 \cdot 2 \cdots k$). Dieser stellt die Anzahl der x -elementigen Teilmengen einer n -elementigen Menge dar (`choose(n=k)`). Diese Verteilung entsteht also allgemein, wenn X die Anzahl “Erfolge” in n Wiederholungen repräsentiert.

Abbildung 4.7 zeigt die Verteilung einer binomialverteilten Zufallsvariablen X bei $n = 15$ für 4 verschiedene Parameter π (Eintretenswahrscheinlichkeiten). Siehe dazu Shiny: <https://rstudio.zhaw.ch/rsconnect/content/87>. Dort kannst Du die Parameter des Modells, n und π , einstellen. Versuche, die Abbildung 4.7 nachzuvollziehen.

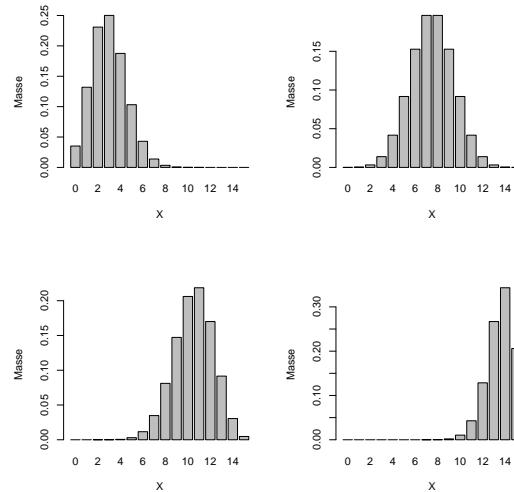


Abbildung 4.7: Binomialverteilung ($n = 15$) mit $\pi = 0.2, 0.5, 0.7$ und $\pi = 0.9$

Wenn $n = 1$, hat man den Spezialfall von einem Bernoulli trial, $X \sim \text{Bin}(\pi, 1)$. Abbildung 4.8 zeigt die Verteilung einer Bernoulli-verteilten Zufallsvariablen X für 4 verschiedene Parameter (Eintretenswahrscheinlichkeiten) π .

$$p(x) = \pi^x (1 - \pi)^{1-x}, \quad x = 0, 1. \quad (4.3.2)$$

Die Summe von n unabhängigen Bernoulli-verteilten Zufallsvariable mit identischem Parameter π ist dann binomialverteilt mit Parametern π und n .

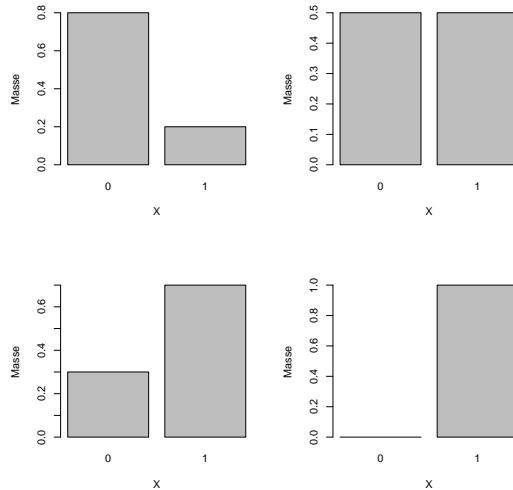


Abbildung 4.8: Bernoulli-Verteilung mit $\pi = 0.2, 0.5, 0.7$ und $\pi = 1$

4.3.2 Poisson-Verteilung

Die *Poisson-Verteilung* mit Parameter λ ist eine diskrete Verteilung mit Gewichtsfunktion

$$p(x) = \frac{\lambda^x}{x!} \exp(-\lambda), \quad x = 0, 1, 2, \dots \quad (4.3.3)$$

Wir schreiben kurz $X \sim \text{Pois}(\lambda)$.

Die Poisson-Verteilung ist nicht auf den Wertebereich $0, 1, 2, \dots, n$ beschränkt. Man erhält die Poisson-Verteilung durch den Grenzübergang aus der Binomialverteilung⁴.

Anzahlen von seltenen Ereignissen werden oft mit einer Poisson-Verteilung modelliert, z.B.

⁴Wenn der Parameter n der Binomialverteilung gross und der zweite Parameter π klein wird, und wenn das Produkt $n\pi$ konstant bleibt, dann ist eine Approximation der Binomialverteilung durch die Poisson-Verteilung mit Parameter $\lambda = n\pi$ möglich.

- die Anzahl Stürze in einer bestimmten Periode
- die Anzahl Erkrankungen in einer bestimmten Periode

Abbildung 4.9 zeigt die Masse von Poissonverteilungen für verschiedene Parameter λ . Siehe dazu: wieder die Abbildung 4.9 mit <https://rstudio.zhaw.ch/rsconnect/content/89>.

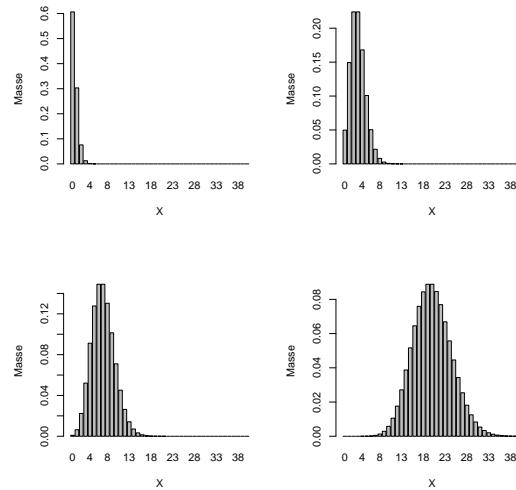


Abbildung 4.9: Poisson Masse mit $\lambda = 0.5, 3, 7, 20$

4.3.3 Normalverteilung

Lassen wir für die Binomialverteilung n gegen unendlich gehen, so wird daraus eine *Normalverteilung*. Wenn n gegen unendlich geht, wenn man also das Zufallsexperiment sehr oft wiederholt, gibt es immer mehr “mögliche Ausprägungsgrade” der Zufallsvariablen, der Wertebereich wird immer “stetiger”. Siehe <https://rstudio.zhaw.ch/rsconnect/content/87>. Lasse n gross werden.

Aus der *diskreten* Binomialverteilung wird dann als Grenzfall eine *stetige* Normalverteilung. So kann z.B. das Merkmal $X = \text{Körpergrösse}$ in einer Population normalverteilt sein. Die Normalverteilung hat zwei Parameter, μ und σ^2 . Wir werden später sehen, dass die Parameter dem sogenannten *Erwartungswert* respektive der sogenannten *Varianz* entsprechen. Man schreibt dann auch

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad (4.3.4)$$

Siehe Shiny: <https://rstudio.zhaw.ch/rsconnect/content/88>. Spiele mit den Parametern und schaue, was passiert.

Die Normalverteilung ist überaus wichtig aus verschiedenen Gründen:

- Empirische Verteilungen haben häufig die Form einer Normalverteilung, viele Merkmale in der Natur sind normalverteilt. Das Grenzwerttheorem (siehe dazu [7.2.1](#)) besagt, dass die Summe einer grossen Anzahl von Zufallsgrössen normalverteilt sind.
- Die Normalverteilung ist ein wichtiges Verteilungsmodell für statistische Kennwerte. So sind z.B. Mittelwerte und viele andere Grössen, die man aus Stichproben von wiederholten Stichprobennahmen berechnet, approximativ normalverteilt.
- Die Normalverteilung ist wichtig in der statistischen Fehlertheorie, z.B. werden Residuen aus Regressionsmodellen als normalverteilt mit Mittelwert 0 angenommen. Wir wissen aus dem Alltag, dass sich viele Zufallseinflüsse gegenseitig annullieren, wenn man sie mittelt.
- Die Normalverteilung ist eine mathematische Basisverteilung für viele andere Verteilungen, insbesondere für die t , die χ^2 und die F -Verteilung.

Eine normalverteilte Variable X mit Parametern μ und σ^2 hat die Dichte

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}. \quad (4.3.5)$$

Standardnormalverteilung. Man kann zeigen, dass wenn X normalverteilt ist mit Parametern μ und σ^2 , also $X \sim \mathcal{N}(\mu, \sigma^2)$, dann ist *standardisierte* Variable $Z = \frac{X-\mu}{\sigma}$ standardnormalverteilt, $Z \sim \mathcal{N}(0, 1)$.

Für $\mu = 0$ und $\sigma^2 = 1$ erhält man also die Standardnormalverteilung mit der Dichte

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right). \quad (4.3.6)$$

Wollen wir die Wahrscheinlichkeit quantifizieren, mit der eine standardnormalverteilte Zufallsvariable X einen Wert x unterschreitet oder Werte zwischen zwei Grössen a und b annimmt, müssen wir die Dichte (Gleichung [4.3.6](#)) *integrieren*, siehe [\(4.2.2\)](#). Dies ergibt die kumulative Verteilungsfunktion der Standardnormalverteilung,

$$\Phi(x) = \Pr(X < x) = \int_{-\infty}^x \phi(t) dt \quad (4.3.7)$$

Die Dichte $p(x)$ und kumulierte Verteilung $P(x)$ einer $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen wird mit $\phi(x)$ respektive $\Phi(x)$ notiert. Diese sind in Abbildung [4.10](#) dargestellt.

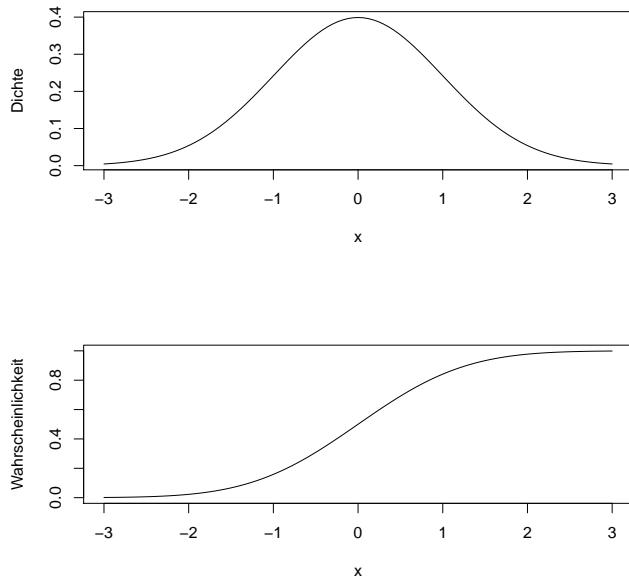


Abbildung 4.10: Wahrscheinlichkeitsdichte $\phi(x)$ (oben) und Verteilungsfunktion $\Phi(x)$ (unten) der Standardnormalverteilung.

Die Wahrscheinlichkeiten der Verteilungsfunktion kann man nicht aus einer geschlossenen Formel berechnen. Sie sind jedoch *tabelliert*. Heute werden sie aber eher in Computerprogrammen abgerufen, wir kommen bald darauf zurück. Aus den tabellierten Werten von $\Phi(x)$ kann man alle Wahrscheinlichkeiten berechnen, mit denen die standardnormalverteilte Variable X einen Wert zwischen a und b annimmt. Für jedes a und $b > a$ gilt

$$\Pr(a < X < b) = \Phi(b) - \Phi(a). \quad (4.3.8)$$

Die Wahrscheinlichkeiten von Gleichung 4.3.7 sind für verschiedene Quantile in der z -Tabelle B.2 tabelliert. $\Phi(0) = 0.5$, $\Phi(1.96) = 0.975$, $\Phi(-1.96) = 1 - \Phi(1.96) = 0.025$. Umgekehrt gilt: $Q_{0.975} = 1.96$, $Q_{0.025} = -1.96$.

Abbildung 4.11 illustriert eine empirische (1000 simulierte standardnormalverteilte Zufallszahlen) und eine theoretische Normalverteilung. Abbildung 4.12 zeigt die Wahrscheinlichkeiten, mit der eine standardnormalverteilte Zufallsvariable Werte in gewissen Bereichen einnimmt.

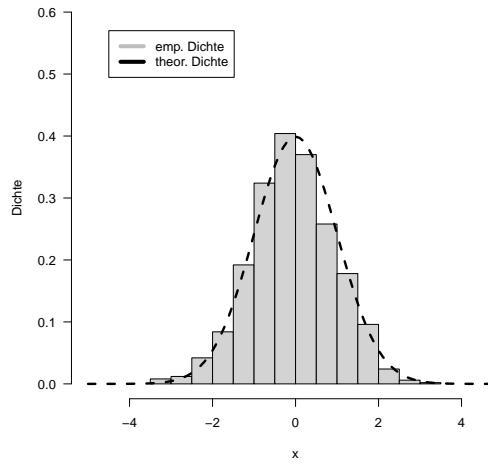


Abbildung 4.11: Empirische und theoretische Normalverteilung

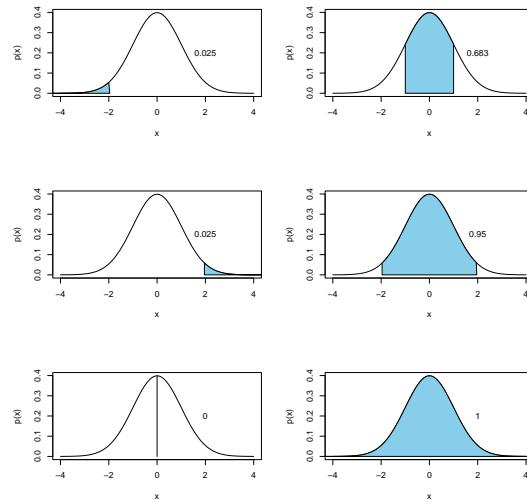


Abbildung 4.12: Wahrscheinlichkeit von verschiedenen Ereignissen

4.4 Andere Verteilungen

Es gibt sehr viele andere diskrete und stetige parametrische Verteilungen; wir wollen nachfolgend nur auf die wichtigsten eingehen.

Die χ^2 -, die t -und die F -Verteilung spielen eine wichtige Rolle, weil viele *Statistiken* χ^2 , t - oder F -verteilt sind. Diese Verteilungen bestehen aus ganzen Familien von Verteilungen, da sie im Gegensatz zur Normalverteilung noch durch sogenannte *Freiheitsgrade* (“Degrees of freedom”) spezifiziert sind. Das ist die Anzahl der Werte in der Berechnung einer Statistik, die frei variieren dürfen. Sie quantifizieren die Anzahl an *unabhängigen* Informationen, die für die Schätzung von Parametern zur Verfügung stehen.

χ^2 -Verteilung. Die Chi-Quadrat-Verteilung kann aus der Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ abgeleitet werden. Hat man n *unabhängige standardnormalverteilte* i.i.d. (independent and identically distributed) Zufallsvariablen $Z_i \sim \mathcal{N}(0, 1)$, so ist die Chi-Quadrat-Verteilung mit n Freiheitsgraden definiert als die Verteilung von

$$Q_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2. \quad (4.4.1)$$

Diese Grösse ist dann χ^2 -verteilt mit n Freiheitsgraden, $Q_n \sim \chi_n^2$. Solche Summen von quadrierten standardnormalverteilten Zufallsvariablen werden wir später bei Schätzfunktionen wie der Stichprobenvarianz antreffen. Wir brauchen diese Verteilung bei sogenannten χ^2 -Verfahren. Die χ^2 -Verteilung hängt ebenfalls zusätzlich von einem Freiheitsgrad ab. Abbildung 4.13 zeigt einige Vertreter der χ^2 -Verteilung mit verschiedenen Freiheitsgraden.

t -Verteilung. Sind X und Y unabhängig mit $X \sim \mathcal{N}(0, 1)$ und $Y \sim \chi_n^2$, so ist

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \quad (4.4.2)$$

t -verteilt mit n Freiheitsgraden. Wir werden im Kapitel 7 sehen, dass die Zufallsvariable *Mittelwert* normalverteilter Daten nicht mehr normalverteilt, sondern t -verteilt ist, wenn die Varianz des Merkmals unbekannt ist und mit der Stichprobenvarianz geschätzt werden muss. Die Abbildung 4.14 zeigt einige Vertreter der t -Verteilung mit verschiedenen Freiheitsgraden. Die t -Verteilung kann bei Freiheitsgraden grösser als 30 praktisch durch die Normalverteilung approximiert werden. Die t -Verteilungen brauchen wir bei sogenannten t -Test-Verfahren. Diese zählen zu den bekanntesten Tests.

F-Verteilung. Die F-Verteilung kann aus der χ^2 -Verteilung abgeleitet werden, sie ist die Verteilung der Zufallsvariable

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}, \quad (4.4.3)$$

mit χ_n^2 und χ_m^2 als unabhängigen Chi-Quadrat-verteilten Zufallsvariablen mit n bzw. m Freiheitsgraden (Beispiel für $n = 3$ und $m = 100$ in Abbildung 4.15). Diese Statistik spielt vor allem in klassischen *Varianzanalysen* eine wichtige Rolle.

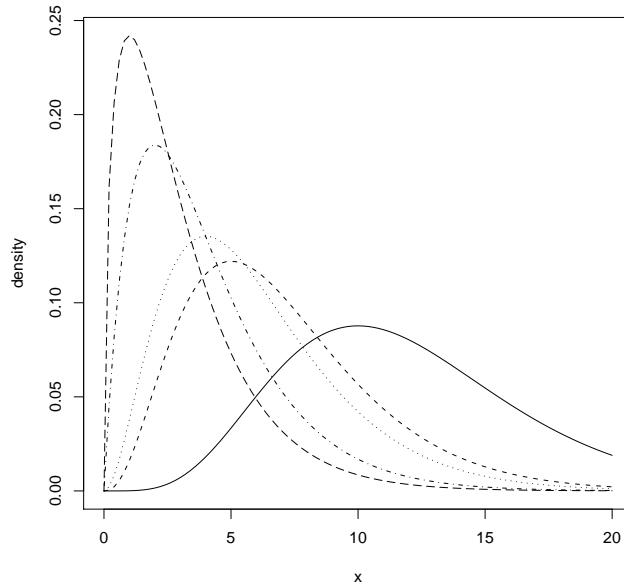


Abbildung 4.13: χ^2 -Verteilungen für $n = 12$ (—), $n = 7$ (---), $n = 6$ (···), $n = 4$ (··) und $n = 3$ (---) Freiheitsgrade.

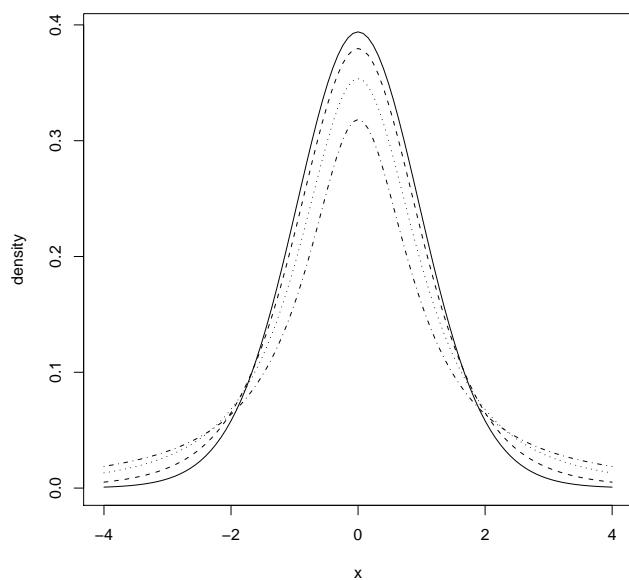


Abbildung 4.14: t -Verteilungen für $n = 1(-)$, $n = 2(\cdots)$, $n = 5(--)$, $n = 20(—)$ Freiheitsgrade.

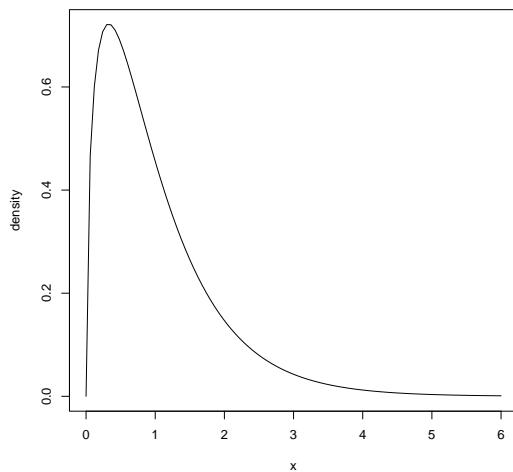


Abbildung 4.15: $F_{3,100}$ -Verteilung von $F = \frac{\chi^2_3}{\frac{\chi^2_{100}}{100}}$

4.5 Verteilungen mit R

Dieser Abschnitt ist für unseren Kurs zentral! Es gibt R-Funktionen, mit denen man

1. Zufallszahlen aus einer Verteilung ziehen kann
2. Die Dichte (oder Masse) bei $X = x$ angeben kann
3. Die kumulative Verteilung bei $X = x$ angeben kann
4. Quantile bei einer gewissen kumulierten Wahrscheinlichkeit herausgeben lassen kann

In R steht dann (unabhängig von der Verteilung)

- `r` für Zufallszahlen (`random`)
- `d` für Dichte **oder** Masse (`density`)
- `p` für Verteilungsfunktion (`probability`)
- `q` für die Quantilfunktion (`quantile`).

Für die Standardnormalverteilung z.B. bekommt man

- n Zufallszahlen mit `rnorm(n=,mean=0,sd=1)`
- Dichte d beim Quantil x mit `dnorm(x=,mean=0,sd=1)`
- Verteilungsfunktion p beim Quantil q mit `pnorm(q=,mean=0,sd=1)`
- Quantil q bei Verteilungsfunktion p mit `qnorm(p=,mean=0,sd=1)`.

Mit dem Kommando `?dnorm` oder `help(dnorm)` findet man sehr instruktive Hilfe zu den entsprechenden Funktionen.

Beispiele zur Normalverteilung. Gebe folgende Befehle in die R-Konsole ein:

- Kumulierte Wahrscheinlichkeit bei $X = 1.96$, wenn $X \sim \mathcal{N}(0, 1)$

```
pnorm(q = 1.96, mean = 0, sd = 1)
## [1] 0.975
```

- Das 0.975-Quantil von $X \sim \mathcal{N}(0, 1)$

```
qnorm(p = 0.975, mean = 0, sd = 1)
## [1] 1.96
```

- Dichte bei $X = 2$, wenn $X \sim \mathcal{N}(0, 1)$

```
dnorm(x = 2, mean = 0, sd = 1)
## [1] 0.054
```

- Wahrscheinlichkeit von $X < 1.96$, wenn $X \sim \mathcal{N}(0, 1)$

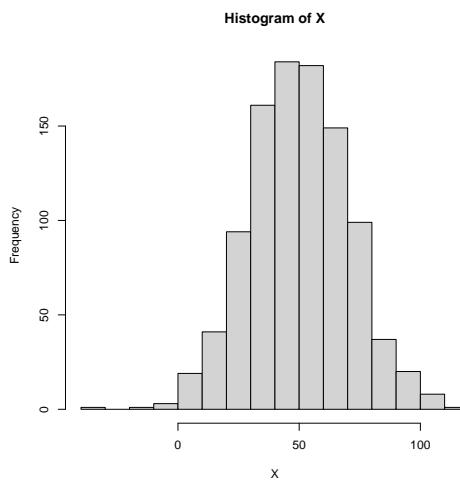
```
pnorm(q = 1.96, mean = 0, sd = 1)
## [1] 0.975
```

- Wahrscheinlichkeit von $-1.96 < X < 1.96$, wenn $X \sim \mathcal{N}(0, 1)$

```
pnorm(q = 1.96, mean = 0, sd = 1) - pnorm(q = -1.96, mean = 0, sd = 1)
## [1] 0.95
```

- Ziehen von $n = 1000$ Zufallszahlen aus $X \sim \mathcal{N}(50, 400)$

```
X <- rnorm(n = 1000, mean = 50, sd = 20) #Achtung: in R müssen wir sigma eingeben, nicht sigma^2
hist(X) #Wir kommen auf diesen Befehl in der deskriptiven Statistik zurück
```



Viele andere Verteilungen sind implementiert.

- `dbinom()`, `pbinom()`, `qbinom()`, `rbinom()` für Binomialverteilung
- `dpois()`, `ppois()`, `qpois()`, `rpois()` für Poissonverteilung
- `dt()`, `pt()`, `qt()`, `rt()` für t -Verteilung
- `dF()`, `pF()`, `qF()`, `rF()` für F -Verteilung
- `dchisq()`, `pchisq()`, `qchisq()`, `rchisq()` für χ^2 -Verteilung

Welche Argumente man in die Funktion einsetzen muss, kann man mit der Hilfefunktion erkennen. Schreibe `?dbinom` und `?dpois` in die Konsole und lese das Hilfesfile, vor allem die Teile **Description**, **Usage**, **Arguments** und **Details**.

Beispiele für andere Verteilungen.

- Masse bei $X = 2$, wenn $X \sim \text{Bin}(\pi = 0.6, n = 20)$

```
dbinom(x = 2, size = 20, prob = 0.6)
## [1] 0.0000047
```

- Wahrscheinlichkeit von $X \leq 1$, wenn $X \sim \text{Pois}(\lambda = 3)$

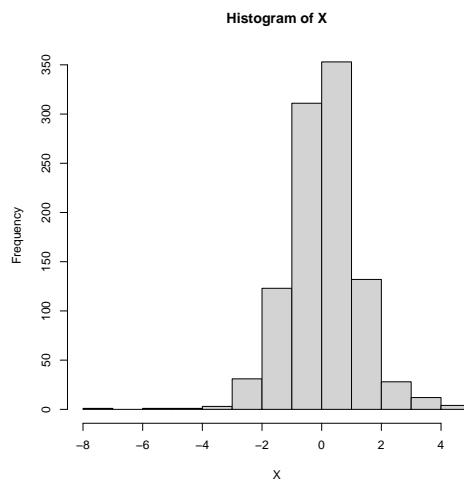
```
ppois(q = 1, lambda = 3)
## [1] 0.199
dpois(x = 0, lambda = 3) + dpois(x = 1, lambda = 3) ##Alternative
## [1] 0.199
```

- 0.5-Quantil von X , wenn $X \sim \text{Bin}(\pi = 0.5, n = 100)$

```
qbinom(p = 0.5, size = 100, prob = 0.5)
## [1] 50
```

- $n = 1000$ Zufallszahlen aus t -Verteilung mit 6 Freiheitsgraden.

```
X <- rt(n = 1000, df = 6)
hist(X)
```



4.6 Erwartungswerte

Grundidee. Wir möchten für eine Zufallsvariable X gewisse *Kennzahlen* finden, die in geeigneter Form das *durchschnittliche Verhalten* von X beschreiben.

4.6.1 Erwartungswert

Definition. Sei X eine diskrete Zufallsvariable mit Werten x_1, \dots, x_k mit Masse $p(x)$. Der *Erwartungswert* von X ist

$$E(X) = \sum_{i=1}^k x_i p(x_i). \quad (4.6.1)$$

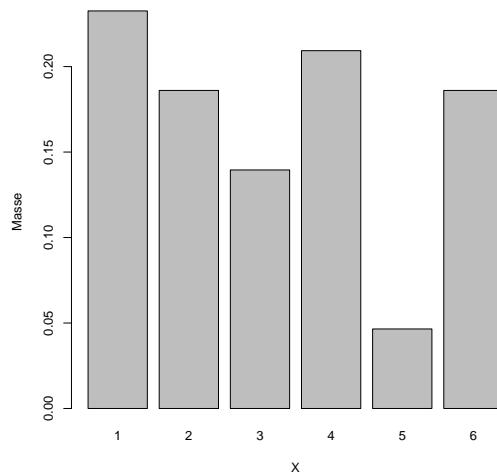
Wenn X kontinuierlich ist, dann ersetzen wir die Summe \sum durch ein Integral \int (und

die Masse durch Dichte.)

$$\mathbb{E}(X) = \int xp(x)dx. \quad (4.6.2)$$

Beispiel. Eine diskrete Zufallsvariable und ihre Verteilung sei gegeben:

```
##  X      p
## 1 0.23256
## 2 0.18605
## 3 0.13953
## 4 0.20930
## 5 0.04651
## 6 0.18605
```



Der Erwartungswert ist $1 \cdot 0.233 + 2 \cdot 0.186 + 3 \cdot 0.14 + 4 \cdot 0.209 + 5 \cdot 0.047 + 6 \cdot 0.186 = 3.209$.

4.6.2 Varianz und Standardabweichung

Definition. Sei X eine Zufallsvariable, dann heisst

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \quad (4.6.3)$$

die *Varianz* von X , und $\sqrt{\text{Var}(X)}$ heisst *Standardabweichung* von X . Man schreibt auch $\text{sd}(X) = \sigma = \sqrt{\text{Var}(X)}$. Sowohl die Varianz als auch die Standardabweichung sind Kennzahlen für die *Streuung* der Verteilung von X . In Worten ist die Varianz ein *erwarteter quadrierter Abstand vom Erwartungswert* oder – leicht weniger präzis – ein “mittlerer quadrierter Abstand”.

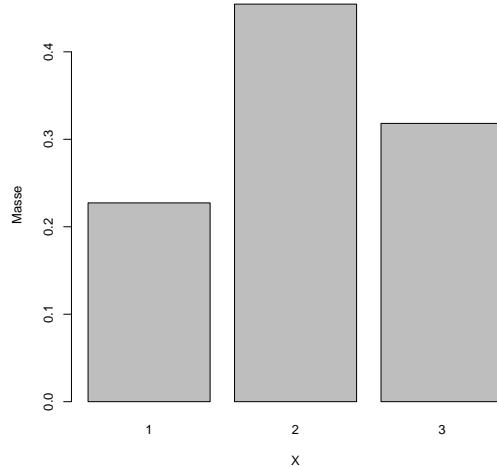
Man kann zeigen, dass

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \quad (4.6.4)$$

Mit dieser Regel kann oft einfacher gerechnet werden.

Beispiel. Eine diskrete Zufallsvariable und ihre Verteilung sei gegeben:

```
##   X      P
## 1 0.227
## 2 0.455
## 3 0.318
```



Der Erwartungswert ist $1 \cdot 0.227 + 2 \cdot 0.455 + 3 \cdot 0.318 = 2.091$. Die Varianz ist $\text{Var}(X) = E(X^2) - E(X)^2 = 1 \cdot 0.227 + 4 \cdot 0.455 + 9 \cdot 0.318 - 2.091^2 = 0.537$.

Varianz einer linearen Funktion einer Zufallsvariable. Wenn $Y = a + bX$ eine lineare Funktion der Zufallsvariable X ist, dann ist

$$\text{Var}(Y) = \text{Var}(a + bX) = b^2 \text{Var}(X) \quad (4.6.5)$$

Beachte das b^2 !

Beispiel Normalverteilung: $X \sim \mathcal{N}(0, 1)$ sei standardnormalverteilt mit Varianz 1. Dann ist $Y = 10 + 12 \times X$ normalverteilt mit Varianz 144, $Y \sim \mathcal{N}(10, 144)$.

```
X <- rnorm(1000)
var(X)

## [1] 0.997

Y <- 10 + 12 * X
var(Y)

## [1] 144
```

4.6.3 Erwartungswerte von wichtigen parametrischen Verteilungen.

Ohne Beweise geben wir in der Tabelle 4.2 Erwartungswert und Varianz an für wichtige parametrische Verteilungen.

Verteilung	Notation	Erwartungswert	Varianz
Bernoulli-Verteilung	$X \sim \text{Bin}(\pi, 1)$	π	$\pi(1 - \pi)$
Binomial-Verteilung:	$X \sim \text{Bin}(\pi, n)$	πn	$n\pi(1 - \pi)$
Poisson-Verteilung:	$X \sim \text{Pois}(\lambda)$	λ	λ
Normalverteilung:	$X \sim \mathcal{N}(\mu, \sigma^2)$	μ	σ^2
Standardnormalverteilung:	$X \sim \mathcal{N}(0, 1)$	0	1
t_n -Verteilung:	$X \sim t_n$	0	$\frac{n}{n-2}$
χ_n^2 -Verteilung:	$X \sim \chi_n^2$	n	$2n$
$F_{n,m}$ -Verteilung:	$X \sim F_{n,m}$	$m/(m-2)$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$

Tabelle 4.2: Erwartungswert und Varianz von wichtigen Verteilungen

Beispiel. Eine Zufallsvariable sei t -verteilt mit 14 Freiheitsgraden, $X \sim t_{14}$. Wir simulieren eine grosse Zufallsstichprobe aus dieser Verteilung und berechnen den (empirischen) Durchschnitt und die (empirische) Varianz⁵. Ersterer sollte nahe bei beim Erwartungswert 0, letzterer nahe bei der Varianz $14/12 = 1.167$ sein.

```
X <- rt(10000, df = 14)
mean(X)

## [1] 0.00216

var(X)

## [1] 1.17
```

⁵Mehr dazu in der beschreibenden Statistik.

4.7 Gemeinsame Verteilungen

Grundidee. Wir möchten das gemeinsame Verhalten von und die Zusammenhänge zwischen mehreren Zufallsvariablen beschreiben und untersuchen. Das erlaubt uns, den Begriff der Unabhängigkeit von Ereignissen auf Zufallsvariablen zu verallgemeinern.

Dazu schauen wir uns direkt ein sogenannte *bivariate* Verteilung an.

Bivariate diskrete Verteilung. Eine gemeinsame diskrete Verteilung von X und Y wäre

```
##      Y=0    Y=1    Y=2    Y=3    Sum
## X=1 0.125 0.250 0.125 0.000 0.500
## X=2 0.000 0.125 0.250 0.125 0.500
## Sum 0.125 0.375 0.375 0.125 1.000
```

Hier könnte z.B. X für “Interventionsart” stehen und Y für eine vierwertige Variable auf dem Konstrukt “Zufriedenheit”.

Gemeinsame Verteilung. Die 2×4 Wahrscheinlichkeiten $p(x, y) = \Pr(X = x, Y = y)$ stellen die *gemeinsame Verteilung* dar. So ist $p(1, 2) = 0.125$.

Randverteilungen. $p(x)$ stellt die *Randverteilung* von X dar, $p(y)$ stellt die Randverteilung von Y dar

- $p_X(1) = 0.5, p_X(2) = 0.5.$
- $p_Y(0) = 0.125, p_Y(1) = 0.375, p_Y(2) = 0.375, p_Y(3) = 0.125.$

Bedingte Verteilung. Die Verteilung $p(x | y)$ heisst *bedingte Verteilung* von X gegeben $Y = y$. Diese ist

$$p(x | y) = \frac{p(x, y)}{p(y)} \quad (4.7.1)$$

So ist z.B. für $Y = 2$: $p(1 | 2) = 0.125 / 0.375 = 0.333, p(2 | 2) = 0.25 / 0.375 = 0.667$.

Unabhängigkeit. X und Y sind genau dann unabhängig, wenn für alle y gilt

$$p(x | y) = p(x). \quad (4.7.2)$$

So sind A und B im folgenden Beispiel – im Gegensatz zu obigem Beispiel – unabhängig.

```
##      B=1    B=2    Sum
## A=1 0.25 0.25 0.50
## A=2 0.25 0.25 0.50
## Sum 0.50 0.50 1.00
```

Gemeinsame Normalverteilung. Eine grundlegende mehrdimensionale Verteilung ist die *zweidimensionale Normalverteilung*. Diese hat zusätzlich zu μ_x, μ_y, σ_x^2 und σ_y^2 noch einen weiteren Parameter ρ (eine Zahl zwischen -1 und 1). Man kann zeigen, dass ρ gerade die *Korrelation* zwischen X und Y darstellt, die wir unten einführen. Die Abbildung 4.16 zeigt die gemeinsame Verteilung von $X \sim \mathcal{N}(10, 9)$ und $Y \sim \mathcal{N}(10, 16)$ mit $\rho = 0$ (Links), $\rho = 0.5$ (Mitte) und $\rho = 0.8$ (Rechts). Die Höhe repräsentiert die gemeinsame Dichte $p(x, y)$. Das Volumen unter der Dichtefunktion ist gleich Eins.

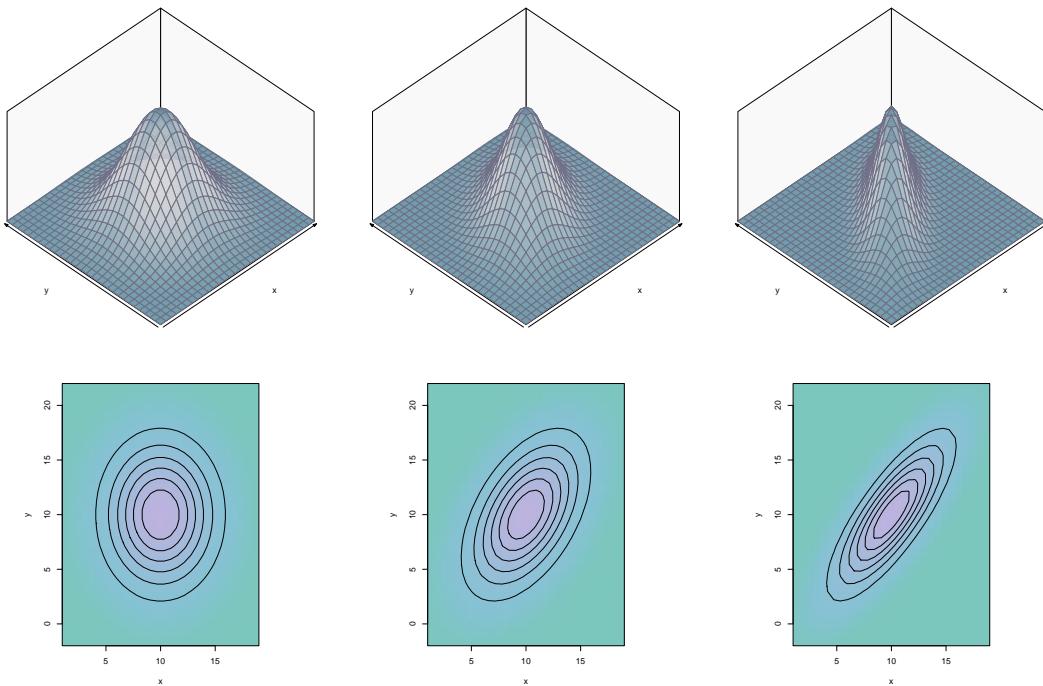


Abbildung 4.16: Dichte von gemeinsamer Normalverteilung

4.8 Kovarianz und Korrelation

Definition. Seien X und Y Zufallsvariablen. Dann heisst

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]\end{aligned}\tag{4.8.1}$$

die *Kovarianz* von X und Y . Die Kovarianz misst, ob – im Schnitt – die Abweichungen der Variable $X - \mu_X$ „zusammen gehen“ mit den Abweichungen der Variable $Y - \mu_Y$. Dann ist

$$\rho_{X,Y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}\tag{4.8.2}$$

die *Korrelation* von X und Y . Die Korrelation zwischen X und Y ist ein standardisiertes Mass für die Stärke des *linearen Zusammenhangs* zwischen den beiden Zufallsvariablen. Ist $\rho = 0$, so heissen X und Y unkorreliert.

- Korrelationen sind Zahlen zwischen -1 und 1, $|\text{Corr}(X, Y)| \leq 1$.
- $|\text{Corr}(X, Y)| = 1$ genau dann, wenn $Y = a + bX$ ist, mit $b \neq 0$ (Wenn also Y eine lineare Funktion von X ist⁶). Man nennt dann X und Y perfekt korreliert.
- Eine Kovarianz ist also eine mehrdimensionale Verallgemeinerung der Varianz. Die Kovarianz einer Variablen mit sich selber ist gerade die Varianz.

Wichtig. Unabhängige Zufallsvariablen X und Y sind immer unkorreliert, das Umgekehrte gilt aber in der Regel nicht. Wenn aber X und Y gemeinsam normalverteilt sind, dann impliziert Unkorreliertheit auch Unabhängigkeit.

Wie bei der Varianz (4.6.4) gibt es bei der Kovarianz eine Rechenregel. Man kann zeigen, dass (bei diskreten Zufallsvariablen)

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \sum_i \sum_j x_i y_j p_{ij} - \sum_i x_i p_i \times \sum_j y_j p_j.\tag{4.8.3}$$

Beispiel. Kommen wir zurück auf obiges Beispiel der diskreten bivariaten Verteilung.

```
##      Y=0    Y=1    Y=2    Y=3    Sum
## X=1 0.125  0.250  0.125  0.000  0.500
## X=2 0.000  0.125  0.250  0.125  0.500
## Sum 0.125  0.375  0.375  0.125  1.000
```

Die Kovarianz zwischen X und Y ist dann gemäss (4.8.3): $\text{Cov}(X, Y) = 0.125 \cdot 1 \cdot 0 + 0 \cdot 2 \cdot 0 + \dots + 0.125 \cdot 2 \cdot 3 - (0.5 \cdot 1 + 0.5 \cdot 2)(0.125 \times 0 + 0.375 \times 1 + \dots + 0.125 \cdot 3) = 0.25$. Aus den Randverteilungen kommen wir auf $\sigma_X = 0.5$ und $\sigma_Y = 0.866$, und $\rho_{X,Y} = 0.577$.

⁶Aus der Schule kennen wir vielleicht noch die einfache lineare Funktion, die durch $Y = a+bX$ gegebene Gerade.

4.8.1 Rechenregeln

Für Zufallsvariable X, Y, Z und Zahlen $a, b, c, d \in \mathbb{R}$ gilt:

1. $\text{Var}(a + bX) = b^2 \text{Var}(X)$
2. $\text{Cov}(X, X) = \text{Var}(X)$
3. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$
4. $\text{Cov}(X, a) = 0$
5. $\text{Cov}(X, bY) = b \text{Cov}(X, Y)$
6. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
7. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$
8. $\text{Cov}(a + \sum_i b_i X_i, c + \sum_j d_j Y_j) = \sum_i \sum_j b_i d_j \text{Cov}(X_i, Y_j)$
9. $\text{Var}(a + \sum_i b_i X_i) = \sum_{i,j} b_i b_j \text{Cov}(X_i, X_j)$
10. Sind X und Y unabhängig, dann ist $\text{Cov}(X, Y) = 0$
11. Sind X_1, \dots, X_n unabhängig, dann $\text{Var}(a + \sum_i b_i X_i) = \sum_i b_i^2 \text{Var}(X_i)$

Zum Schluss zwei wichtige Beispiele:

Beispiel 1. Varianz der Summe und vom Durchschnitt von unabhängigen Variablen X_1, X_2, X_3 mit $\text{Var}(X_1) = 10$, $\text{Var}(X_2) = 30$ und $\text{Var}(X_3) = 50$ sowie $\text{Cov}(X_i, X_j) = 0$ für $i \neq j$. Dann ist

$$\text{Var}(X_1 + X_2 + X_3) = \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) = 90.$$

Die Varianz vom Durchschnitt ist dann

$$\text{Var}\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{9}(\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)) = \frac{1}{9}(10 + 30 + 50) = 10$$

Beispiel 2. Wir möchten die Varianz vom Durchschnitt von n i.i.d. Zufallsvariablen (“independent and identically distributed”). Wenn X_1, X_2, \dots, X_n i.i.d., dann ist

$$\text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{n \text{Var}(X)}{n^2} = \frac{1}{n} \text{Var}(X) \quad (4.8.4)$$

Dieses Resultat werden wir sehr oft brauchen, insbesondere im Zusammenhang mit dem *Standardfehler* einer Schätzung. Es besagt, dass Durchschnitte weniger variieren als Einzelmessungen und dass man deren Varianz beliebig klein machen kann (wegmitteln), wenn n gross wird. Das entspricht auch unserer Alltagserfahrung.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999
3	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabelle 4.1: Standardnormalverteilung, Verteilungsfunktion von z .

Teil III

Statistik

Ziel. Man hat beobachtete Daten und will daraus Rückschlüsse ziehen auf den zugrundeliegenden *Mechanismus*, der diese Daten generiert hat.

Grundidee. Man fasst die Daten x_1, \dots, x_n auf als Realisierungen(realisierte Werte) $X_1(\omega), \dots, X_n(\omega)$ von Zufallsvariablen X_1, \dots, X_n . Dann sucht man (unter geeigneten Annahmen) Aussagen über die Verteilung von X_1, \dots, X_n .

Wichtig. Dazu muss man immer sauber unterscheiden zwischen den *Daten* x_1, x_2, \dots, x_n (bezeichnet mit kleinen lateinischen Buchstaben; x_1, \dots, x_n sind *Zahlen*) und dem generierenden Mechanismus X_1, \dots, X_n (bezeichnet mit grossen lateinischen Buchstaben; X_1, \dots, X_n sind Zufallsvariablen). Wir dürfen also Daten nicht durcheinanderbringen mit den Abstraktionen, die wir brauchen, um sie zu analysieren.

Terminologie. Die Gesamtheit der Beobachtungen x_1, \dots, x_n nennt man *Stichprobe*; die Anzahl n heisst Stichprobenumfang.

Kapitel 5

Beschreibende Statistik

Zuerst führen wir Begrifflichkeiten ein, die zur Beschreibung von Merkmalen zentral sind.

5.1 Statistische Einheiten und Variablen

Mit *statistischen Einheiten* meinen wir im Folgenden *Fälle* oder *Cases*, bei denen wir bestimmte *Merkmale* quantifizieren wollen. In unserem Umfeld stellen diese statistischen Einheiten meistens Patienten dar, es können aber prinzipiell alle möglichen Objekte wie Spitäler, Nationen, und Schulklassen statistische Einheiten sein.

An diesen Fällen möchten wir nun *manifeste*, weniger manifeste oder *latente* Merkmale *quantifizieren*. Diesen Merkmalen – dem *empirischen Relativ* – wird ein Korrelat in der Welt der Zahlen – ein *numerisches Relativ*, zugeordnet. Wir haben diese Zuordnung in [4.1](#) mit dem Begriff der *Zufallsvariable* $X : \Omega \rightarrow \mathbb{R}$ eingeführt.

So beschrieb die folgende Zufallsvariable “Anzahl von Therapien bis zur Heilung” im numerischen Relativ:

ω	G	KG	KKG	KKKG	KKKKG	KKKKKG
$X(\omega)$	1	2	3	4	5	6

Wir werden oft nur noch von einer Variable sprechen, wenn wir eigentlich Zufallsvariable meinen.

5.2 Typen von Merkmalen

Bevor wir uns mit Daten befassen, wollen wir verschiedene Arten von (Zufalls)Variablen anschauen. Das *Messen*, die oben erwähnte *Abbildung* eines empirischen in ein numerisches Relativ, ist nicht immer trivial, manchmal vielleicht sehr schwierig. Dann kommen sogenannte *qualitative Methoden* zum Zuge.

Nehmen wir aber nun an, das Messen, die erwähnte Abbildung, sei möglich. Welche Relationen bleiben nun bei dieser Abbildung, beim Messen, erhalten?

5.2.1 Skalenniveau

Nominalskala. Wenn nur *Gleich-Ungleich*-Relationen oder *nominale* Relationen bei dieser Abbildung erhalten bleiben, spricht man von einer kategorialen oder *nominal skalierten* Variable, z.B. die Variable *Nationalität*. Eine solche Variable wird – im numerischen Relativ – mit mehreren *Indikatorvariablen* beschrieben. Es braucht dabei so viele Zufallsvariablen wie Anzahl Kategorien. Wenn Nationalität dreiwertig ist, dann haben wir

ω	CH	AU	DE
$X_1(\omega)$	1	0	0
$X_2(\omega)$	0	1	0
$X_3(\omega)$	0	0	1

Ordinalskala. Manchmal können *Grösser-Kleiner*-Relationen oder Ordnungsrelationen erhalten bleiben. Dann sprechen wir von einer *ordinal skalierten* Variablen mit geordneten Kategorien. Ein Beispiel wäre ein Fragebogen zu *Selbständigkeit*. Diese Variablen haben eine Zwischenstellung zwischen nicht geordneten kategorialen Variablen und den unten folgenden metrischen Variablen: Man kann also die Ordnung ignorieren und die Variable behandeln wie eine kategoriale Variable:

ω	wenig	mittel	hoch
$X_1(\omega)$	1	0	0
$X_2(\omega)$	0	1	0
$X_3(\omega)$	0	0	1

Man kann die Skala auch “überinterpretieren” und unbescheiden als *metrisch* behandeln. Das führt uns zur nächsten Skala, der Intervallskala¹.

Intervallskala. Können zusätzlich *Distanzen* homomorph² abgebildet werden, haben wir es mit einer *intervallskalierten* Variable zu tun, und die *Addition* und *Subtraktion* von Messwerten ergibt einen Sinn. Behandeln wir obiges Beispiel als solches, dann wäre die Zufallsvariable (Wir brauchen dann nur eine!):

ω	wenig	mittel	hoch
$X(\omega)$	1	2	3

Die Celsius-Skala für Temperatur wäre ein anderes, diesmal unzweideutiges Beispiel einer intervallskalierten Variable:

ω	keine Temperatur	...	angenehm	...	unendlich
$X(\omega)$	-273	...	20	...	∞

¹Andere Prozeduren für die Behandlung von ordinalen Variablen brauchen viel Mathematik, wir gehen hier nicht darauf ein.

²Das heißt, die *Metrik* bleibt erhalten.

Ratioskala. Hat das Merkmal schliesslich auch einen *absoluten Nullpunkt*, haben wir eine *Ratio-Skala* und wir können auch *Verhältnisse* interpretieren; dies wäre bei der Kelvin-Skala der Fall sowie für viele physikalische und physiologische Merkmale.

ω	keine Temperatur	...	angenehm	...	unendlich
$X(\omega)$	0	...	293	...	∞

Wichtig. Das Skalenniveau beschreibt *nicht*, wie das Messinstrument *aussieht*, sondern welche Relationen während der beschriebenen Abbildung erhalten bleiben. In diesem Sinne ist das Skalenniveau vor allem bei latenten Merkmalen oft *hypothetisch*.

Je nachdem, welche Strukturen bei der Abbildung erhalten bleiben (Kategorien oder Ordnungen, Äquidistanz oder Verhältnisse), nennen wir die Variable also nominal-, ordinal-, intervall-, oder ratioskaliert. Tabelle 5.1 fasst das Gesagte zusammen.

<i>Skala</i>	<i>Definition</i>	<i>Beispiel</i>	<i>Operation</i>
Nominal	Kategorien	Geschlecht	$=, \neq$
Ordinal	Ordnung	Selbständigkeit	\leq, \geq
Intervall	Äquidistanz	Celsius	$+, -$
Ratio	Verhältnisse	Kelvin	$\cdot, /$

Tabelle 5.1: Skalenniveau von Variablen

5.2.2 Quantitative versus qualitative Merkmale

Nominalskalierte Variablen wie das biologische *Geschlecht* sind prinzipiell *qualitativ*, da eine *Qualität* und nicht ein *Ausmass* beschrieben wird. Metrische Variablen wie das *Gewicht* sind immer *quantitative Merkmale*, da ja eine *Metrik* gerade ein Ausmass, eine Quantität, beinhaltet. Ordinalskalierte Merkmale wie *Selbständigkeit* besitzen je nach Interpretation einen quantitativen oder qualitativen Charakter, eigentlich sind sie aber qualitativ. Sie nehmen also eine Zwischenstellung ein, wie wir schon gesehen haben.

5.2.3 Stetige versus diskrete Merkmale

Wenn ein Merkmal nur *endlich* viele oder *abzählbar unendlich* viele Ausprägungen annehmen kann, wie z.B. die Anzahl Studierenden in einer Klasse, nennen wir das Merkmal *diskret*. Dies ist bei *Zähldaten* der Fall.

Kann ein Merkmal hingegen *reelle Zahlen* innerhalb eines Intervalls annehmen kann (wie eine physikalischen Grösse), heisst das Merkmal *stetig* oder *kontinuierlich*. Stetige Merkmale können aber letztendlich nur *diskret gemessen* werden, da ja Messinstrumente immer eine gewisse Abstufung haben.

Oft werden Beobachtungen von stetigen Merkmalen zu *Klassen* zusammengefasst, gruppiert oder kategorisiert, das heisst, die *Rohdaten* werden in *Klassen* eingeteilt. Das

führt dann zu *Häufigkeitsdaten*, auf die wir später eingehen.

5.3 Stichprobe

Eine Stichprobe ist eine *Teilmenge* einer Grundgesamtheit, der *Population*. *Zufällige* Stichproben erhält man, wenn jede statistische Einheit dieselbe Chance hat, in die Stichprobe aufgenommen zu werden. Eine solche Stichprobe ergibt dann ein getreues, ein repräsentatives Abbild der Population. Interessiert nun eine Variable wie z.B. die Körpergrösse X , dann besteht eine *einfache Zufallsstichprobe* aus n Zufallsvariablen

$$X_1, X_2, \dots, X_i, \dots, X_n, \quad (5.3.1)$$

X_i beschreibt den Mechanismus, das *Modell* für die Körpergrösse für das i -te Individuum. X_i ist also eine *Abstraktion*, die a priori die Verteilung von X_i darstellt. Oft nimmt man an, dass die X_i *unabhängig und gleichverteilt* (independent and identically distributed, i.i.d.) sind.

Im Gegensatz dazu ist x_i (jetzt klein geschrieben!) der i -te *gemessene Wert* (a posteriori), dieser ist dann nicht mehr eine Zufallsgrösse oder ein Modell, sondern eine Zahl. Im Gegensatz zu den Zufallsvariablen X_i stellen also die x_i

$$x_1, x_2, \dots, x_i, \dots, x_n \quad (5.3.2)$$

n *Beobachtungen* oder *Realisierungen* der Zufallsgrösse X dar. Wir nennen diese n Werte auch die *konkrete Stichprobe*.

Die deskriptive Statistik befasst sich mit dem Beschreiben der konkreten Stichprobe.

5.4 Empirische Verteilungen

Wir werden nun die Messwerte oder die beobachteten Werte *beschreiben*. Wie kann man die Werte x_1, \dots, x_n eines Merkmals oder Variable X darstellen? Wir bezeichnen die Werte

$$x_1, \dots, x_n \quad (5.4.1)$$

als *Urliste*, *Roh-* oder *Primärdaten*. Falls n gross ist, sind übersichtlichere Darstellungen der Daten nötig. Dazu müssen wir zuerst den Begriff der *Häufigkeit* und der *empirischen Verteilung* von Beobachtungen einführen.

5.4.1 Häufigkeiten

Um Daten zusammenzufassen, werden die Rohdaten nach verschiedenen Werten oder Ausprägungen durchsucht. Als absolute oder relative Häufigkeit eines Wertes bezeichnet man die *Anzahl* bzw. den *Anteil* von Werten der Urliste.

Beispiel. Gegeben seien $n = 150$ Beobachtungen der Variable **Alter**, also 150 Werte x_1, x_2, \dots, x_{150} . Diese Daten liegen in einer *.csv*-Datei (*comma-separated values*) als *firstdata.csv* auf einem *GitHub-Repository* bereit, zusammen mit anderen Datensätzen, die wir im Laufe des Kurses brauchen werden.

Wir lesen die Datei in R ein mit der Funktion `read.csv()` und nennen das Objekt **firstdata**:

```
firstdata <- read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/firstdata.csv")
```

Zuerst schauen uns die Struktur von diesem Objekt an (`str()`):

```
str(firstdata)

## 'data.frame': 150 obs. of 1 variable:
## $ Alter: int 18 19 19 19 19 19 19 19 19 ...
```

Das Objekt **firstdata** ist ein data frame mit $n = 150$ Beobachtungen von einer einzelnen numerischen Variablen. Wir speichern diese Variable in ein Objekt mit dem Namen **Alter**. Mit der `$`-Syntax kann man bei data frames Kolonnen auswählen:

```
Alter <- firstdata$Alter
## Alter<-firstdata[,1] ## Alternative
```

Tabelle 5.2 zeigt eine tabellarische Darstellung der Häufigkeiten der $n = 150$ Werte. So kommen z.B. 18-jährige (der Werte 18) 1-mal, 19-jährige 24-mal vor, usw. 1 und 24 sind die absoluten Häufigkeiten (h) der Werte 18 und 19.

Wenn die Variable X mindestens ordinalskaliert ist, macht auch der Begriff der *kumulierten Häufigkeit* Sinn. Die absolute und relative kumulierte Häufigkeit eines Wertes ist gleich der Anzahl (bzw. Anteil) der Werte, die *kleiner oder gleich* diesem Wert sind. Die kumulierte Häufigkeit (H) beim Alter 18 ist 1, beim Alter 19 ist sie 25, usw.

Die Kolonnen f und F stellen die relativen und kumulierten relativen Häufigkeiten dar. Sie sind nichts anderes als h/n und H/n .

Alter	h	H	f	F
18	1	1	0.0067	0.0067
19	24	25	0.1600	0.1667
20	39	64	0.2600	0.4267
21	22	86	0.1467	0.5733
22	19	105	0.1267	0.7000
23	12	117	0.0800	0.7800
24	9	126	0.0600	0.8400
25	8	134	0.0533	0.8933
26	7	141	0.0467	0.9400
27	3	144	0.0200	0.9600
28	2	146	0.0133	0.9733
29	2	148	0.0133	0.9867
35	1	149	0.0067	0.9933
42	1	150	0.0067	1.0000

Tabelle 5.2: Häufigkeitsverteilung von $n = 150$ Beobachtungen der Variable Alter: h: absolute Häufigkeit, H: kumulierte absolute Häufigkeit, f: relative Häufigkeit, F: kumulierte relative Häufigkeit.

Wir wollen jetzt diese verschiedenen Begriffe präziser definieren. Häufig werden nicht die *Rohdaten* selber betrachtet, sondern Beobachtungen werden zuerst *gruppiert*, sie werden in k Klassen $c_j, j = 1, \dots, k$ eingeteilt, in *links geschlossene* Intervalle

$$[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k), \quad (5.4.2)$$

also mit Klassen $[18, 20), [20, 22), \dots$. Die *Häufigkeiten* der Werte in den k Klassen werden dann dargestellt in Form einer *Häufigkeitsverteilung* h_j und einer *relativen Häufigkeitsverteilung* f_j . Die Gruppierung ist natürlich nicht zwingend. Wir könnten direkt die Rohdaten aus Tabelle 5.2 nehmen, also mit einer Klassenbreite von 1 Jahr.

Wenn wir nun nach der *Anzahl* an Beobachtungen (#) in einer bestimmten Klasse c_j fragen, dann bestimmen wir die *absolute Häufigkeit* h_j ,

$$h_j = \#\{i \mid x_i \in c_j\}. \quad (5.4.3)$$

Darstellungen von *gruppierten* absoluten oder relativen Häufigkeiten von metrischen Variablen nennt man *Histogramme*. Abbildung 5.1 zeigt ein Histogramm der Variable **Alter**. Diese Variable wurde in $k = 12$ Klassen $[c_{j-1}, c_j), j = 1, \dots, 12$ unterteilt mit *Klassenbreite* = 2 Jahre. Die Klassenbreite wird in R automatisch eingestellt. Man kann sie aber natürlich über ein Argument ändern (`breaks=`). Meistens ist aber die Default-Klassenbreite gut zu gebrauchen. Die Anzahl der Klassen sollte von der Größenordnung \sqrt{n} sein. Das ist hier mit $n = 150$ gegeben. Ebenso kann man statt

links geschlossene rechts geschlossene Intervalle, $(c_{j-1}, c_j]$, einstellen (dies ist defaultmäßig eingestellt, `right=TRUE`) mit den Klassen $(18, 20], (20, 22], \dots$. Der kleinste Wert ist zwar ausserhalb des ersten Intervalls, wird aber dort gezählt³.

Unter <https://rstudio.zhaw.ch/rsconnect/content/86> kann man allgemein die Wirkung von Klassenbreite auf das Histogramm studieren.

```
hist(Alter, labels = TRUE, freq = TRUE, right = FALSE)
hist(Alter, labels = TRUE, freq = TRUE, right = FALSE, breaks = seq(18, 42, by = 1))
hist(Alter, labels = TRUE, freq = TRUE, right = TRUE)
```

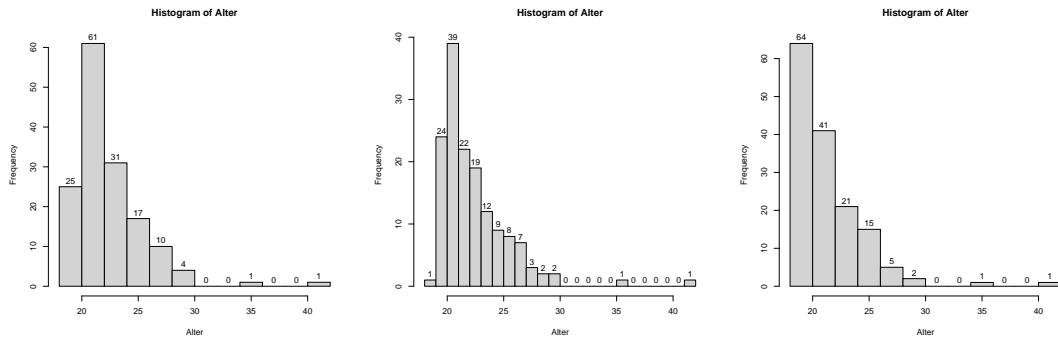


Abbildung 5.1: Absolute Häufigkeitsverteilung, Klassenbreite 2 Jahre vs. 1 Jahr, rechts: links offene Intervalle

Wenn wir nach der *Proportion*, *relativen Häufigkeit* oder *Anteil* von Beobachtungen in einer bestimmten Klasse c_j fragen, bestimmen wir die *relative Häufigkeit*

$$f_j = \frac{h_j}{n}. \quad (5.4.4)$$

Abbildung 5.2 zeigt die entsprechende Verteilung. Es ändert sich nur die Beschriftung der y -Achse. Das absolute Mass wird durch ein relatives Mass ersetzt. Da die Klassenbreite 2 Jahre ist, muss man die angegebenen *Dichten* mit 2 multiplizieren (siehe (4.1.5)). Mit Dichte ist ja die *relative Häufigkeit pro Jahr* gemeint. Wichtig: Die relative Häufigkeit als solche wird also in einem Histogramm durch das *Balkenvolumen* beschrieben. Mit `freq=FALSE` werden Dichten geplottet. Das Histogramm hat dann eine totale Fläche von eins.

```
hist(Alter, labels = TRUE, freq = FALSE, right = FALSE)
```

³Argument `include.lowest` defaultmäßig `TRUE`, siehe `?hist`.

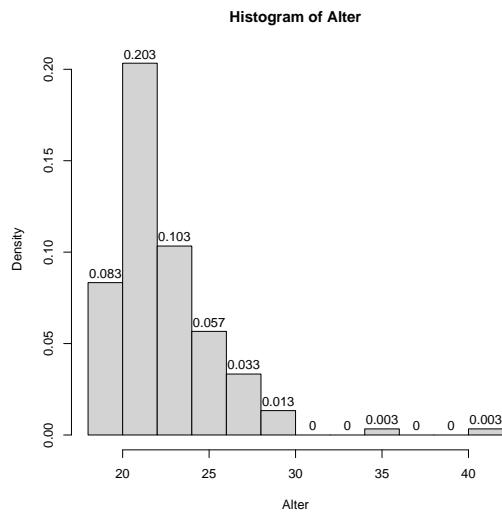
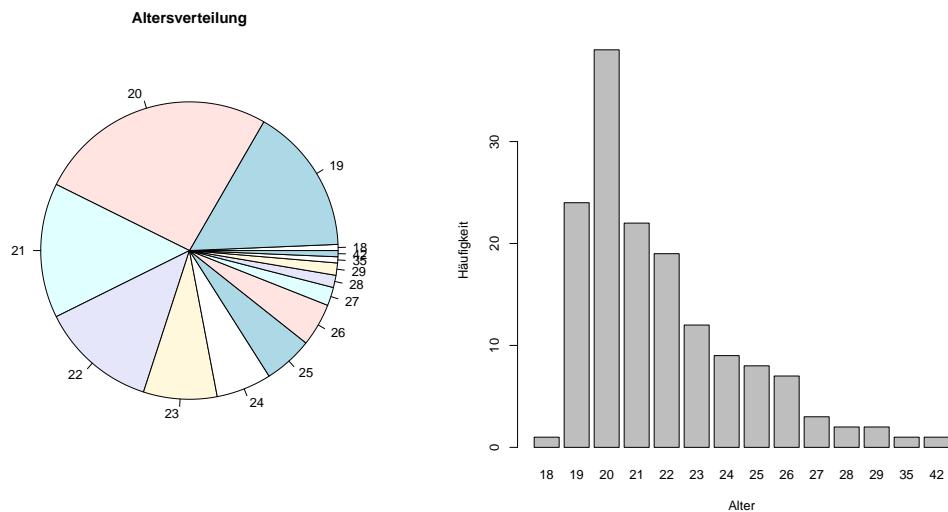


Abbildung 5.2: Relative Häufigkeitsverteilung (links geschlossene Klassen, Klassenbreite 2)

Andere Darstellungen. Ansonsten wird die Darstellung von Häufigkeiten auch mit den bekannten *Stab-, Säulen-, Balken- und Kuchendiagrammen* bewerkstelligt. Diese Darstellungen sind dann sinnvoll, wenn die Anzahl an Ausprägungen des Merkmals klein ist. Für unser Merkmal, das ja eigentlich kontinuierlich ist, ist das grenzwertig und nicht zu empfehlen.

```
pie(table(Alter), main = "Altersverteilung")
barplot(table(Alter), xlab = "Alter", ylab = "Häufigkeit")
```



Wir sehen, dass im Balkendiagramm die Variable wirklich als kategorial behandelt wird. So hat es im Gegensatz zum Histogramm – dort wurde die Variable als stetig behandelt – Abstände zwischen den Balken.

Für eine fünfwertige Variable z.B. eignen sich Kuchen- und Balkendiagramme besser. Mit `sample()` kann man eine solche Variable simulieren:

```
set.seed(30) ##Zufallsgenerator wird fixiert, damit wir alle dasselbe Ergebnis haben
Interesse <- sample(x = c("kein", "wenig", "mittel", "gross", "sehr gross"), prob = c(1, 1, 3, 3, 1),
size = 1000, replace = TRUE)
```

Das ist ein character-Vektor mit 1000 Werten gezogen mit Zurücklegen mit gewichteten Wahrscheinlichkeiten.

```
Interesse
str(Interesse)
```

Daraus machen wir zuerst einen *Faktor*. Damit die Ordnung erhalten bleibt, müssen wir die *vorkommenden Stufen* der Variable in der Reihenfolge eingeben, wie wir sie wollen. Ansonsten wird diese alphabetisch bestimmt.

```
Interesse0 <- factor(x = Interesse) #Falsch. Wenn wir levels nicht spezifizieren, wird die Reihenfolge alphabetisch gerichtet
levels(Interesse0)

## [1] "gross"      "kein"       "mittel"     "sehr gross" "wenig"
```

Richtig ist also

```
Interesse <- factor(x = Interesse, levels = c("kein", "wenig", "mittel", "gross", "sehr gross"))
levels(Interesse)

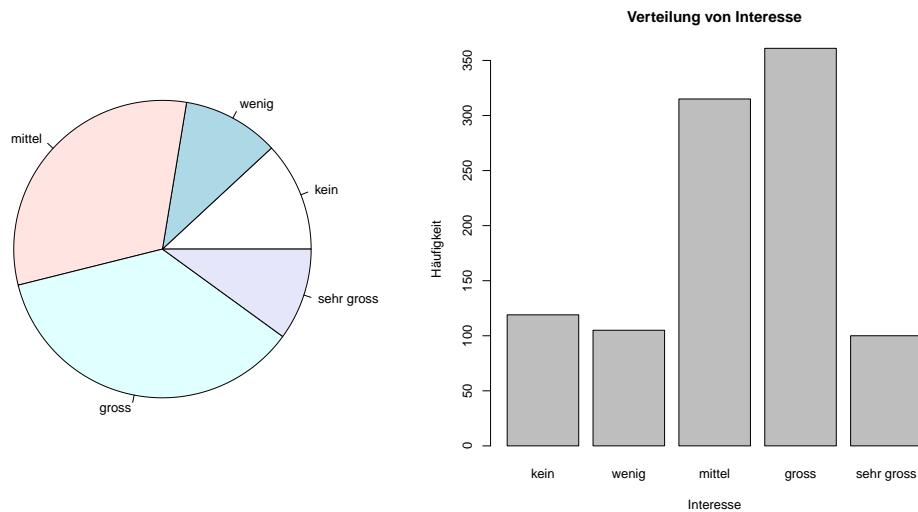
## [1] "kein"       "wenig"      "mittel"     "gross"      "sehr gross"
```

Mit `table()` können wir die Häufigkeiten abzählen

```
## Interesse
##      kein      wenig      mittel      gross      sehr gross
##      119       105       315       361        100
```

und die Verteilung plotten

```
pie(table(Interesse))
barplot(table(Interesse), xlab = "Interesse", ylab = "Häufigkeit", main = "Verteilung von Interesse")
## plot(Interesse) #Alternative. Plot erkennt Faktoren und macht einen Barplot.
```



In einem Balkendiagramm werden die einzelnen Kategorien sichtbar getrennt dargestellt, es wird also kein unterliegendes Kontinuum suggeriert. Für kontinuierliche Variablen ist das Histogramm besser geeignet. Sehr beliebt – bei geordneten Daten – ist auch der *Boxplot*, den wir unten einführen.

5.4.2 Kumulierte Häufigkeiten

Wir können auch nach der *kumulierten Häufigkeit* fragen. Die kumulierte Häufigkeit eines Wertes x ist die *Anzahl an Beobachtungen*, die *kleiner oder gleich* x sind wie x , d.h., wir *summieren* die absoluten Häufigkeiten *auf* bis zum Wert x ,

$$H(x) = \sum_{i:x_i \leq x} h_i(x). \quad (5.4.5)$$

Wir können das wieder relativ betrachten. Wie gross ist die *Proportion* oder relative Häufigkeit an Beobachtungen, die einen Wert *kleiner oder gleich* x haben? Dies ist dann gegeben durch

$$F(x) = \frac{H(x)}{n} \text{ oder } F(x) = \sum_{i,x_i \leq x} f_i(x). \quad (5.4.6)$$

So ist z.B. die kumulierte relative Häufigkeit beim Wert 21 (Alter = 21) 0.57 (siehe Tabelle 5.2 letzte Kolonne). Das bedeutet, dass 57% der Personen 21 Jahre alt oder jünger sind. Eine 21-jährige Person liegt also auf der 57. *Perzentile* oder dem 0.57-*Quantil*.

Abbildung 5.3 zeigt die *empirische kumulative Verteilungsfunktion* (`plot.ecdf()`) der Variable `Alter`, eine graphische Darstellung der letzten Kolonne der Tabelle 5.2. Eine kumulierte Verteilung ist eine monoton steigende *Treppenfunktion*. Anschaulich entsteht

sie durch Aufsummieren der einzelnen relativen Häufigkeiten.

```
plot.ecdf(Alter, xlab = "Alter", ylab = "F(After)", main = "kumulierte Verteilung Alter")
abline(v = 21, h = 0.57, lty = 3)
```

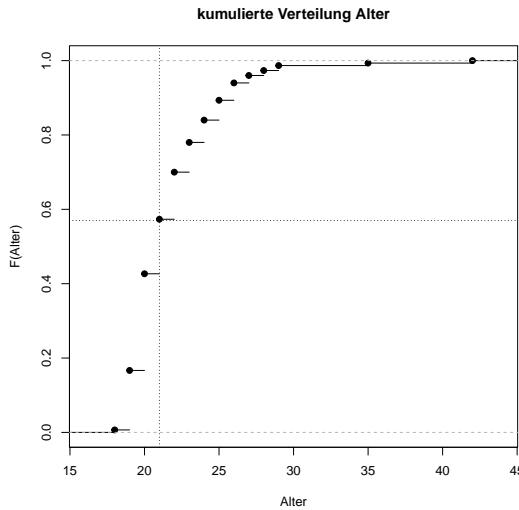


Abbildung 5.3: Kumulierte relative Häufigkeit

5.4.3 Quantile

Die Umkehrfunktion der empirischen kumulierten relativen Häufigkeit ist die empirische Quantilfunktion, in R ist diese zu haben über die Funktion `quantile()`.

```
quantile(Alter)
##   0%   25%   50%   75% 100%
## 18    20    21    23   42
```

Die Standardeinstellung gibt *Quartile* aus. Will man spezifische Quantile (z.B. Perzentile oder Dezile), dann kann man das Argument `probs` anpassen,

```
quantile(Alter, probs = seq(0, 1, 0.1)) ## für Dezile
quantile(Alter, probs = seq(0, 1, 0.01)) ## für Perzentile
```

Übung zur Häufigkeitstabelle*. Zuletzt wollen wir noch den R-Code für die Tabelle 5.2 angeben und diese reproduzieren.

```

h <- as.vector(table(Alter)) ## absolute Häufigkeiten
f <- as.vector(proportions(h)) ## relative Häufigkeiten
Werte <- unique(Alter) ## vorkommende Alterswerte
H <- cumsum(h) ## kumuliert Häufigkeiten
F <- cumsum(f) ## kumuliert relative Häufigkeiten
TabelleAlter <- data.frame(Werte, h, H, f, F)
str(TabelleAlter)

```

f und F sind nicht ganzzahlig. Beim Ausdrucken können wir einstellen, wie viele Nachkommastellen wir möchten.

```

round(TabelleAlter, digits = 4)

##   Werte   h   H     f      F
## 1     18   1   1 0.0067 0.0067
## 2     19  24  25 0.1600 0.1667
## 3     20  39  64 0.2600 0.4267
## 4     21  22  86 0.1467 0.5733
## 5     22  19 105 0.1267 0.7000
## 6     23  12 117 0.0800 0.7800
## 7     24   9 126 0.0600 0.8400
## 8     25   8 134 0.0533 0.8933
## 9     26   7 141 0.0467 0.9400
## 10    27   3 144 0.0200 0.9600
## 11    28   2 146 0.0133 0.9733
## 12    29   2 148 0.0133 0.9867
## 13    35   1 149 0.0067 0.9933
## 14    42   1 150 0.0067 1.0000

```

5.5 Kennwerte von Verteilungen

Die empirische Verteilung einer Variablen X kann mit *empirischen Kennwerten* zusammengefasst werden. Es gibt zwei Arten von Kennwerten, *Lagemasse* oder Masse für die *zentrale Tendenz* einerseits und *Streuungsmasse* anderseits.

Tabelle 5.3 zeigt eine Übersicht der Kennwerte; diese werden in den folgenden Abschnitten eingeführt. Wie man sehen kann, hängen mögliche Kennwerte von der Skalierung der Variable ab.

5.5.1 Lagemasse

Empirischer Mittelwert

Das gebräuchlichste Lagemasse oder Mass für die zentrale Tendenz ist der *empirische Mittelwert* oder der *arithmetische Mittelwert*, (`mean()`). Wir notieren ihn mit \bar{x} . Der empirische Mittelwert von n Beobachtungen einer Variable X ist

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i. \quad (5.5.1)$$

Kennwert / Skala	nominal	ordinal	metrisch
<i>Lage</i>	Modus	Modus	Modus
		Median	Median
			Mittelwert
<i>Streuung</i>		Range Interquartilrange	Range Interquartilrange Standardabweichung Varianz

Tabelle 5.3: Masse für die zentrale Tendenz und für die Variabilität bei verschiedenen Skalierungen.

Die Summe aller Abstände vom Mittelwert ist immer 0,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0, \quad (5.5.2)$$

d.h. der Mittelwert ist zu interpretieren als der *Schwerpunkt* der Daten, im Gegensatz zum *Median*, den wir unten einführen.

Der arithmetische Mittelwert einer Stichprobe ist der am häufigsten benutzte Kennwert und der bekannteste Kennwert für die zentrale Tendenz. Jedoch ist der arithmetische Mittelwert nicht immer der sinnvollste Wert. Wenn ein paar extreme Werte den empirischen Mittelwert stark in eine Richtung drängen, ist der Mittelwert nicht immer das sinnvolle Lagemaß.

```
mean(Alter)

## [1] 21.9
```

Falls die Variable fehlende Werte hat wie

```
DatenMitMissings <- c(2, 4, 3, 2, NA, 5)
DatenMitMissings

## [1] 2 4 3 2 NA 5
```

dann muss das Argument `na.rm=TRUE` gesetzt werden.

```
mean(DatenMitMissings)

## [1] NA

mean(DatenMitMissings, na.rm = TRUE)

## [1] 3.2
```

Empirischer versus wahrer Mittelwert. Der *wahre Mittelwert* μ ist der Mittelwert in der Population, aus der die Stichprobe kommt. μ ist eine *abstrakte Grösse*. Man nennt diese Grösse auch den Erwartungswert E von X , $E(X) = \mu$. Wir haben diesen in 4.6.1 und 4.6.2 eingeführt als ein mit den Wahrscheinlichkeiten der Werte der Zufallsvariablen gewichtetes Mittel. Man kann sich den Erwartungswert auch als ein Mittel über unendlich viel Information aus der Verteilung in der Population vorstellen, da ja eine Wahrscheinlichkeit ein Grenzwert der relativen Häufigkeit ist und der Grenzwert von \bar{x} damit gerade dem Erwartungswert entspricht⁴. Mit dem *Gesetz der grossen Zahlen* kann man zeigen, dass $\lim_{n \rightarrow \infty} \bar{x} = \mu$, dass der Durchschnitt mit wachsendem n gegen μ strebt. In der deskriptiven Statistik ist vorerst aber \bar{x} , nicht μ von Interesse.

Median

Es gibt Lagemasche, die *robust* gegenüber Extremwerten sind. Ein Beispiel dafür ist der *Median* (`median()`). Man braucht ihn vor allem bei *nicht-symmetrischen Verteilungen* sowie bei *nicht intervallskalierten* Merkmalen.

Einen Firmenmitarbeiter z.B. interessiert es nicht nur, welches der *mittlere Lohn* im Sinne des arithmetischen Mittels ist, weil der Mitarbeiter weiss, dass der *Schwerpunkt* des Lohns von extremen Salären nach oben gezogen wird. Den Mitarbeiter interessiert vor allem, welcher Lohn alle Mitarbeiter in zwei Hälften teilt, in eine, die mehr verdient, und in eine, die weniger verdient. Dieser Lohn entspricht dem Median. Bei sogenannten *linkssteilen* Verteilungen (die Lohnverteilung in einer Bank wird eher linkssteil sein) wird der Mittelwert hin zu den extremen Werten gezogen. Der Mittelwert ist also *nicht robust* gegenüber – sprichwörtlichen – *Ausreissern*. (Bei rechtssteilen Verteilungen verhält es sich gerade umgekehrt.)

Für die Berechnung des Medians müssen die Primärdaten zuerst *geordnet* werden. Der Median einer *geordneten Stichprobe* $x_{(1)} \leq x_{(i)} \dots \leq x_{(n)}$ von n Beobachtungen ist gleich

$$\text{median}(x) = \begin{cases} \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{für } n \text{ gerade,} \\ x_{((n+1)/2)} & \text{für } n \text{ ungerade.} \end{cases} \quad (5.5.3)$$

Der Median entspricht der Perzentile $P50$ oder dem 0.5-Quantil $Q_{0.5}$.

```
median(Alter)
## [1] 21
```

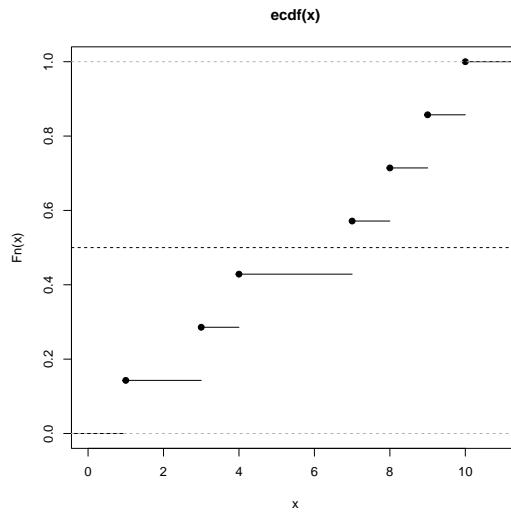
Beispiel 1. Der Median der Werte $\{10, 1, 7, 4, 9, 3, 8\}$ ist gleich “der Mitte” der geordneten Stichprobe $1 \leq 3 \leq 4 \leq 7 \leq 8 \leq 9 \leq 10$, also 7.

⁴Bei einer diskreten k -wertigen Variablen: $E(X) = \sum_1^k x_i p(x_i)$. Der empirische Mittelwert ist $\bar{x} = \sum_1^k x_i f(x_i) = \sum_1^k x_i h(x_i) / n = 1/n \sum_1^k x_i h(x_i)$.

```
median(ungeradeAnzahl <- c(10, 1, 7, 4, 9, 3, 8))

## [1] 7

plot.ecdf(ungeradeAnzahl, verticals = FALSE)
abline(h = 0.5, lty = 2)
```

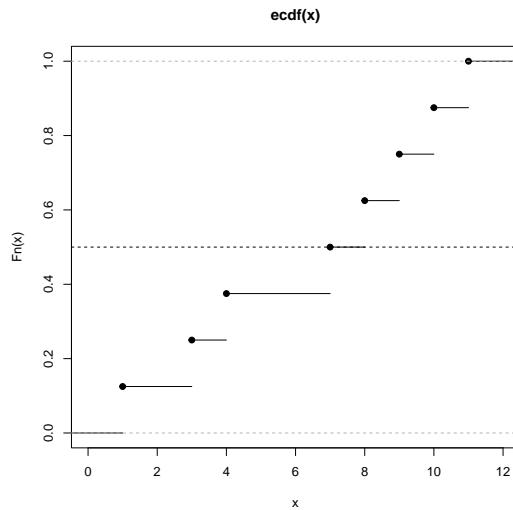


Beispiel 2. Bei gerader Anzahl an Beobachtungen gibt es zwei “Mitten” der geordneten Stichprobe, dann wird der Durchschnitt gemäss (14.2.2) genommen. Der Median der Werte $\{10, 1, 11, 7, 4, 9, 3, 8\}$ gleich “der Mitte” der geordneten Stichprobe $1 \leq 3 \leq 4 \leq 7 \leq 8 \leq 9 \leq 10 \leq 11$, also 7 (“Untermedian”) oder 8 (“Obermedian”).

```
median(gerAnzahl <- c(10, 1, 11, 7, 4, 9, 3, 8))

## [1] 7.5

plot.ecdf(gerAnzahl, verticals = FALSE)
abline(h = 0.5, lty = 2)
```



Modus

Schlussendlich gibt es noch ein drittes Mass für die zentrale Tendenz, den *Modus*. Der Modus entspricht *dem Wert* oder Ausprägung mit *maximaler Häufigkeit*, also dem Wert, der in der Stichprobe am *häufigsten* vorkommt. Der Modus macht Sinn ab Nominalskalenniveau,

$$\text{modus}(x) = \arg \max_x h(x), \quad (5.5.4)$$

d.h., der Modus ist der Wert von X , der $h(x)$ maximiert. Der Modus ist aber *nicht* immer *eindeutig*. Er ist dann eindeutig, wenn die Häufigkeitsverteilung ein eindeutiges Maximum besitzt, ansonsten haben wir eine bimodale oder multimodale Verteilung (Abbildung 5.4).

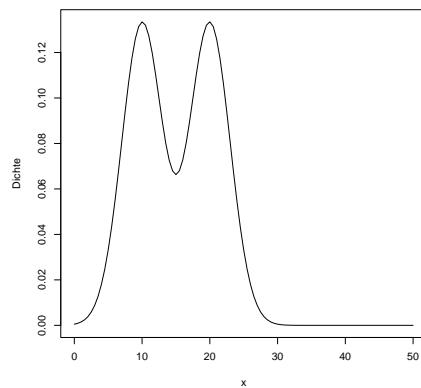


Abbildung 5.4: bimodale Verteilung

Eigenschaften der Lagemasse. Sei a ein Zentrum auf der x -Achse. \bar{x} ist das Lagemass, das die Summe der quadrierten Abweichungen $\sum_{i=1}^n (x_i - a)^2$ minimiert, da immer gilt $\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$. Diese Eigenschaft des Mittelwerts werden wir später für die *Methoden der kleinsten Quadrate* brauchen.

Der Median optimiert eine andere Grösse. Er minimiert die Summe der absoluten Abweichungen $\sum_{i=1}^n |x_i - a|$, weil immer gilt $\sum_{i=1}^n |x_i - \text{median}| \leq \sum_{i=1}^n |x_i - a|$. Extremwerte beeinflussen diese Summe der absoluten Abweichungen weniger als die Summe der quadrierten Abweichungen, daher ist der Median robuster gegenüber Extremwerten.

Auch der Modus optimiert eine Grösse. Er minimiert die Summe $\sum_{i=1}^n I(x_i, a)$, wobei $I(x, a) = 1$ für $x \neq a$ und $I(x, a) = 0$ für $x = a$. Diese Summe macht Sinn ab Nominalskalenniveau, da durch $I(\cdot)$ nur Gleichheit/Ungleichheit quantifiziert wird.

5.5.2 Streuungsmasse

Wenn wir eine Stichprobe beschreiben wollen, brauchen wir neben einer Grösse, die die zentrale Tendenz beschreibt, noch eine Grösse, die die *Streuung* oder die *Variabilität* der Daten beschreibt. Solche Kennwerte werden im Folgenden eingeführt.

Range

Der *Range* (`range()`) ist das einfachste Streuungsmass und beschreibt den Abstand zwischen dem kleinsten und dem grössten Wert. Er macht Sinn ab Ordinalskalenniveau.

```
range(Alter)
## [1] 18 42

diff(range(Alter))
## [1] 24
```

Interquartilrange und Boxplot

Die Umkehrfunktion der kumulativen Verteilungsfunktion hiess Quantilfunktion 4.2.3. Das gilt auch für die empirische Welt 5.4.3. Quantile sind wichtige Grössen, da mit ihnen quantifiziert wird, wie gross die Wahrscheinlichkeit (relative Häufigkeit) ist, dass sich Werte in gewissen Intervallen befinden. Ein besonderes Quantil ist der Median. Er ist identisch mit $Q_{0.5}$. Andere spezifische Quantile sind auch die Quartile, $Q_{0.25}$ und $Q_{0.75}$. Perzentile sind Spezialfälle der p -Quantile, nämlich wenn $p \cdot 100\%$ ganzzahlig ist.

```
quantile(Alter, probs = seq(0, 1, 0.01)) ## für Perzentile
```

Der *Interquartilabstand (IQR)* (`IQR()`) ist der Abstand zwischen dem ersten ($Q_{0.25}$) und dem dritten Quartil ($Q_{0.75}$). Interquartilabstände werden oft mit einem sogenannten *Boxplot* visualisiert (`boxplot()`). Die Abbildung 5.5 zeigt die Verteilung der Variable `Alter` als Boxplot. In jedem Boxplot ist $Q_{0.25}$ die untere Begrenzung der *Schachtel* (Box) und $Q_{0.75}$ ist die obere Begrenzung der Schachtel. Man erkennt den Interquartilabstand (die Höhe der Box) und den Median (= 21) (mittlere horizontale Linie in der Box). Alle anderen Werte sind entweder oberhalb $Q_{0.75} = 23$ oder unterhalb $Q_{0.25} = 20$. 50% der Beobachtungen sind also zwischen 20 und 23 Jahren alt. Am Boxplot ist ersichtlich, dass die Verteilung nicht symmetrisch um den Median ist. Sie ist linkssteil, das war bereits in Abbildung 5.1 ersichtlich. Die vertikale Ausdehnung der Box hat keine Bedeutung.

```
IQR(Alter)

## [1] 3

boxplot(Alter, horizontal = TRUE, xlab = "Alter")
```

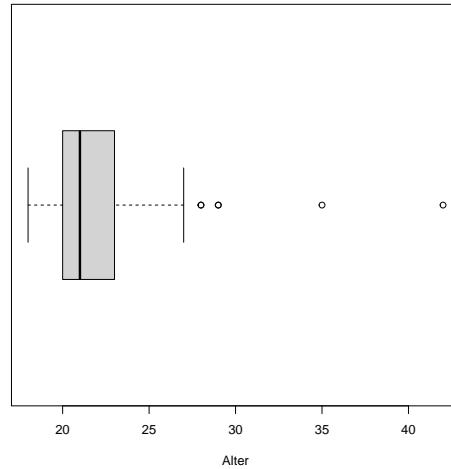


Abbildung 5.5: Boxplot Alter

In R ist `summary()` eine sogenannte *generische* Funktion, die – je nach Objekt – verschiedene Werte ausgibt. Es ist eine der wichtigsten R-Funktionen. Für numerische Vektoren gibt die Funktion Minimum, Maximum, Quartile und Mittelwert der Variable heraus. Wir werden diese Funktion später auch auf andere Objekte anwenden.

```
summary(Alter)

##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   18.0   20.0   21.0   21.9   23.0   42.0
```

Empirische Varianz und empirische Standardabweichung

Sobald eine Variable X mindestens intervallskaliert ist, also bei sogenannten metrischen Daten, wird als Mass für die Streuung der Verteilung oft die empirische *Standardabweichung* (`sd()`) oder die empirische *Varianz* (`var()`) angegeben.

Die *Stichprobenvarianz* oder *empirische Varianz* entspricht einer *mittleren quadrierten Abweichung der Daten vom Mittelwert*. Eine solche Grösse macht Sinn bei metrischen Daten. Um die Variabilität zu quantifizieren, werden die *quadrierten Abstände* aller Werte summiert und dann durch $n - 1$ geteilt. Man quadriert die Abstände, weil der mittlere nicht-quadrierte Abstand immer Null wäre, da sich positive und negative Abstände gegenseitig annullieren, siehe (5.5.2). Ein Mass, das unabhängig von den Daten immer 0 ist, ist aber unbrauchbar.

Die empirische Stichprobenvarianz s^2 ist gegeben durch

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (5.5.5)$$

Die Stichprobenvarianz (5.5.5) hat quadrierte Einheiten. Wurden z.B. Körpergrössen in Meter gemessen, so hat die Varianz der Variable Körpergrösse die Einheit m^2 . Um zur ursprünglichen Einheit der Messgrösse zurückzukehren, wird aus der Varianz die Wurzel gezogen. Diese Grösse bezeichnet man dann als die *empirische Standardabweichung* s (“standard deviation”). Die empirische Standardabweichung ist also

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (5.5.6)$$

Rechenbeispiel. Für Variablen mit n gross lässt man sich die Varianz und Standardabweichung von einem Computer berechnen. Um das Prinzip zu verstehen, schauen wir uns die kleine Stichprobe $\{x_1 = 2, x_2 = 0, x_3 = -2\}$ an. Der empirische Durchschnitt, der Schwerpunkt der Daten \bar{x} ist dann 0. Die drei Abstände sind $\{2, 0, -2\}$, die drei quadrierten Abstände sind $\{4, 0, 4\}$. Somit ist die empirische Varianz $s^2 = \frac{4+0+4}{3-1} = 4$. Die empirische Standardabweichung schlussendlich ist $\sqrt{4} = 2$.

```
smallsample <- c(2, 0, -2)
var(smallsample)

## [1] 4
```

```
sd(smallsample)
```

```
## [1] 2
```

Freiheitsgrade. Wieso teilen durch $n-1$ statt durch n ? s^2 ist eine *empirische* Grösse, die man direkt aus Daten berechnen kann. Die *wahre Varianz*, die (unbekannte) Varianz in der Population, wurde oben eingeführt als 4.6.3

$$\text{Var}(X) = \sigma^2 = E[(X - \mu)^2], \quad (5.5.7)$$

als der erwartete quadrierte Abstand. Der Leser versuche, die Analogie zwischen (5.5.5) und (5.5.7) zu “sehen”.

In (5.5.5) teilt man nicht durch n , sondern durch die *Anzahl Freiheitsgrade* ($n - 1$). Dies hat nämlich zur Folge, dass die Stichprobenvarianz s^2 eine *erwartungstreue* oder *unverzerrte* Schätzung der Populationsvarianz σ^2 ist, kurz damit $E(s^2) = \sigma^2$ gilt⁵.

Das lässt sich so erklären: In die Berechnung von s^2 fliessen n Werte mit ein. Als Zwischenschritt muss aber der Mittelwert geschätzt werden und somit geht ein Freiheitsgrad verloren. Man kann sich das auch so veranschaulichen, dass nur $n - 1$ von n Abweichungen $(x_i - \bar{x})$ frei variieren können, die letzte $(x_n - \bar{x})$ ist durch diese bestimmt, weil $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

5.6 Standardisierung

Um Abweichungen vom Mittelwert gut vergleichbar machen zu können, werden sie oft zuvor an der Unterschiedlichkeit aller Werte der jeweiligen Stichprobe relativiert. Dies geschieht durch die *z-Transformation* (`scale()`). Diese Transformation ordnet jeder Beobachtung x_i aus X einen neuen Wert z_i auf der *standardisierten Variable* Z zu. Die n Beobachtungen einer Stichprobe werden standardisiert gemäss der Regel

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \dots, n. \quad (5.6.1)$$

z_i quantifiziert, wie viele Standardabweichungen in der Differenz zwischen dem Messwert x_i und dem Durchschnitt \bar{x} enthalten sind. Die $z_i, i = 1, \dots, n$ haben dann nach Definition einen Mittelwert von 0 und eine Standardabweichung von 1.

Beispiel. Wir simulieren Poisson-verteilte Daten und standardisieren sie (von “Hand” und mit implementiertem Befehl).

⁵Streng genommen müsste man S^2 statt s^2 schreiben, aber meistens wird die *Variable der Stichprobenvarianz* klein geschrieben.

```
X <- rpois(n = 500, lambda = 8)
mean(X) #E(X)=lambda, mean ist nahe bei lambda
var(X) #Var(X)=lambda, idem
Z <- (X - mean(X))/sd(X) #von Hand
Z <- scale(X) #implementierte Funktion
mean(Z)
sd(Z)
```

Die Einheit der z_i ist also das Streuungsmass selber und nicht mehr die Einheit der Messung! z -Werte helfen uns z.B. bei der Frage, was als *extrem* oder als *normal* angesehen werden darf.

Bei Normalverteilung. Standardisierte Werte bekommen dann ein spezielle Bedeutung, wenn die Beobachtungen x_1, x_2, \dots, x_n aus einer normalverteilten Variablen X stammen, dann kommen die standardisierten Werte z_1, z_2, \dots, z_n aus einer standardnormalverteilten Variablen Z (4.3.6, Abbildung 4.10). Aufgrund der Symmetrie der Normalverteilung gilt dann $z_p = -z_{1-p}$, also ist z.B. $z_{0.1} = -z_{0.9} = -1.28$.

Das Quantil x_p erhält man dann aus dem Quantil z_p durch die inverse Transformation bezüglich Gleichung (5.6.1),

$$x_p = \bar{x} + s \cdot z_p. \quad (5.6.2)$$

Die Quantile z_p der Standardnormalverteilung haben wir in der Tabelle B.2 dargestellt. Wir haben auch bereits gesehen, dass wir uns in R gewünschte Quantile mit `qnorm(p)` und gewünschte kumulierte Wahrscheinlichkeit mit `pnorm()` herausgeben lassen können.

Der Übergang zur standardisierten Variable Z bewirkt, dass Beobachtungen als Abweichungen vom Mittel und mit der Standardabweichung als Maßeinheit gemessen werden. Mit Hilfe der Standardisierung können also alle Berechnungen für X , etwa die Berechnung von Quantile, auf Berechnungen für die Standardnormalverteilung zurückgeführt werden.

Die Standardabweichung wird also oft zusammen mit dem Mittelwert benutzt, um Intervalle der Form $\bar{x} \pm s$, $\bar{x} \pm 2s$ oder $\bar{x} \pm 3s$ anzugeben. Wenn das Merkmal normalverteilt ist, so erhält man Anteilsraten für Intervalle von X aus entsprechenden Intervallen für Z . Bei normalverteilter Variable X liegen

- ca. 68% der Beobachtungen im Intervall $\bar{x} \pm s$

```
pnorm(q = 1) - pnorm(q = -1)
```

- ca. 95% der Beobachtungen im Intervall $\bar{x} \pm 2s$

```
pnorm(q = 2) - pnorm(q = -2)
```

- 95% der Beobachtungen im Intervall $\bar{x} \pm 1.96s$

```
pnorm(q = 1.96) - pnorm(q = -1.96)
qnorm(p = c(0.025, 0.975))
```

- 99% der Beobachtungen im Intervall $\bar{x} \pm 2.58s$

```
pnorm(q = 2.58) - pnorm(q = -2.58)
qnorm(p = c(0.005, 0.995))
```

- Basierend auf der Normalverteilung kann man also den range einer normalverteilten Variablen grob abschätzen mit $Range = 6 \times sd$.

```
pnorm(q = 3) - pnorm(q = -3)
```

Empirische versus theoretische Quantile einer Normalverteilung. Abbildung 5.6 zeigt die empirische Verteilung von $n = 1000$ Beobachtungen der normalverteilten Variablen $X \sim \mathcal{N}(\mu = 100, \sigma^2 = 400)$. Die empirischen Beobachtungen sind annähernd normalverteilt, die standardisierten Beobachtungen annähernd standardnormalverteilt, $Z \sim \mathcal{N}(0, 1)$. Die Quantile der theoretischen Standardnormalverteilung würden hier also näherungsweise gelten.

Wir können die empirischen Quantile gegen die theoretischen Quantile einer Normalverteilung plotten (*Quantil-Quantil-Plot*) und so die Approximation visuell beurteilen (`qqnorm()`). Bei normalverteilten Daten geht dieser Plot in Richtung einer Geraden mit Achsenabschnitt μ und Steigung σ (`qqline()`).

```
hist(X, freq = FALSE)
curve(dnorm(x, mu, sigma), col = 2, lty = 2, add = TRUE)
qqnorm(X)
qqline(X)
```

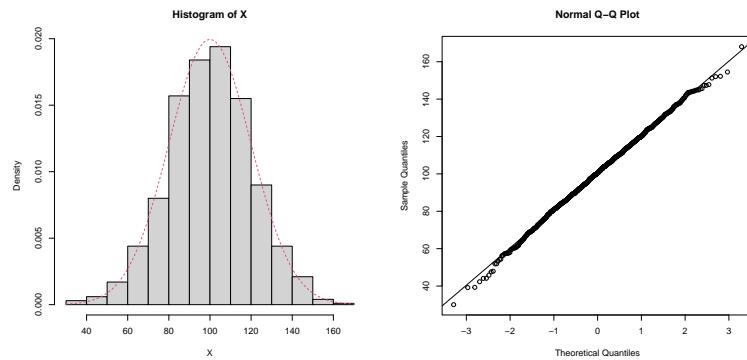


Abbildung 5.6: Normalverteilte Daten. Gestrichelt: theoretische Verteilung. Rechts: Quantil-Quantil-Plot

Die Variable `Alter` hingegen ist offensichtlich nicht normalverteilt (Abbildung 5.7):

```
hist(Alter)
qqnorm(Alter)
qqline(Alter)
```

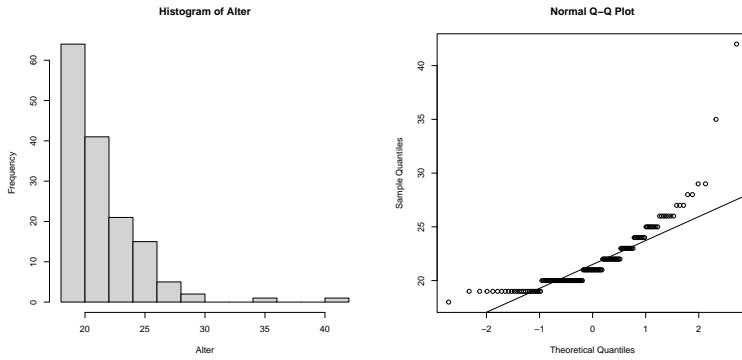


Abbildung 5.7: Nicht normalverteilte Variable

In Abbildung 5.8 sehen wir eine symmetrische, aber langschwänzige Verteilung (Daten aus einer t -Verteilung mit 3 Freiheitsgraden) mit dem entsprechenden Vergleich der empirischen Quantile gegen die theoretischen Quantile einer Normalverteilung.

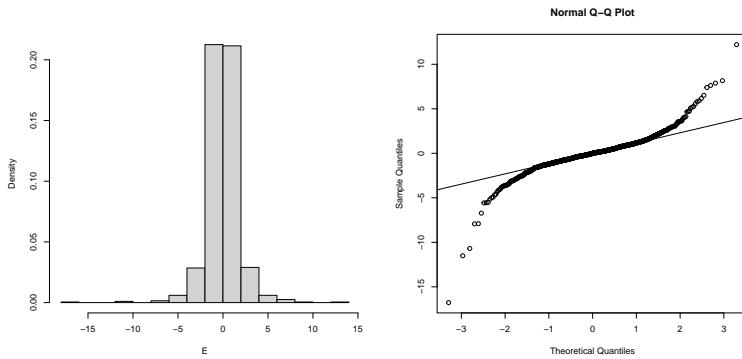


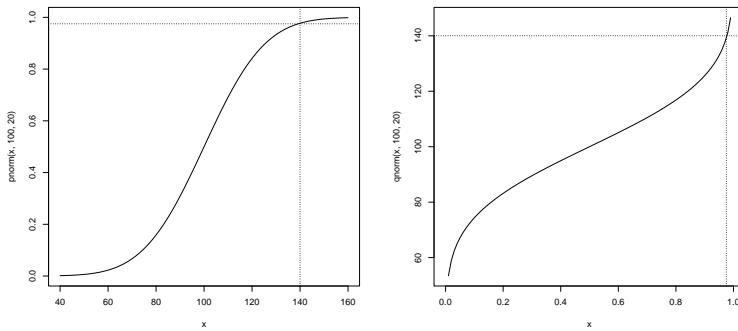
Abbildung 5.8: Nicht normalverteilte Daten: Sample versus theoretische Quantile einer Normalverteilung.

Beispiel. Gegeben seien folgende empirischen Kennwerte einer normalverteilten Variablen X : $\bar{x} = 100$, $s = 20$. Nehmen wir nun an, Hansi hat auf der Variable X den Wert 140. Dann ist der (bezüglich dieser Stichprobe) standardisierte Wert von Hansi gemäss (5.6.1) $z_{Hansi} = \frac{140-100}{20} = 2$. Hansi befindet sich also bezüglich der Stichprobe gerundet auf dem 0.975-Quantil der z -Verteilung. Das bedeutet, dass 2.5% der Werte grösser sind als der von Hansi und dass 97.5% der Werte kleiner sind als der von Hansi.

```
pnorm(q = 140, mean = 100, sd = 20)
pnorm(q = 2)  ##äquivalent, da z=(140-100)/20=2
```

Verteilungsfunktion und Quantilfunktion: Zur Wiederholung nochmals `pnorm()` und die Umkehrfunktion `qnorm()` mit Kennzeichnung bezogen auf obiges Beispiel (Code für die Interessierten).

```
curve(pnorm(x, 100, 20), from = 100 - 3 * 20, to = 100 + 3 * 20)
abline(h = 0.975, v = 100 + 2 * 20, lty = 3)
curve(qnorm(x, 100, 20), from = 0, to = 1)
abline(v = 0.975, h = 100 + 2 * 20, lty = 3)
```



5.7 Deskriptive Zusammenfassungen in R

Summaries. `summary()` ist eine generische Funktion, die je nach Objekt, das man als Argument übergibt, etwas anderes macht. Bei numerischen Vektoren sind das die deskriptiven Kennwerte:

```
summary(Alter)

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##     18.0    20.0    21.0    21.9    23.0    42.0
```

Zu empfehlen ist auch die `describe()`-Funktion aus der package `psych`. Diese package beinhaltet verschiedene Funktionen, die wir auch später noch brauchen werden.

Damit man mit zusätzlichen packages arbeiten kann, muss man diese package zuerst **installieren**, indem wir in die Konsole schreiben

```
install.packages("psych")
```

Wichtig: Diesen Befehl in die Konsole und nicht in ein Skript schreiben, die package muss nur einmal installiert werden. Wenn die package installiert ist, muss diese für den Gebrauch in der aktuellen R-Sitzung **geladen** werden mit

```
library(psych)
```

Am besten schreibt man das immer am Anfang vom Skript. Dann kann man eine gewünschte Funktion aus der package benutzen, z.B. `describe()`

```
describe(Alter)

##   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1     1 150 21.9 3.12      21    21.4 1.48  18   42    24 2.62     11.7 0.26
```

Eine Alternative, die **kein explizites Laden der package erforderlich** macht, ist der Gebrauch der `::` Funktion (double colon), mit der man dann Funktionen der package direkt brauchen kann:

```
psych::describe(Alter)
```

Wir lesen das Help-File für diese Funktion, um uns damit vertraut zu machen

```
help(describe)
```

Diese Funktion gibt zusätzliche Kennwerte heraus:

- *mad* steht für median absolute deviation
- *skew* steht *Schiefe* mit positiven Zahlen für rechtsschiefe und negativen Zahlen für linksschiefe Verteilungen. Die *kurtosis* ist ein Mass für den *Exzess* mit 0 für “normalgipflig”, positiv für “steilgipflig” und negativ für “flachgipflig”⁶
- *se* steht für Standardfehler, dazu kommen wir später zurück
- *trimmed* ist ein für Extremwerte korrigierter Mittelwert

Abbildung 5.9 zeigt drei Verteilungen mit positiver, keiner und negativer Schiefe.

```
hist(X1 <- rchisq(1000, 3), main = "positive Schiefe")
hist(X2 <- rnorm(1000), main = "keine Schiefe")
hist(X3 <- -rchisq(1000, 3), main = "negative Schiefe")
```

⁶Für die Interessierten: *skew* und *kurtosis* sind das dritte und vierte (normierte) *zentrale Moment*: Mittlere absolute Abweichung: $E(|X - \mu|)$, Varianz: $E((X - \mu)^2)$, Schiefe: $\frac{E((X - \mu)^3)}{\sigma^3}$, Exzess: $\frac{E((X - \mu)^4)}{\sigma^4} - 3$

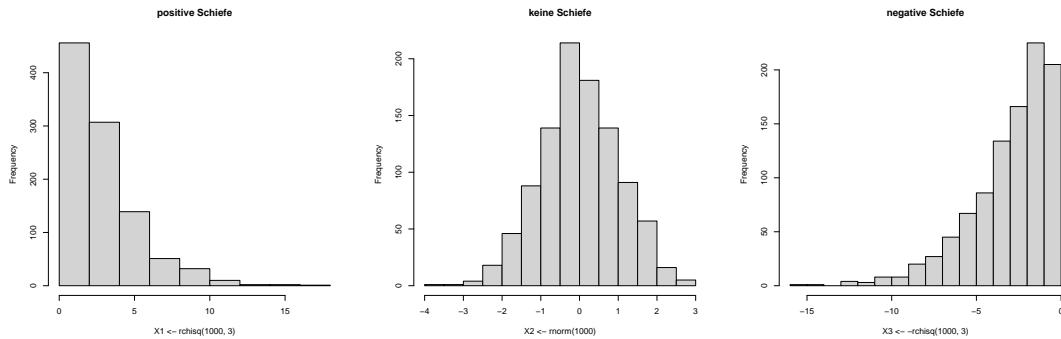


Abbildung 5.9: Verteilungen mit positiver, keiner und negativer Schiefe

Man kann als Argument in `describe()` (oder in `summary()`) auch ein data frame übergeben. Das wird sich später als sehr nützlich erweisen. Vergleiche die drei Schiefen.

```
mydatframe <- data.frame(X1, X2, X3)
describe(mydatframe)

##      vars     n   mean    sd median trimmed  mad    min   max range skew kurtosis    se
## X1      1 1000  2.88  2.36   2.23   2.52 1.84   0.01 16.63 16.62  1.65    3.79  0.07
## X2      2 1000  0.01  1.01  -0.02   0.01 0.99  -3.53  2.99  6.52 -0.03   -0.07  0.03
## X3      3 1000 -3.01  2.41  -2.38  -2.69 2.07 -15.38 -0.03 15.34 -1.36    2.21  0.08
```

Wenn man noch die Quartile will, bietet sich wieder `summary()` an.

```
summary(mydatframe)

##          X1              X2              X3
##  Min.   : 0.0136   Min.   :-3.53156   Min.   :-15.3768
##  1st Qu.: 1.2295   1st Qu.:-0.65136   1st Qu.:-4.2135
##  Median : 2.2337   Median :-0.01667   Median :-2.3824
##  Mean   : 2.8765   Mean   : 0.00681   Mean   : -3.0145
##  3rd Qu.: 3.8810   3rd Qu.: 0.69790   3rd Qu.: -1.1977
##  Max.   :16.6344   Max.   : 2.98997   Max.   : -0.0333
```

Übung: Häufige deskriptive Analysen, klassischer Workflow. Wir lesen Daten ein aus einer Textdatei *Davis.csv*. Das Dateiformat *.csv* steht wieder für *comma-separated values* und beschreibt den Aufbau einer einfachen Textdatei zur Speicherung oder zum Austausch einfacher strukturierter Daten. Die Datei ist abgelegt unter

<https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/Davis.csv>.

Diese kann man in R einlesen und als Objekt abspeichern mit

`read.csv(..., stringsAsFactors=TRUE)`

oder alternativ mit dem allgemeineren

```
read.table(...,header=TRUE,sep=",",stringsAsFactors=TRUE).
```

Das Argument **stringsAsFactors=TRUE** muss ab **R-Version 4.0 neu** gesetzt werden, damit character-Vektoren direkt in Faktoren umgewandelt werden. In **read.csv()** sind die beiden letzten Argumente der zweiten Funktion als Default eingestellt (dass die erste Datenzeile eine Kopfzeile darstellt und dass die Daten durch Kommas getrennt sind).

```
d.dav<-read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/Davis.csv",stringsAsFactors=TRUE)
```

d.dav ist nun ein data frame mit $n = 200$ Beobachtungen einer kategorialen Variable **sex** (“Faktor” in R) und 4 kontinuierlichen, diskret gemessenen Variablen (**weight**, **height**, **repwt**, **reph**) (ganzzahlig, “integer”). Diese Information erhalten wir über **str()**.

```
str(d.dav)

## 'data.frame': 200 obs. of 5 variables:
## $ sex : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
## $ weight: int 77 58 53 68 59 76 76 69 71 65 ...
## $ height: int 182 161 161 177 157 170 167 186 178 171 ...
## $ repwt : int 77 51 54 70 59 76 77 73 71 64 ...
## $ repht : int 180 159 158 175 155 165 165 180 175 170 ...
```

Bei data frames mit vielen Zeilen ist die **head()** Funktion sinnvoll, um die ersten Zeilen anzuschauen.

```
head(d.dav)

##   sex weight height repwt repht
## 1   M     77    182     77    180
## 2   F     58    161     51    159
## 3   F     53    161     54    158
## 4   M     68    177     70    175
## 5   F     59    157     59    155
## 6   M     76    170     76    165
```

Eine der wichtigsten Funktionen ist die schon bekannte **summary()**-Funktion, die man auf viele Objekte anwenden kann. Bei data frames bekommen wir

```
summary(d.dav)

##   sex      weight      height      repwt      repht
##   F:112  Min.   :39.0  Min.   :148  Min.   :41.0  Min.   :148
##   M: 88  1st Qu.:55.0  1st Qu.:164  1st Qu.:55.0  1st Qu.:160
##             Median :63.0  Median :170  Median :63.0  Median :168
##             Mean   :65.3  Mean   :171  Mean   :65.6  Mean   :168
##             3rd Qu.:73.2  3rd Qu.:177  3rd Qu.:73.5  3rd Qu.:175
##             Max.   :119.0  Max.   :197  Max.   :124.0  Max.   :200
##                               NA's   :17    NA's   :17
```

`summary()` erkennt kategoriale Variablen (weil Zeichen (character) statt Zahlen (integer oder numeric)) und wird dort statt `mean()`, `median()` usw. die Häufigkeiten angeben.

Mit dem \$-Zeichen kann man aus einer Liste (und ein data frame ist ein Spezialfall davon), einzelne Variablen anwählen, z.B. `d.dav$sex`. Eine wichtige Funktion, um Häufigkeiten auszuzählen, ist `table()` und `proportions()`.

```
table(d.dav$sex)

##
##   F     M
## 112   88

proportions(table(d.dav$sex))

##
##   F     M
## 0.56 0.44
```

Wenn man `summary()` ein data frame übergibt, dann wird für kategoriale Variablen direkt `table()` angewandt.

Die `describe()` Funktion (aus `psych`) kann man auch auf data frames anwenden

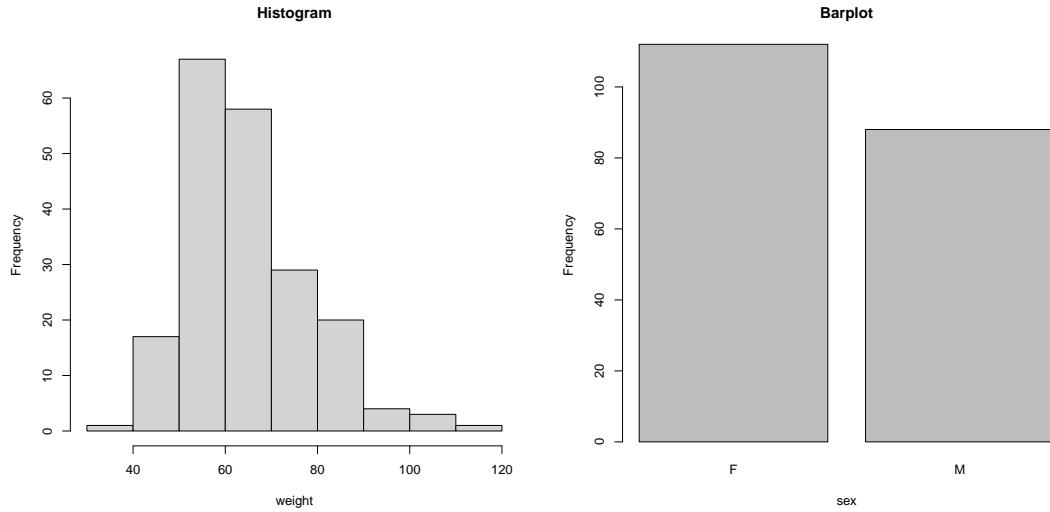
```
describe(d.dav)

##      vars   n   mean      sd median trimmed   mad min max range skew kurtosis    se
## sex*     1 200  1.44  0.50       1    1.43  0.00    1   2     1 0.24 -1.95  0.04
## weight   2 200 65.25 13.32      63   64.04 11.86   39 119    80 0.91  0.86  0.94
## height   3 200 170.56  8.93     170  170.36  9.64 148 197    49 0.22 -0.37  0.63
## repwt    4 183  65.62 13.78      63   64.27 11.86   41 124    83 1.03  1.33  1.02
## reptht   5 183 168.50  9.47     168  168.19 10.38 148 200    52 0.33 -0.36  0.70
```

`describe()` macht ein Sternchen bei kategorialen Variablen, da dort `mean()` usw. eigentlich keinen Sinn machen.

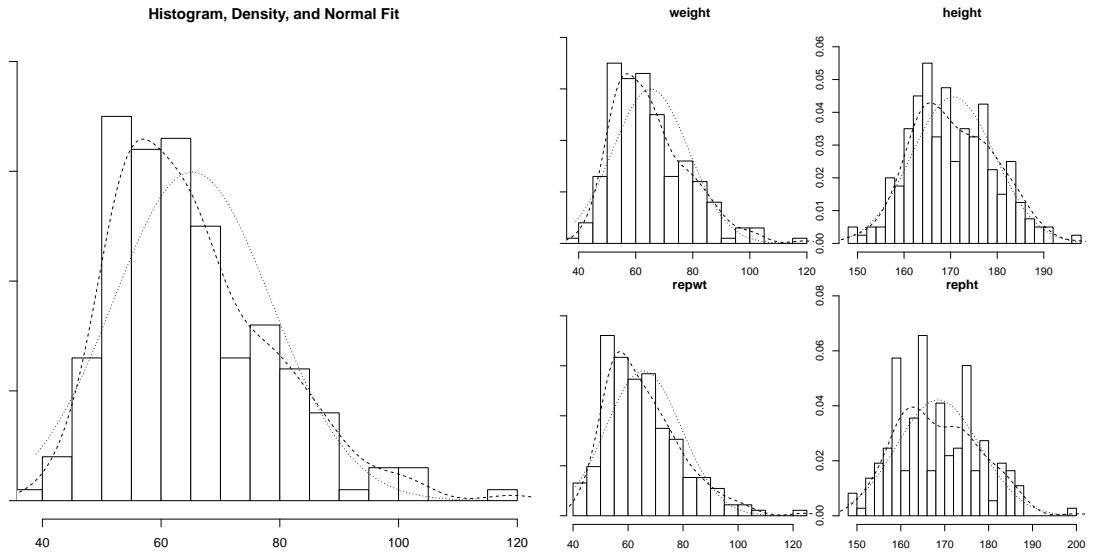
Histogramme und Barplots machen wir mit

```
hist(d.dav$weight, xlab = "weight", main = "Histogram")
barplot(table(d.dav$sex), xlab = "sex", ylab = "Frequency", main = "Barplot")
```



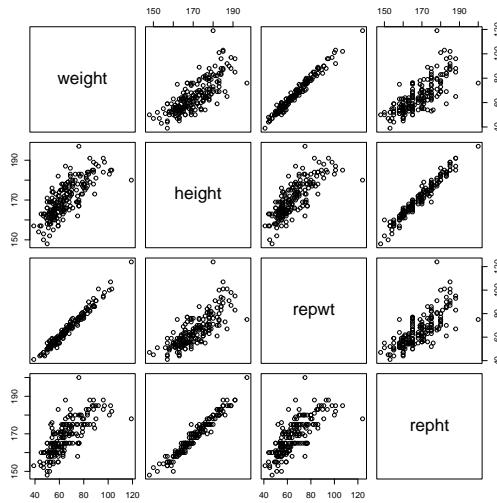
Es gibt immer auch Alternativen, z.B. mit `psych::multi.hist()` und `psych::histBy()`. Mit `multi.hist()` wird zusätzlich eine Dichteschätzung und ein Normalverteilungsmodell über das Histogramm aufgetragen.

```
multi.hist(d.dav$weight) #nur die zweite Variable
multi.hist(d.dav[, -1], global = FALSE) #alle ohne die erste Variable (Faktor Sex)
```



Mit `pairs()` können auch alle *bivariate* Verteilungen anschauen (dazu mehr im nächsten Kapitel).

```
pairs(d.dav[, -1])
```



Es ist in R einfach, neue Variablen zu kreieren oder eine alte zu transformieren. So können wir den BMI berechnen mit

```
d.dav$height <- d.dav$height/100 ## überschreiben der Variablen, in Meter
d.dav$replt <- d.dav$replt/100
d.dav$BMI <- (d.dav$weight)/(d.dav$height)^2
str(d.dav)

## 'data.frame': 200 obs. of  6 variables:
## $ sex      : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
## $ weight    : int  77 58 53 68 59 76 76 69 71 65 ...
## $ height   : num  1.82 1.61 1.61 1.77 1.57 1.7 1.67 1.86 1.78 1.71 ...
## $ repwt    : int  77 51 54 70 59 76 77 73 71 64 ...
## $ rept     : num  1.8 1.59 1.58 1.75 1.55 1.65 1.65 1.8 1.75 1.7 ...
## $ BMI      : num  23.2 22.4 20.4 21.7 23.9 ...
```

Zusätzlich kreieren wir eine kategoriale Variable aus dem kontinuierlichen BMI. Mit `cut()` kann man mit dem Argument `breaks=` Schwellen angeben.

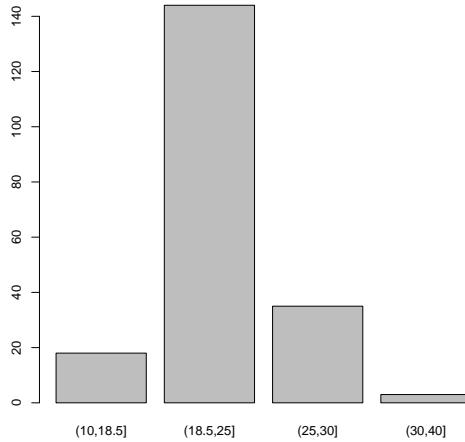
```
d.dav$BMICat <- cut(d.dav$BMI, breaks = c(10, 18.5, 25, 30, 40))
str(d.dav)

## 'data.frame': 200 obs. of  7 variables:
## $ sex      : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
## $ weight    : int  77 58 53 68 59 76 76 69 71 65 ...
## $ height   : num  1.82 1.61 1.61 1.77 1.57 1.7 1.67 1.86 1.78 1.71 ...
## $ repwt    : int  77 51 54 70 59 76 77 73 71 64 ...
## $ rept     : num  1.8 1.59 1.58 1.75 1.55 1.65 1.65 1.8 1.75 1.7 ...
## $ BMI      : num  23.2 22.4 20.4 21.7 23.9 ...
## $ BMICat: Factor w/ 4 levels "(10,18.5]","(18.5,25]",...: 2 2 2 2 2 3 3 2 2 2 ...

table(d.dav$BMICat)

##
## (10,18.5] (18.5,25] (25,30] (30,40]
##          18        144       35       3
```

```
plot(d.dav$BMICat)
```



Mit `levels()` kann man die Levels noch labeln.

```
levels(d.dav$BMICat) <- c("underweight", "normal", "overweight", "obese")
table(d.dav$BMICat)

## 
##   underweight      normal   overweight       obese
##           18          144          35            3
```

Dasselbe hätten wir bekommen, wenn wir die labels direkt in `cut()` gesetzt hätten.

```
d.dav$BMICat <- cut(d.dav$BMI, breaks = c(10, 18.5, 25, 30, 40), labels = c("underweight", "normal",
"overweight", "obese"))
```

Mit `by()` kann man eine Funktion auf ein data frame, *gesplittet bezüglich Faktoren*, anwenden.

```
by(d.dav, d.dav$sex, summary)

## d.dav$sex: F
##   sex      weight      height      repwt      repht      BMI
##   F:112  Min.   :39.0   Min.   :1.48   Min.   :41.0   Min.   :1.48   Min.   :15.8
##   M:  O  1st Qu.:52.8   1st Qu.:1.62   1st Qu.:53.0   1st Qu.:1.59   1st Qu.:19.7
##          Median :56.0   Median :1.65   Median :56.0   Median :1.61   Median :20.6
##          Mean   :56.9   Mean   :1.65   Mean   :56.7   Mean   :1.62   Mean   :21.0
##          3rd Qu.:62.0   3rd Qu.:1.69   3rd Qu.:61.0   3rd Qu.:1.65   3rd Qu.:22.3
##          Max.   :78.0   Max.   :1.78   Max.   :77.0   Max.   :1.76   Max.   :28.6
##          NA's    :11     NA's    :11     NA's    :11     NA's    :11
## 
##   BMICat
##   underweight:17
##   normal     :91
```

```

##  overweight : 4
##  obese      : 0
##
##
##
## -----
## d.dav$sex: M
##   sex       weight      height      repwt      repht        BMI
##   F: 0    Min.   :54.0   Min.   :1.63   Min.   :56.0   Min.   :1.61   Min.   :17.8
##   M:88   1st Qu.:67.8   1st Qu.:1.73   1st Qu.:68.0   1st Qu.:1.71   1st Qu.:21.7
##   Median  :75.0   Median :1.78   Median :75.0   Median :1.75   Median :23.5
##   Mean    :75.9   Mean   :1.78   Mean   :76.6   Mean   :1.76   Mean   :23.9
##   3rd Qu.:83.0   3rd Qu.:1.83   3rd Qu.:83.0   3rd Qu.:1.80   3rd Qu.:25.8
##   Max.    :119.0  Max.   :1.97   Max.   :124.0  Max.   :2.00   Max.   :36.7
##   NA's    :6        NA's   :6        NA's   :6
##
##   BMIcat
##   underweight: 1
##   normal     :53
##   overweight :31
##   obese      : 3
##
##
## 
```

by(d.dav, d.dav\$sex, describe)

```

## d.dav$sex: F
##   vars n mean sd median trimmed mad min max range skew kurtosis se
##   sex*   1 112 1.00 0.00 1.00 1.00 0.00 1.00 1.00 0.00 NaN  NaN 0.00
##   weight 2 112 56.89 6.86 56.00 56.67 5.93 39.00 78.00 39.00 0.41 0.56 0.65
##   height 3 112 1.65 0.06 1.65 1.65 0.06 1.48 1.78 0.30 -0.23 0.15 0.01
##   repwt  4 101 56.74 6.74 56.00 56.52 5.93 41.00 77.00 36.00 0.45 0.57 0.67
##   repht  5 101 1.62 0.06 1.61 1.62 0.06 1.48 1.76 0.28 0.10 -0.28 0.01
##   BMI    6 112 20.95 2.17 20.63 20.88 1.85 15.82 28.58 12.76 0.52 0.87 0.20
##   BMIcat* 7 112 1.88 0.42 2.00 1.93 0.00 1.00 3.00 2.00 -0.73 1.81 0.04
## -----
## d.dav$sex: M
##   vars n mean sd median trimmed mad min max range skew kurtosis se
##   sex*   1 88 2.00 0.00 2.00 2.00 0.00 2.00 2.00 0.00 NaN  NaN 0.00
##   weight 2 88 75.90 11.89 75.00 75.10 11.86 54.00 119.00 65.00 0.78 0.88 1.27
##   height 3 88 1.78 0.06 1.78 1.78 0.07 1.63 1.97 0.34 0.15 -0.09 0.01
##   repwt  4 82 76.56 12.29 75.00 75.44 10.38 56.00 124.00 68.00 1.06 1.57 1.36
##   repht  5 82 1.76 0.07 1.75 1.76 0.07 1.61 2.00 0.39 0.30 0.35 0.01
##   BMI    6 88 23.90 3.12 23.53 23.75 3.27 17.81 36.73 18.92 0.78 1.81 0.33
##   BMIcat* 7 88 2.41 0.58 2.00 2.36 0.00 1.00 4.00 3.00 0.70 -0.15 0.06

```

Man kann auch mit `split()` direkt ein Objekt (Liste mit zwei data frames) kreieren

```

subgr <- split(d.dav, f = d.dav$sex)
str(subgr)

```

und mit `lapply()` können wir eine Funktion *über eine Liste* ausführen

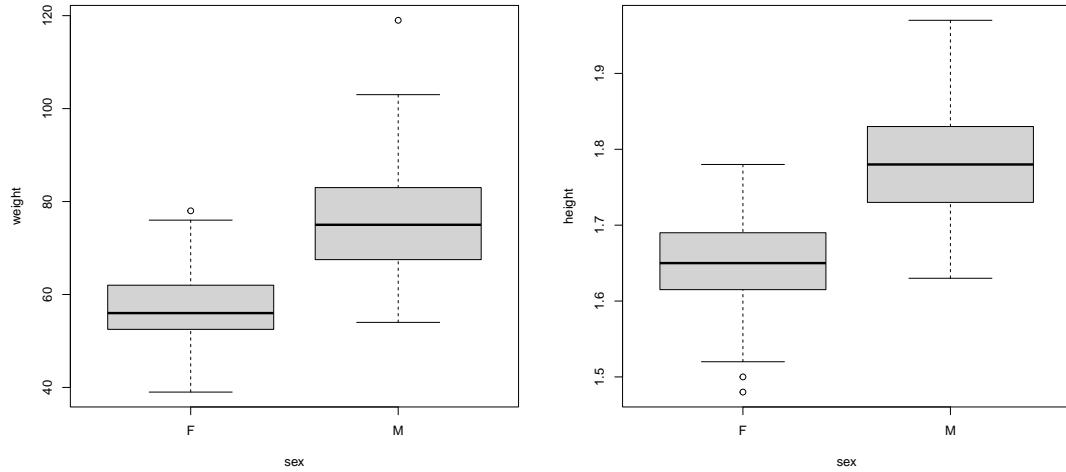
```

lapply(subgr, summary)
lapply(subgr, describe)

```

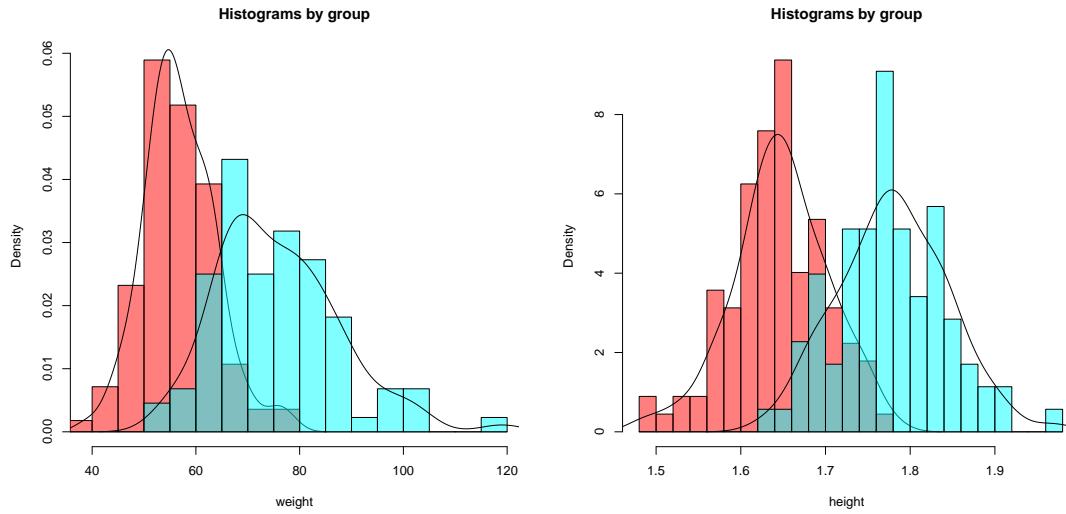
Boxplots sind dann auch gesplittet bezüglich einem Faktor möglich. Dazu brauchen wir die **formula**-Schreibweise mit dem \sim -Zeichen: Abhängige Variable \sim unabhängige Variable.

```
boxplot(weight ~ sex, d.dav)
boxplot(height ~ sex, d.dav)
```



Mit **psych::histBy()** wird ein Histogramm pro Gruppe gemacht.

```
histBy(weight ~ sex, data = d.dav)
histBy(height ~ sex, data = d.dav)
```



Übung: Reproduzieren von deskriptiven Resultaten. Konsultiere die Tabelle 1 der folgenden Studie an:

<https://hqlo.biomedcentral.com/articles/10.1186/s12955-020-01576-w/tables/1>

an. Lese dazu folgende Datei mit einem Teil der involvierten Variablen ein und reproduziere die entsprechenden Kennwerte in der Tabelle. Brauche dazu die Funktion `by()` zusammen mit `summary()` und/oder `describe()`.

```
d.brud <- read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/table1Brud.csv",
  stringsAsFactors = TRUE)
str(d.brud)

## 'data.frame': 96 obs. of 7 variables:
## $ ageP      : int 74 69 81 88 69 83 79 72 86 82 ...
## $ ageI      : int 67 34 85 62 37 81 74 36 40 45 ...
## $ sexP      : int 2 1 2 1 2 2 2 2 2 ...
## $ sexI      : int 1 1 1 1 1 1 2 1 ...
## $ diagnosis: int 3 2 3 2 2 3 2 3 3 3 ...
## $ Tscore    : num 57.7 61.8 60.2 43.2 63.8 ...
## $ group     : Factor w/ 2 levels "H","P": 2 2 2 2 2 2 2 2 2 ...
```

Ein wichtiger Schritt ist immer die Kontrolle, ob R Faktoren als solche erkennt, sonst muss man das manuell machen. `sexP`, `sexI` und `diagnosis` sind offensichtlich Faktoren.

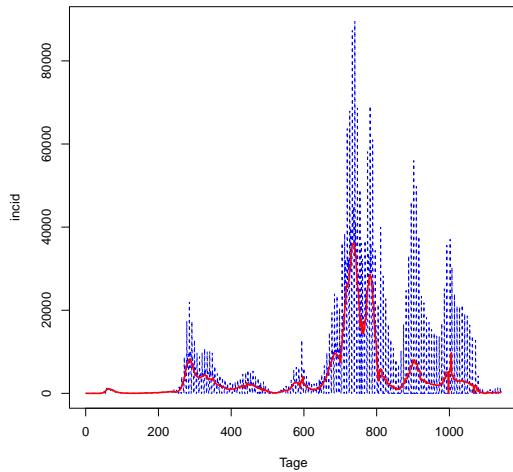
```
d.brud$sexP <- as.factor(d.brud$sexP)
d.brud$sexI <- as.factor(d.brud$sexI)
d.brud$diagnosis <- as.factor(d.brud$diagnosis)
str(d.brud)

## 'data.frame': 96 obs. of 7 variables:
## $ ageP      : int 74 69 81 88 69 83 79 72 86 82 ...
## $ ageI      : int 67 34 85 62 37 81 74 36 40 45 ...
## $ sexP      : Factor w/ 2 levels "1","2": 2 1 2 1 2 2 2 2 2 ...
## $ sexI      : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 2 1 ...
## $ diagnosis: Factor w/ 3 levels "1","2","3": 3 2 3 2 2 3 2 3 3 3 ...
## $ Tscore    : num 57.7 61.8 60.2 43.2 63.8 ...
## $ group     : Factor w/ 2 levels "H","P": 2 2 2 2 2 2 2 2 2 ...
```

Darstellung einer Zeitreihe*. Folgender Code aktualisiert die aktuellen Fallzahlen von Covid-19 in der Schweiz. Die Daten bestehen aus einem *Zeitreihen*-Objekt. Wir stellen die kumulativen Häufigkeiten und deren Veränderung (Inzidenz) dar.

```
myurl<-paste(
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/",
  "time_series_covid19_confirmed_global.csv",sep="")
```

```
data <- read.csv(myurl, stringsAsFactors = TRUE)
swissCovid <- data[data$Country.Region == "Switzerland", -c(1, 2, 3, 4)]
cases <- as.ts(as.numeric(swissCovid)) #times series object
incid <- diff(cases) #Zuwachs
t <- 1:length(incid)
incidAv <- broman::runningmean(t, incid, window = 7, what = "mean") #7-Tage-Schnitt
plot(t, incid, type = "l", col = "blue", lty = 2, xlab = "Tage")
lines(t, incidAv, col = "red", lwd = 2)
```



Kapitel 6

Multivariate Beschreibung

Wir nennen statistische Methoden, die nach Zusammenhängen in Daten suchen, *explorative* Methoden. Explorative Methoden gehen über rein deskriptive Methoden hinaus. Explorative Methoden gehören aber noch nicht zur *induktiven* Statistik. Es geht noch nicht um die Verallgemeinerung von Stichprobenresultaten, sondern um die Suche nach Strukturen und Besonderheiten in den Daten. Die Methoden der explorativen Datenanalyse werden zudem häufig eingesetzt, wenn die Fragestellung nicht genau definiert ist, z.B. bei Fragen, die sich *a posteriori*, also nach der Datensammlung ergeben.

Eine einfache Exploration ist die Suche nach *Zusammenhängen* oder *Korrelationen* zwischen zwei Variablen X und Y .

Wir beginnen also die *multivariate Betrachtung* mit dem *zweidimensionalen Fall*.

6.1 Empirische Kovarianz und Korrelation

Wir haben uns bis jetzt mit der univariaten Verteilung *einer* Variablen X befasst. Wir können uns auch zwei Variablen X und Y *gemeinsam* anschauen. X sei die Variable *Alkoholkonzentration* mit den Realisationen x_1, x_2, \dots, x_n . Y sei die Variable *Reaktionszeit* mit den Realisationen y_1, y_2, \dots, y_n . Die Stichprobengröße sei $n = 9$. Die Daten stammen also aus einer Querschnittsstudie; die Rohdaten sind in Tabelle 6.1 dargestellt.

ID	Alc(Promille)	Rct(ms)
1	0.00	554
2	0.20	581
3	0.50	589
4	0.70	628
5	1.00	623
6	1.40	687
7	1.80	692
8	2.25	734
9	2.50	812

Tabelle 6.1: Messwerte auf Alkoholkonzentration (Promille) und Reaktionszeit (ms).

Reproduzieren wir diese Daten mit R:

```
A <- c(0, 0.2, 0.5, 0.7, 1, 1.4, 1.8, 2.25, 2.5)
R <- c(554, 581, 589, 628, 623, 687, 692, 734, 812)
ARdata <- data.frame(Alkohol = A, Reaktionszeit = R)
```

Die univariante Betrachtung von X und Y gibt folgende Kennwerte für die beiden Variablen: $\bar{x} = 1.15$ Promille, $s_x = 0.894$ Promille, $\bar{y} = 655.556$ ms, $s_y = 82.966$ ms.

```
psych::describe(ARdata)

##           vars n   mean     sd median trimmed   mad min    max range skew kurtosis    se
## Alkohol      1 9  1.15  0.89      1  1.15  1.19  0  2.5  2.5 0.21 -1.65  0.3
## Reaktionszeit 2 9 655.56 82.97    628 655.56 87.47 554 812.0 258.0 0.50 -1.13 27.7
```

Wollen wir die *gemeinsame* Verteilung dieser beiden Variablen anschauen, machen wir dies über ein sogenanntes *Streudiagramm*. Abbildung 6.1 zeigt das Streudiagramm für diese beiden Variablen (für die bivariate Verteilung). Die Werte der Variablen X sind horizontal und die Werte der Variablen Y sind vertikal aufgetragen.

```
plot(A, R, xlab = "Alkoholkonzentration [Promille]", ylab = "Reaktionszeit [ms]")
```

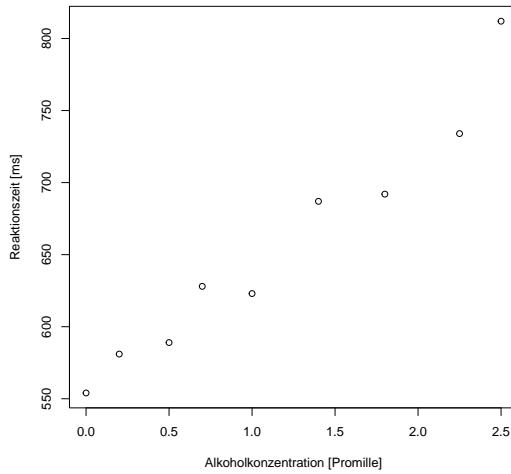


Abbildung 6.1: Streudiagramm

Man sieht, dass die Reaktionszeit im Schnitt grösser wird, wenn die Alkoholkonzentration steigt. Man sagt dann, dass die Variablen *kovariieren*. Dies wird dann über die Grösse der (aus den Daten geschätzten) *Kovarianz* zwischen den Variablen X und Y quantifiziert. Wir notieren dies mit $\text{Cov}(X, Y)$. Wir haben diese Grösse eingeführt in 4.8.1. Die Kovarianz ist eine Verallgemeinerung der Varianz auf den mehrdimensionalen Fall. Die empirische Kovarianz $\widehat{\text{Cov}}(X, Y)$ (die, die man aus den Daten berechnen kann), ist

$$\widehat{\text{Cov}}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}. \quad (6.1.1)$$

In unserem Fall ist die empirische Kovarianz $\widehat{\text{Cov}}(X, Y) = 72.15$.

```
cov(A, R)
## [1] 72.2
## sum((A-mean(A))*(R-mean(R)))/(length(A)-1) #Alternative 'by hand'
```

Korrelationskoeffizient. Die Grösse der Kovarianz ist schwierig zu interpretieren. Darum *standardisieren* wir diese Grösse. Dies tun wir, indem wir sie durch die beiden Standardabweichungen s_x und s_y teilen. Dies ergibt den *Korrelationskoeffizienten* nach *Pearson*, $\text{Corr}(X, Y)$. Diese Grösse haben wir eingeführt in 4.8.2. Aus den Daten wird

diese Grösse geschätzt mit der empirischen Korrelation,

$$\widehat{\text{Corr}}(X, Y) = \frac{\widehat{\text{Cov}}(X, Y)}{s_X \cdot s_Y}. \quad (6.1.2)$$

Man schreibt für die *empirische* Korrelation oft kurz r und für die *wahre* Korrelation kurz ρ .

Der Betrag von r ist wie bei ρ durch 1 beschränkt ($|r| \leq 1$). Die Pearson-Korrelation stellt ein Mass dar für Stärke des linearen Zusammenhangs. Das Vorzeichen gibt an, ob die Korrelation positiv oder negativ ist, gibt also die *Richtung* des Zusammenhangs an. Ist die Zahl positiv, bedeutet dies: je grösser X , desto grösser ist im Schnitt Y ; ist die Zahl negativ, dementsprechend: je grösser X , desto kleiner ist im Schnitt Y . Der Betrag der Korrelation gibt die *Stärke* des Zusammenhangs zwischen den beiden Variablen an. In unserem Beispiel ist die empirische Korrelation

$$r_{X,Y} = \frac{72.15}{0.894 \cdot 82.966} = 0.973$$

Diese beobachtete (empirische) Korrelation ist wie erwartet positiv und betragsmässig sehr gross, da nahe bei 1.

```
cor(A, R)
## [1] 0.973
# cov(A,R)/(sd(A)*sd(R))
```

Abbildung 6.2 zeigt drei verschiedene bivariate Verteilungen und ihre Korrelationen.

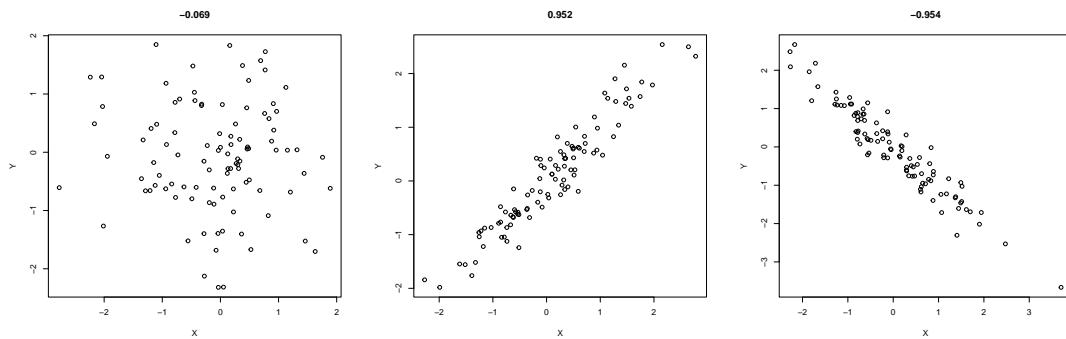


Abbildung 6.2: Korrelationen

Andere Korrelationstechniken* Den Pearson-Korrelationskoeffizienten (Mass für den *linearen* Zusammenhang) haben wir bereits eingeführt. Tabelle 6.2 zeigt

Korrelationstechniken auf, wenn eines von beiden Merkmalen nicht intervallskaliert ist.

Y / X	intervall	ordinal	dichotom künstlich	dichotom natürlich
Intervall	r_P	r_S , Kendall polychorisch	punktbiserial biserial	punktbis. punktbis.
Ordinal		r_S , Kendall polychorisch	biserial Rang polychorisch	biserial Rang biserial
Dichotom künstlich			ϕ -Koef., tetrachorisch	ϕ -Koef.
Dichotom natürlich				ϕ -Koef.

Tabelle 6.2: Korrelationstechniken: r_P : Pearson, r_S : Spearman, ϕ : Phi-Koeffizient.

Biseriale Korrelationen kommen zur Anwendung, wenn ein Merkmal intervall- oder ordinalskaliert und das zweite Merkmal dichotom nominalskaliert ist. Für das nominalskalierte Merkmal unterscheidet man noch zwischen 1) Echt dichotome Variable: natürlich vorkommende Gruppenteilung wie z.B. männlich/weiblich, etc. Der Zusammenhang einer solchen mit einer intervallskalierten Variablen wird durch die punktbiserialen Korrelation beschrieben. Das ist ein Spezialfall der Pearson-Korrelation, `cor()`. 2) Künstlich dichotome Variable: wird eine kontinuierliche Variable in zwei Gruppen aufgeteilt, dann spricht man von einer künstlich dichotomen Variablen. Zusammenhänge dieser mit einer intervallskalierten Variablen werden durch die biseriale Korrelation beschrieben, `psych::biserial()`.

Der ϕ -Koeffizient ist ein Mass für den Zusammenhang zwischen zwei dichotomen Merkmalen. Auch er ist ein Spezialfall der Pearson-Korrelation, `cor()`. Soll der Zusammenhang zwischen zwei künstlich-dichotomen Variablen berechnet werden, die aus stetigen, normalverteilten latenten Variablen abgeleitet wurden (z.B. Intelligenz und Leistung in Mathematik), verwendet man die tetrachorische Korrelation, `psych::tetrachoric()`.

Rangkorrelationen: 1) Die Spearman-Rangkorrelation ρ_S ist nichts anderes als die Pearson Korrelation angewendet auf rangtransformierte Daten (über `rank()`), in R mit `cor(X, Y, method='spearman')`. 2) Kendall τ : Ränge müssen nicht wie bei der Spearman-Korrelation gleichabständig sein, `cor(X, Y, method='kendall')`. 3) Bei der polychorischen Korrelation, `psych::polychoric()`, sind die manifesten ordinalen Variablen Kategorisierungen von unterliegenden latenten normalverteilten Variablen.

Wir lesen nochmals *Davis.csv* ein:

```
d.dav <- read.csv("https://raw.githubusercontent.com/mcdrl65/StatsResource/master/Data/Davis.csv", stringsAsFactors = TRUE)

cor(as.numeric(d.dav$sex), d.dav$weight) ## point-biserial, braucht 'as.numeric', da sex Faktor!
cor(d.dav$weight, d.dav$height, method = "pearson") ##pearson
cor(d.dav$weight, d.dav$height) ##pearson ist default
cor(d.dav$weight, d.dav$height, method = "spearman") ##spearman
cor(d.dav$weight, d.dav$height, method = "kendall") ##kendall
cor(rank(d.dav$weight), rank(d.dav$height)) ##=spearman
```

Korrelation versus Kausalität. (Pearson) Korrelationen quantifizieren den (linearen) Zusammenhang zwischen zwei Merkmalen. Dies bedeutet aber nicht, dass es einen *kausalen* Zusammenhang gibt zwischen den beiden Merkmalen, dass ein Merkmal eine Ursache-Wirkungs-Beziehung hat mit dem zweiten Merkmal. Es ist immer möglich, dass eine dritte Variable – eine *Kovariable* – beide Merkmale so beeinflusst, dass diese beiden Merkmale statistisch korrelieren, obwohl sie nicht kausal miteinander zusammenhängen. Ein solcher *Confounder* ist also eine Variable, die mit der unabhängigen und der abhängigen Variable korreliert.

Zusammenhänge beschreiben also zuerst nur Assoziationen und nicht Kausalität! Dies ist eine der wichtigsten Aspekte in der Wissenschaft und bei der Auseinandersetzung mit Wissenschaft. Allzu oft werden Korrelationen überinterpretiert und es wird fälschlicherweise Kausalität suggeriert, wo keine ist. Die Abbildung 6.3 zeigt die Problematik auf. Der Effekt von X (unabhängige Variable) auf Y (abhängige Variable) wird durch C konfundiert. Eine Variable C ist ein Confounder, wenn sie mit X assoziiert ist (ohne eine Wirkung von X zu sein) und zusammenhängt mit Y (unabhängig von X).

Bei einer randomisierten kontrollierten Studie sind alle potentiellen C 's keine Störgrößen mehr, da C dann von der Intervention X unabhängig ist. In Beobachtungsstudien aber müssen wir z.B. in der Analyse für C kontrollieren.

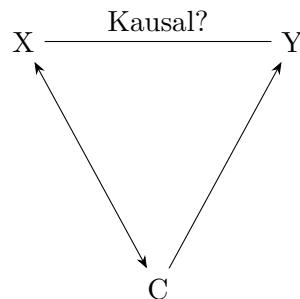


Abbildung 6.3: Potentielle Ursache X , Wirkung Y und Störgrösse C .

In Abbildung 6.4 ist Rauchen als ein potentieller Confounder für den Effekt von Kaffeekonsum auf Krebs dargestellt. Eine diesbezügliche Studie (prospektive Kohorte, retrospektive Fall-Kontroll-Studie oder Querschnittsstudie) muss also für Rauchen kontrollieren (*a priori* über das *Design* (Selektion, Inklusion, Exklusion, Randomisation, Matching) oder *a posteriori* über die *Analyse* (Stratifikation oder Analyse, wo man Rauchen als Kovariable in einem Modell berücksichtigt)).

Korrelationsmatrix. Wir können uns Korrelationen anschauen zwischen mehr als zwei Variablen. Als Beispiel betrachten wir Daten aus dem Jahr 1888. Es ist ein Datensatz, der in R schon als Objekt verfügbar ist. Das data frame heisst **swiss**.

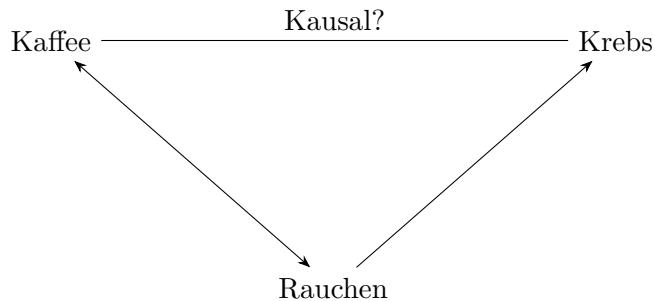


Abbildung 6.4: Rauchen als Confounder für den Effekt von Kaffee auf Krebs

```

str(swiss)
head(swiss)
swiss
  
```

Es sind Daten zu einer standardisierten Messung von **Fertilität** und von sozio-ökonomischen Indikatoren in 47 Regionen der Westschweiz um 1888. Um 1888 fiel die Fertilität der Schweiz erstmals unter den für unterentwickelte Länder typischen hohen Level. Die statistische Einheit ist hier die Region. Alle Variablen ausser **Fertilität** geben Anteile in Prozent in der Population an, siehe `help(swiss)`. Abbildung 6.5 zeigt alle paarweisen bivariaten Verteilungen als Streudiagramme. Jede Variable ist gegen jede andere aufgetragen. Das gibt $(6 \times 6 - 6)/2 = 15$ Streudiagramme. Wir haben diese Funktion `pairs()` bereits angetroffen. Eine Alternative ist `pairs.panels()` aus `psych`, diese sehr nützliche Funktion gibt oberhalb der Diagonalen die Korrelationen zurück und in der Diagonale die univariaten Histogramme.

`cor()` übernimmt auch ein data frame als Argument. Wir untersuchen explorativ Pearson-Korrelationen mit

```

cor(swiss)
  
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
## Fertility	1.000	0.3531	-0.646	-0.6638	0.464	0.4166
## Agriculture	0.353	1.0000	-0.687	-0.6395	0.401	-0.0609
## Examination	-0.646	-0.6865	1.000	0.6984	-0.573	-0.1140
## Education	-0.664	-0.6395	0.698	1.0000	-0.154	-0.0993
## Catholic	0.464	0.4011	-0.573	-0.1539	1.000	0.1755
## Infant.Mortality	0.417	-0.0609	-0.114	-0.0993	0.175	1.0000

6.2 Lineare Regression

Eine (Pearson)-Korrelation ist ein Mass für die Stärke des (linearen) Zusammenhangs zwischen zwei Variablen X und Y . Wenn wir wissen wollen, wie die *Art* des

```
pairs(swiss)
psych::pairs.panels(swiss, smooth = FALSE, ellipses = FALSE)
```

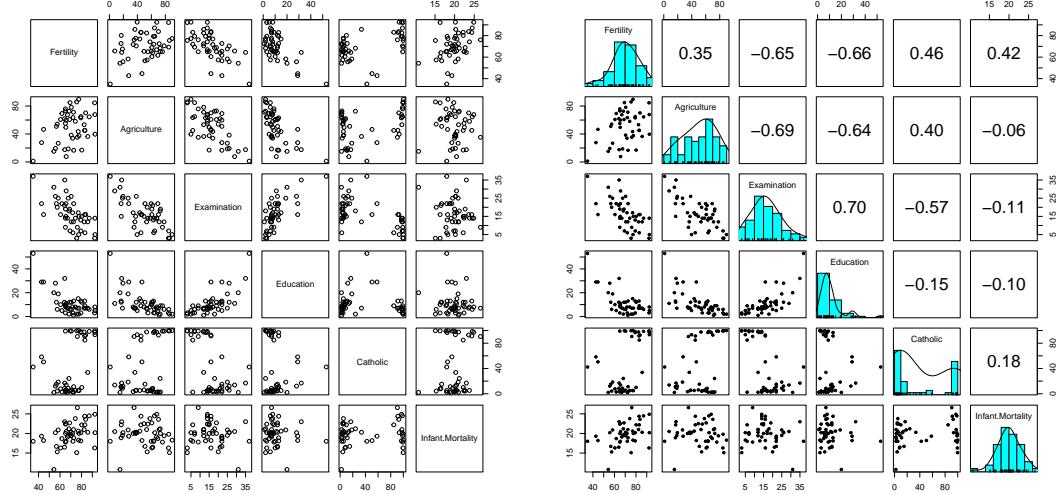


Abbildung 6.5: Bivariate Verteilungen

Zusammenhangs ist, müssen wir den Daten ein (lineares) *Modell* zugrunde legen, das uns erlaubt, Y durch X zu erklären.

Man sagt dann, man macht eine *Regression* von der *Zielgröße* Y auf die *Eingangsgröße* X ¹. Die Zielgröße Y nennen wir manchmal auch *Kriterium*, *Outcome* oder *Response*, die Eingangsgröße X manchmal auch *Prädiktor* oder *Kovariable*.

Mathematische Grundlage: Lineare Funktionen. Aus der Schule kennen wir noch die einfache *lineare Funktion*, die durch

$$y = a + b \cdot x \quad (6.2.1)$$

gegebene *Gerade*. Dabei wurde der y -Wert auf der Vertikalen, der x -Wert auf der Horizontalen aufgezeichnet, a war der “ y -Achsenabschnitt” der Geraden und b war die *Steigung* der Geraden. Diese **lineare Funktion – und Verallgemeinerungen davon – spielen eine ganz zentrale Rolle in der Wissenschaft**. Abbildung 6.6 zeigt eine lineare Funktion $y(x) = 4 + 5x$ und eine nichtlineare Funktion $y(x) = \exp(x)$. Insbesondere ist bei der linearen Funktion die Steigung $b = 5$ eine zentrale Größe, die wir später in linearen Modellen oft antreffen werden. Damit wir diese Größe auch allgemein gut verstehen, brauchen wir den Begriff der *Ableitung*.

¹regredi=zurückgehen

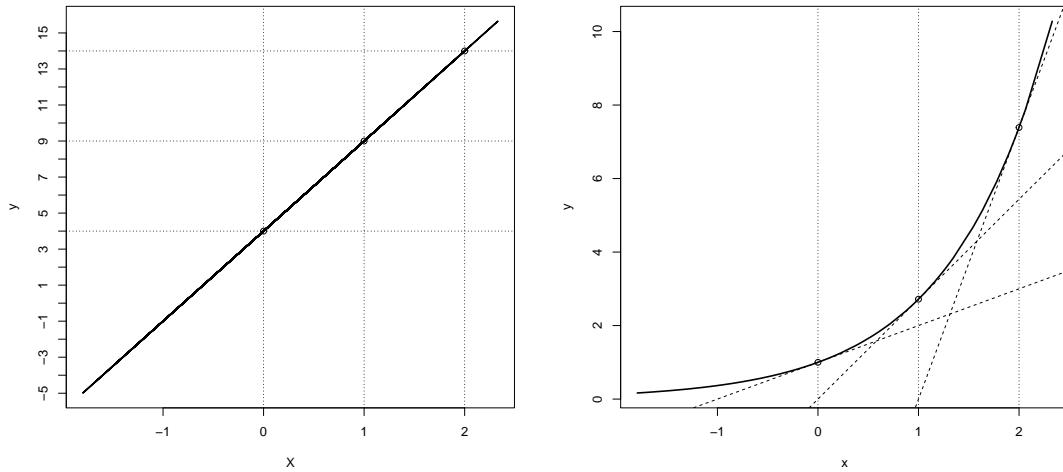


Abbildung 6.6: Links: Lineare Funktion $y = 4 + 5x$, $a = 4$ und $b = 5$ können abgelesen werden. Rechts: Exponentialfunktion $y = \exp(x)$ mit nicht konstanter Steigung.

Definition. Die *Ableitung* $f'(x_0)$ einer Funktion f an einem Punkt x_0 ist

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}. \quad (6.2.2)$$

Anschaulich: Abbildung 6.7 zeigt die Ableitung einer Funktion bei $x_0 = 0$ als Grenzwert von Steigungen von Sekanten, wenn $x \rightarrow x_0$ (Steigung der roten Tangente).

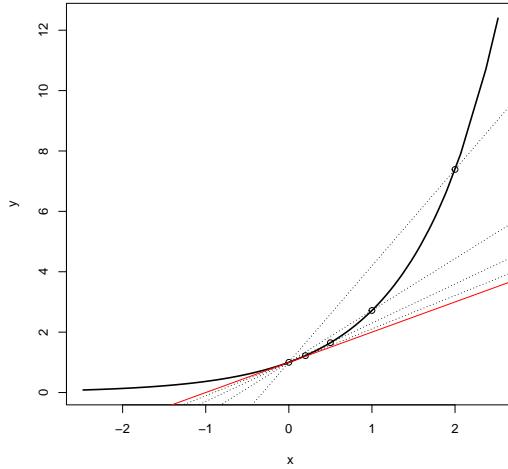


Abbildung 6.7: Ableitung einer Funktion an der Stelle $x_0 = 0$

Bei einer linearen Funktion $f(x) = a + bx$ ist die Ableitung oder Steigung $f'(x) = b$, also über alle x konstant². Diese Grösse ist dann zu interpretieren als die **Veränderung auf y pro Einheit Veränderung auf x** ,

$$f'(x) = b = \frac{\Delta y}{\Delta x}. \quad (6.2.3)$$

Diese Grösse ist in der Abbildung 6.6 auf der Y -Achse ablesbar. Bei einer Einheit Veränderung auf x verändert sich y um 5, bei zwei Einheiten um 10, usw. Ebenso ist a als der Wert der Funktion bei $x = 0$ erkennbar. Bei einer linearen Funktion ist die Steigung, der Zuwachs pro Einheit, überall gleich gross, nämlich b . Im Allgemeinen aber, wie z.B. bei der Exponentialfunktion, ist das nicht der Fall³. Lineare Funktionen nehmen also einen ganz speziellen Platz ein in der Menge aller Funktionen.

Lineare Regression. Zurück zur Statistik. Das grundlegende Regressionsmodell ist das *einfache lineare Modell*, ein Modell mit *einer* Eingangsgrösse X . In diesem Modell ist die Zielgrösse Y die Summe aus einer linearen Funktion von x und einem *zufälligen* Messfehler ϵ . Die Zielgrösse Y_i wird modelliert mit

$$Y_i = \underbrace{\alpha + \beta \cdot x_i}_{\text{linearer Prediktor}} + \underbrace{\epsilon_i}_{\text{Messfehler}}, \quad i = 1, \dots, n, \quad (6.2.4)$$

² $f'(x_0) = \lim_{x \rightarrow x_0} \frac{a+bx-(a+bx_0)}{x-x_0} = \lim_{x \rightarrow x_0} \frac{b(x-x_0)}{x-x_0} = \lim_{x \rightarrow x_0} b = b$.

³ Jede Exponentialfunktion steigt irgendwann **viel schneller** als jede lineare Funktion mit noch so grosser Steigung. Die Ableitung der Exponentialfunktion ist $f'(x) = \exp(x)$. Der Zuwachs wächst also selber exponentiell.

Im Gegensatz zu einer linearen *Funktion* hat es also in einem linearen *Modell* zusätzlich eine stochastische, eine Zufallsgrösse. Oft wird angenommen, dass die Abweichungen, die stochastischen Fehler ϵ_i , $i = 1, \dots, n$, eine bestimmte Verteilung haben, z.B. eine *Normalverteilung*, und dass sie *stochastisch unabhängig*, also nicht korreliert sind. Sie bilden also eine i.i.d. Zufalls-Stichprobe⁴. Da der Erwartungswert des Fehlers ϵ_i Null ist, ist

$$E(Y_i) = \alpha + \beta x_i. \quad (6.2.5)$$

Diese Grösse nennt man den *linearen Prädiktor*. Somit ist die Steigung, die erwartete Veränderung auf Y bei einer Einheit Zuwachs auf X

$$E(Y_i | x_i + 1) - E(Y_i | x_i) = \alpha + \beta(x_i + 1) - (\alpha + \beta(x_i)) = \beta. \quad (6.2.6)$$

Die beiden **unbekannten** Parameter α und β heissen *Regressionskoeffizienten* und sind aus den Daten zu *schätzen*. Die geschätzten Grössen bezeichnen wir dann mit $\hat{\alpha}$ und $\hat{\beta}$. Wie kann man die Parameter oder Regressionskoeffizienten α und β quantifizieren?⁵ Dazu müssen wir eine Gerade so durch die Punktewolke legen, dass diese in einem gewissen Sinn *optimal* ist, dass diese Gerade “im Schnitt am nächsten zu allen Punkten ist” (Abbildung 6.8). Dies bewerkstelligt man mit der sogenannten *Methode der kleinsten Quadrate* (Least Squares).

Kleinste-Quadrat Schätzer. Diese Methode wählt diejenigen Parameter α und β , die die *Summe der quadrierten Residuen* (“Reste”) des Modells minimiert: Das i -te Residuum r_i ist ja

$$r_i = y_i - \hat{y}_i. \quad (6.2.7)$$

Die \hat{y}_i sind die *angepassten* Werte (markiert durch Kreuze). Die Residuen sind durch die gestrichelten Linien markiert. Man bestimmt dann diejenigen $\hat{\alpha}$ und $\hat{\beta}$, für die die Quadratsumme der Residuen

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6.2.8)$$

minimal wird (durch Ableiten und Nullsetzen). Die *Kleinste-Quadrat Schätzer* sind dann (ohne Herleitung):

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Diese Berechnungen wird natürlich R für uns machen. In R werden lineare Modelle mit `lm()` (Linear Model) angepasst. `lm()` schätzt aus den Daten $\hat{\alpha}$ und $\hat{\beta}$ mit der Methode der kleinsten Quadrate.

Folgender Code macht eine Regression von Reaktionszeit (abhängige Variable, vor dem

⁴Wir notieren dann $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, die ϵ_i sind i.i.d. (für *independent and identically distributed*).

⁵Auf die allgemeinen Grundlagen des statistischen Schätzens kommen wir in Kapitel 7 zurück.

~ Zeichen) auf Alkoholkonzentration (unabhängige Variable, nach dem ~ Zeichen). Angepasste Modelle werden in R auch als Objekte abgespeichert, hier im Objekt `myfirstmodel`.

```
myfirstmodel <- lm(R ~ A)
myfirstmodel

##
## Call:
## lm(formula = R ~ A)
##
## Coefficients:
## (Intercept)          A
##      551.7        90.3
```

Im Output steht zuerst nur das wichtigste, nämlich die beiden geschätzten Parameter. `Intercept` steht für $\hat{\alpha}$, unter `A` steht der Parameter für Alkoholkonzentration, also $\hat{\beta}$. In einem Modellobjekt wie `myfirstmodel` ist aber noch viel mehr Information enthalten, die wir dann später brauchen. Im Moment brauchen wir aber diese Information (noch) nicht.

```
str(myfirstmodel)
```

Schauen wir jetzt das Modell graphisch an, indem wir die Regressionsgerade hinzufügen. Das machen wir mit `abline()`, indem man das Modellobjekt als Argument übergibt (`a,b` steht gerade für die Parameter des Modells). Mit `abline()` kann man aber auch vertikale und horizontale Linien in einer Graphik hinzufügen (`?abline`). Mit `fitted()` kann man die angepassten Werte des Modells ausgeben lassen:

```
plot(A, R, xlab = "Alkoholkonzentration [Promille]", ylab = "Reaktionszeit [ms]")
abline(myfirstmodel) ## fügt die Regressionsgerade hinzu, siehe ?abline
points(A, fitted(myfirstmodel), pch = 3) ## plottet angepasste Werte als Kreuze
segments(x0 = A, y0 = fitted(myfirstmodel), x1 = A, y1 = R, lty = 2) ## Zur Hilfe: plottet Residuen.
```

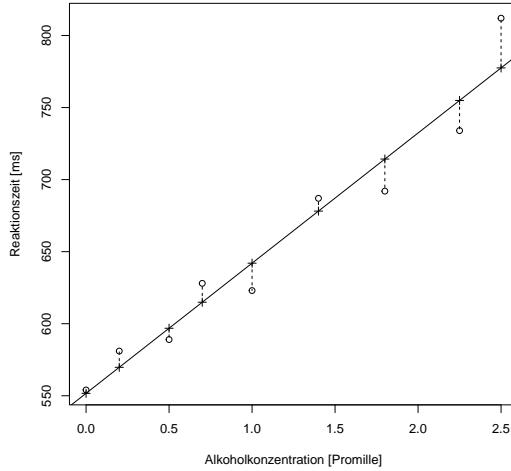


Abbildung 6.8: Regression von Reaktionszeit auf Alkoholkonzentration. Kreuze: Angepasste Werte, Punkte: Beobachtete Werte, Gestrichelt: Residuen.

In unserem Beispiel ist also $\hat{\beta} = 90.329$ die Zunahme an Reaktionszeit in Millisekunden pro Promille Zunahme in der Alkoholkonzentration. Mit diesen Grössen kann man für *neue Werte* der Eingangsgrösse die Zielgrösse *vorhersagen*, der vorhergesagte Wert für ein x_{neu} (Alkoholkonzentration) wäre dann

$$\hat{Y}_{neu} = \hat{\alpha} + \hat{\beta}x_{neu} = 551.678 + 90.329x_{neu}.$$

Insbesondere war hier der Parameter β von Interesse. Er ist wie das b aus der Schulzeit als Steigung zu interpretieren. β kann theoretisch Werte zwischen $-\infty$ und $+\infty$ einnehmen. Wir werden sehen, dass viele Quantitäten von Interesse (*Quantities of interest*) in der Wissenschaft solche Steigungen (slopes) sind.

Wichtig. Wir sind immer noch im Stadium der deskriptiven Statistik. Wir haben die Steigung $\hat{\beta}$ (den Effekt von Alkoholkonzentration auf Reaktionszeit) für *diese* Daten quantifiziert. Später werden wir uns auch mit der Quantifizierung der *Unsicherheit* dieser Schätzung befassen. Dann wird die Frage lauten: In welchem Bereich oder Intervall liegt – mit grosser Wahrscheinlichkeit – der wahre Effekt (z.B. die wahre Steigung β), derjenige, den man hätte, wenn man die ganze Population gemessen hätte?

Multiple Korrelation. Der *angepasste* Wert \hat{y}_i ist i.A. nicht identisch mit dem beobachteten Wert y_i , ausser unser Modell hat gleich viele Parameter wie Daten. Es ist immer möglich, eine komplexe Kurve genau durch die 9 Datenpunkte zu zeichnen, dieses *komplexere* Modell hätte dann aber gleich viele Parameter wie Daten. Abbildung

6.9 zeigt das lineare Modell und zwei komplexere Modelle mit 4 respektive 9 Parametern (Polynome 3. und 8. Grades). Letzteres hat gleiche viele Parameter wie Daten und passt dann perfekt zu *diesen* Daten, ist aber für die *Vorhersage* offensichtlich schlecht. Solche Modelle haben eine *Überanpassung (overfit)*. Darauf kommen wir später zurück. Einfache Modelle sind oft besser, auch wenn sie nicht perfekt zu den verfügbaren Daten “passen”.

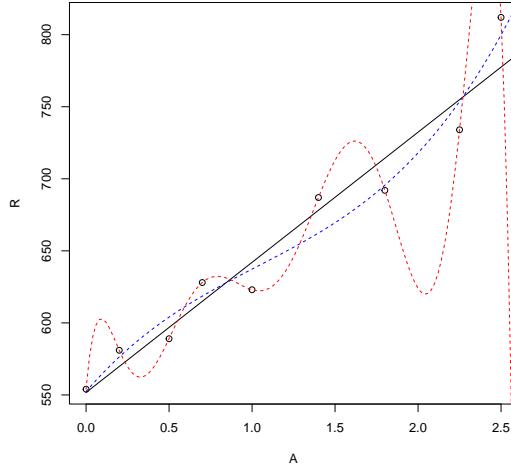


Abbildung 6.9: Einfaches lineares Modell (schwarz, 2 Parameter), Polynom 3. Grades (blau, 4 Parameter) und Polynom 8. Grades (rot, $n = 9$ Parameter)

Zurück zum linearen Modell. Die Abweichung $y_i - \hat{y}_i$ nannten wir oben das i -te Residuum. Wir haben n Daten, n Residuen und n angepasste Werte, wir nennen dann

- $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ die **Residuen-Quadratsumme**. (Das wäre in Abbildung 6.10 die Summe aller quadrierten Längen der roten Strichlinien).
- $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ als der durch das Modell **erklärte Quadratsumme**. (analog, punktierte, blaue Linien).
- $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ die **totale Quadratsumme** (analog, grüne Strichpunktlinien)

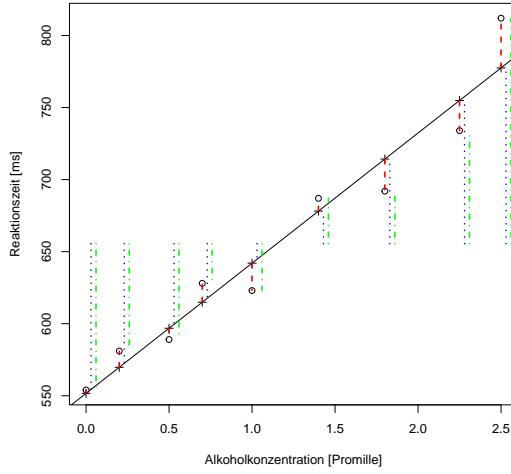


Abbildung 6.10: Residuen: $y_i - \hat{y}_i$ (rot gestrichelt). Durch Modell erklärt: $\hat{y}_i - \bar{y}$ (blau gepunktet). Total: $y_i - \bar{y}$ (grün strichpunkt)

Man kann zeigen dass man die totale Quadratsumme zerlegen kann gemäss

$$TSS = ESS + RSS. \quad (6.2.9)$$

Die Korrelation zwischen den Beobachtungen y_i und den angepassten Werten \hat{y}_i nennt man *multiple Korrelation*. Die quadrierte multiple Korrelation R^2 wird auch *Bestimmtheitsmaß* genannt, da sie den Anteil der Streuung der y -Werte bestimmt, der durch die Regression *bestimmt* wird. R^2 wird berechnet als Verhältnis von erklärter zu totaler Streuung,

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS}. \quad (6.2.10)$$

Die angepassten Werte kann man mit `fitted()` extrahieren (die Kreuze in Abbildung 6.10)

```
fitted(myfirstmodel)

##   1   2   3   4   5   6   7   8   9
## 552 570 597 615 642 678 714 755 777
```

Die durch das Modell erklärte Quadratsumme ist dann

```
sum((fitted(myfirstmodel) - mean(R))^2)

## [1] 52138
```

und die Residuen-Quadratsumme ist

```
sum((R - fitted(myfirstmodel))^2)

## [1] 2929
```

und die totale Quadratsumme ist

```
sum((R - mean(R))^2)

## [1] 55066
```

Also ist $R^2 = \frac{52137.69}{52137.69+2928.532} = 0.947$.

Unser lineares Modell (mit zwei Parametern α und β) erklärt also 94.6% der Streuung der beobachteten Werte. Man kann zeigen, dass bei der einfachen Regression (nur eine erklärende Variable) R^2 identisch ist mit dem Korrelationskoeffizienten im Quadrat.

```
cor(A, R)^2
```

Einfache Regression und Kausalität. Wie bei der Korrelation gilt auch für die einfache Regression: Die Tatsache, dass es einen linearen Zusammenhang gibt, bedeutet i.A. nicht, dass die Eingangsgröße auch eine *Ursache* ist für das *Kriterium*. Es könnte sein, dass die wahre Ursache gar keine Eingangsgröße im Modell war.

Das führt uns nun zur *multiplen Regression*, mit der wir uns auch im nächsten Modul viel beschäftigen werden. In einem multiplen Modell werden zusätzliche potentielle Ursachen als Eingangsgrößen ins Modell aufgenommen. Mit dem einfachen linearen Modell haben wir dafür die Grundlage geschaffen.

6.3 Multiple Regression*

Mit linearen Modellen, die nicht nur einen, sondern *mehrere Eingangsgrößen* oder *Kovariablen* in das Modell einbeziehen, versucht man den Effekt jedes einzelnen Prädiktors auf die Zielgröße zu quantifizieren. Wir möchten dann den Effekt jeder einzelnen Eingangsgröße für den Effekt von anderen potentiellen Kovariablen oder Störgrößen *korrigieren*. In Kontext der Abbildung 6.3 möchten wir den Effekt von X auf Y schätzen, wenn wir den Effekt von C berücksichtigen. C wäre dann eine zweite Eingangsgröße, deren Effekt uns vielleicht gar nicht interessiert, aber die Integration von C als Eingangsgröße oder potentieller Prädiktor ins Modell bewirkt, dass unsere Schätzung des Effekts von X auf Y nicht durch C verzerrt wird.

Wir haben es dann mit einer *multiplen Regression* zu tun. Dieser Begriff taucht in Studien sehr oft als Analysemethode auf, z.B., wenn Störgrößen nicht über das Design kontrolliert werden können. In einem solchen Modell wird aus den Daten ein Effekt-Parameter (eine *Steigung* oder *Regressionskoeffizient*) für jeden einzelnen Prädiktor geschätzt. Diese sind dann wieder zu interpretieren als Veränderung des Kriteriums pro Einheit Veränderung auf jeder entsprechenden Eingangsgröße, oder als die Differenz in mittlerem Y für zwei Subpopulationen, die sich auf der entsprechenden Eingangsgröße um eine Einheit unterscheiden, **wenn alle anderen Eingangsgrößen konstant bleiben (ceteris paribus)**.

Wie sieht ganz allgemein ein solches (multiples) lineares Modell aus? Der lineare Prädiktor in 11.2.1 wird erweitert mit $\beta_j x_{ij}$ -Summanden für zusätzliche Eingangsgrößen x_{ij} .

Bei p Eingangsgrößen schreiben wir das Modell für die Zielgröße Y

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n. \quad (6.3.1)$$

Die erste Eingangsgröße ist meistens eine Konstante, $x_{i1} = 1$, wir haben dann wie bei der einfachen Regression ein *Intercept* im Modell, dass jetzt β_1 heißt statt α .

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n. \quad (6.3.2)$$

Wir haben also wieder einen linearen Prädiktor (jetzt mit $p - 1$ Eingangsgrößen x_{i2}, \dots, x_{ip}) und einen zusätzlichem stochastischen Fehlerterm. Die Fehler $\epsilon_1, \dots, \epsilon_n$ bilden dann wieder eine i.i.d. Zufallstichprobe.

Sind dann $\hat{\beta}_j$ die geschätzten Parameter – wieder über die Minimierung der Summe der quadrierten Residuen des Modells, können wir mit dem *angepassten* Modell die Zielgröße für eine *neue Daten* $x_{neu,2}, \dots, x_{neu,p}$ abschätzen mit

$$\hat{Y}_{neu} = \hat{\beta}_1 + \hat{\beta}_2 x_{neu,2} + \cdots + \hat{\beta}_j x_{neu,j} + \cdots + \hat{\beta}_p x_{neu,p}. \quad (6.3.3)$$

Beispiel: Swiss fertility data. Multiple Regression mit R. Der Datensatz `swiss` ist aus vorigem Kapitel bereits bekannt. Fertilität ist aber jetzt die abhängige Variable, die Zielgröße. Eine einfache Regression von Fertilität auf Catholic wäre

```
model1 <- lm(Fertility ~ Catholic, data = swiss)
model1

##
## Call:
## lm(formula = Fertility ~ Catholic, data = swiss)
##
## Coefficients:
## (Intercept)      Catholic
##       64.428        0.139
```

Wollen wir den Effekt von Catholic kontrollieren für den Effekt von Examination, machen wir eine multiple Regression von Fertilität auf Catholic *und* Examination.

```
model2 <- lm(Fertility ~ Catholic + Examination, data = swiss)
model2

##
## Call:
## lm(formula = Fertility ~ Catholic + Examination, data = swiss)
##
## Coefficients:
## (Intercept)      Catholic  Examination
##       83.0357        0.0418     -0.8862
```

Hier ist $p = 3$ und R gibt die Schätzungen für β_1 (Intercept), β_2 und β_3 im Output. Der für den Effekt von Examination kontrollierte Effekt von Catholic (0.042) ist kleiner als der nicht kontrollierte Effekt (0.139). Für eine multiple Regression auf *alle* Eingangsgrößen im data frame wird, das abgekürzt mit einem Punkt notiert rechts vom \sim . Das gibt ein Modell mit $p = 6$ Parametern.

```
model3 <- lm(Fertility ~ ., data = swiss)
model3

##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Coefficients:
## (Intercept)      Agriculture    Examination     Education      Catholic
##       66.915          -0.172         -0.258        -0.871        0.104
## Infant.Mortality
##           1.077
```

Interpretation. Eine Zunahme auf Agriculture, Examination oder Education ist assoziiert mit einer Verminderung der Fertilität, eine Zunahme von Catholic oder

Infant Mortality mit einer Erhöhung der Fertilität. Wenn wir z.B. den Koeffizienten von Catholic betrachten, dann ist die Veränderung vom Fertilitäts-Score pro Einheit Veränderung auf Catholic gerade 0.1041. Also können wir sagen: Die Veränderung auf Fertilität ist *pro Prozent* mehr Katholiken gerade 0.1041. Alle Effekte sind immer kontrolliert für die Effekte der jeweils anderen im Modell enthaltenen Prädiktoren.

Bei einem Modellobjekt kann sich mit `model$coef` die Parameterschätzungen direkt herausgeben lassen. Wir machen damit eine Zusammenstellung der Effekte von Catholic in den drei angepassten Modellen.

```
data.frame(unadjusted = model1$coef[2], adjustForExam = model2$coef[2], adjustForAllOther = model3$coef[5])  
  
##           unadjusted adjustForExam adjustForAllOther  
## Catholic      0.139       0.0418        0.104
```

6.4 Ziele von Regressionsanalysen

In Studien treffen wir sehr oft auf multiple Regressionsmodelle. Immer dann, wenn man Effekte für Störgrößen oder andere Kovariablen korrigieren will, kommen Regressionsmodelle zum Zug. Man will den Effekt einer Eingangsgröße quantifizieren, und zwar so, dass der Effekt von anderen Eingangsgrößen bereits “wegaddiert” wurde. Multiple Regressionen sagen also *viel mehr* aus als einfache Regressionen.

Allgemeine Ziele von Regressionsanalysen. Die Ziele von Regressionsanalysen sind: Man will eine gute *Anpassung* des Modells an die Daten, d.h. die “Reste” des Modells sollen klein sein. Zudem will man “gute” Schätzungen der *Parameter* des Modells. Ein weiteres Ziel ist die *Vorhersage* der abhängigen Variablen bei neuen Daten als Eingangsgrößen.

Später: Unsicherheit der Schätzung und Vorhersage. Wie oben bereits betont wurde, sind wir immer noch im Stadium der deskriptiven Statistik. Wir haben die Effekte auf Fertilität für *nur diese* Daten quantifiziert. Das Problem der Quantifizierung von Unsicherheiten und Signifikanz bezüglich der Parameterschätzung und Vorhersage werden wir im Kapitel 7 behandeln.

Kapitel 7

Von der Stichprobe zur Population

Schätzverfahren zielen darauf ab, von einer *Zufallsstichprobe* auf die *Grundgesamtheit* oder Population zurückzuschliessen. Wir wollen also über die Stichprobe hinausgehen. Wir wollen versuchen, das *Induktionsproblem* anzugehen, das Problem der *Verallgemeinerung* von dem, was wir in Stichproben empirisch ermittelt haben.

Beim Schätzen unterscheiden wir zwischen der Punktschätzung und der Intervallschätzung. Bei der Punktschätzung fragen wir nach dem *plausibelsten* Wert für eine unbekannte Grösse, bei der Intervallschätzung fragen wir nach einem *Bereich* von plausiblen Werten für eine unbekannte Grösse.

7.1 Punktschätzung

Mit der *Punktschätzung* wollen wir möglichst genaue *Näherungen* für unbekannte Grössen in einer Population oder in der Zukunft angeben.

Eine solche unbekannte Grösse kann z.B. der Parameter θ eines statistischen (parametrischen) Modells sein. Beispiele für solche Parameter wären

- der Erwartungswert μ einer Variablen,
- die Eintretenswahrscheinlichkeit π einer binomialverteilten Variablen,
- der Parameter λ einer Poisson-verteilten Variablen,
- ein Unterschied zwischen zwei Erwartungswerten $\Delta = \mu_1 - \mu_2$,

und viele andere, je nachdem, welche Grösse von Interesse ist.

Wir wollen also etwas über die unbekannten Grössen aussagen, die den **Mechanismus der Datengenerierung**, also die Verteilung, spezifizieren. So spezifizieren die Parameter μ und σ^2 normalverteilte Daten, λ Poisson-verteilte Daten usw. Die Punktschätzung sucht dann nach Schätzungen $\hat{\mu}$ und $\hat{\sigma}^2$ und $\hat{\lambda}$, respektive.

Die unbekannten Grössen müssen aber nicht Parameter eines (parametrischen) Modells sein, sondern können allgemeinere Grössen sein. In nicht-parametrischen Modellen stehen nicht Parameter im Vordergrund, sondern generelle Aspekte der Verteilung wie Median oder ganz allgemein Quantile.

Notation. Unbekannte Größen notieren wir in der Regel mit *griechischen* Symbolen. Damit grenzen wir diese von den Größen ab, die wir direkt beobachten oder aus Daten berechnen können, wie z.B. den *empirischen* Mittelwert \bar{x} usw. Im Gegensatz zu einer empirischen Größe sind unbekannte $\theta, \pi, \mu, \lambda$ usw. theoretische, *abstrakte* Größen.

Damit wir eine unbekannte Größe schätzen können, brauchen wir empirische Daten. Für die Punktschätzung beginnt man mit n Stichprobenziehungs, die durch die Zufallsvariablen oder Stichprobenvariablen X_1, \dots, X_n repräsentiert werden. Die Stichprobenvariable X_i ist ein Modell, eine Abstraktion für die i -te Beobachtung, wie wir das in (5.3.1) dargestellt haben. Im Moment nehmen wir an, dass die Zufallsvariablen X_i i.i.d. sind. Später werden es auch mit korrelierten Daten zu tun haben, z.B. bei Messwiederholungen, dort sind Beobachtungen innerhalb einer Person nicht mehr voneinander unabhängig.

7.1.1 Schätzstatistik und Schätzwert

Allgemein ist eine *Schätzfunktion*, ein *Schätzer* oder eine *Statistik* T für einen unbekannten Parameter θ eine Funktion g der Stichprobenvariablen X_1, \dots, X_n , kurz

$$T = g(X_1, X_2, \dots, X_n). \quad (7.1.1)$$

Alles, was man aus Daten berechnen kann, nennt man eine Statistik, also jede Funktion $g()$ in 7.1.1. Wie wir sehen werden, gibt es natürlich “vernünftige” und weniger vernünftige Statistiken. Der aus den konkreten Beobachtungen x_1, \dots, x_n geschätzte Wert t ist dann der *Schätzwert*, also

$$t = g(x_1, x_2, \dots, x_n). \quad (7.1.2)$$

Zur Erinnerung: Große Buchstaben wie T sind Zufallsvariable und kleine Buchstaben wie t stellen Werte dar, die diese Variable einnimmt.

Statistiken liefern Information über die Verteilung von X und sind die Bausteine beim *Schätzen* von Parametern und beim *Testen von Hypothesen* über diese Parameter.

Wir schauen uns jetzt eine grundlegende Statistik an, den arithmetischen Durchschnitt $T = \bar{X}$.

7.1.2 Schätzen von μ

Wir betrachten im Folgenden eine grundlegende Schätzung. Wir wollen μ – den Erwartungswert einer Variablen – aus den Daten einer Zufallsstichprobe schätzen.

Zur Erinnerung: Der Erwartungswert ist ein mit den Wahrscheinlichkeiten gewichtetes Mittel der Werte einer Variablen, bei einer diskreten k -wertigen Variablen $E(X) = \sum_{i=1}^k x_i p(x_i)$. Der empirische Mittelwert ist analog (wir ersetzen die

theoretische Wahrscheinlichkeit p durch die empirische relative Häufigkeit f)
 $\bar{x} = \sum_{i=1}^k x_i f(x_i) = \frac{1}{n} \sum_{i=1}^k x_i h(x_i) = \frac{1}{n} \sum_{i=1}^n x_i$.

Für die Schätzung von μ ist dann sicher das arithmetische Mittel, jetzt als Variable betrachtet,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (7.1.3)$$

eine vernünftige *Schätzstatistik*, also ist $T = \bar{X}$. Gleichzeitig ist

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.1.4)$$

der zugehörige *Schätzwert* $t = \bar{x}$. Dieser Schätzwert ist i.A. nicht mit μ identisch. Wollen wir z.B. die wahre mittlere Körpergrösse μ einer Population schätzen, so nehmen wir dazu einfach den beobachteten Durchschnitt aus den n Beobachtungen als Schätzwert für μ .

Gedankenexperiment. Der Durchschnitt \bar{X} ist jetzt also eine Zufallsvariable, im Gegensatz zu \bar{x} , der eine Zahl ist! Wir können die Variable \bar{X} auch folgendermassen an einem *Gedankenexperiment* veranschaulichen:

1. Ziehe eine Stichprobe der Grösse n aus $X: X_1, \dots, X_n$.
2. Betrachte den mit dem Schätzer berechneten Schätzwert in der Stichprobe (d.h. „berechne“ den empirischen Mittelwert \bar{x}).
3. Wiederhole 1 – 2 sehr viele Male.
4. Die Verteilung der so ermittelten empirischen Mittelwerte ist die Verteilung von \bar{X} , die sogenannte *Stichprobenverteilung* der Mittelwerte.

Erwartungstreue und Konsistenz eines Schätzers. Diese abstrakte Variable \bar{X} hat dann als Mittelwert (oder Erwartungswert) $E(\bar{X}) = \mu$. Das bedeutet, dass man *im Schnitt* mit \bar{X} richtig liegt, und zwar unabhängig von n . Man sagt dann, dass \bar{X} ein *unverzerrter* Schätzer ist; man nennt diese Eigenschaft einer Statistik dann auch *Erwartungstreue*.

Das sogenannte *Gesetz der grossen Zahlen* besagt zudem, dass \bar{X} mit wachsendem n gegen μ strebt. Man nennt diese Eigenschaft einer Statistik *Konsistenz*.

7.1.3 Standardfehler

Zur Quantifizierung der Variabilität eines Schätzers T – und damit wir später sogenannte *Konfidenzintervalle* konstruieren können – brauchen wir die Varianz $\text{Var}(T)$ respektive die Standardabweichung $\sqrt{\text{Var}(T)}$ von T .

Wir führen das Konzept wieder ein für die Schätzung von einem Erwartungswert μ , also über die Statistik $T = \bar{X}$. Wir haben bereits gezeigt (siehe 4.8.1), dass die Varianz von \bar{X} bei unabhängigen und identisch verteilten X_i

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\sigma^2}{n} \quad (7.1.5)$$

ist, mit σ^2 als der (wahren) Varianz von X .

Die geschätzte Standardabweichung der Schätzers \bar{X} , also $\sqrt{\text{Var}(\bar{X})}$ heisst *Standardfehler* (standard error se) von \bar{X} . Wenn wir σ^2 mit s^2 (5.5.5) schätzen, dann ist

$$\text{se}(\bar{X}) = \frac{s}{\sqrt{n}} \quad (7.1.6)$$

ein Standardfehler für \bar{X} . In Studien sieht man sehr häufig neben dem Wert der Punktschätzung eines Parameters auch den zugehörigen Standardfehler der entsprechenden Schätzung. Es geht dann natürlich nicht immer um die Schätzung von einem wahren Mittelwert μ , aber vielleicht um die Schätzung einer wahren Proportion π (wie bei Abstimmungen oder Wahlen) oder einen Unterschied Δ (wie bei klinischen Experimenten). Letzteren häufigen Fall werden wir in Kapitel 7.5.1 vertiefen.

Standardfehler und n . Schauen wir uns 7.1.6 näher an. Diese Gleichung ist fundamental in der Statistik. Mit wachsendem n wird der Standardfehler eines Schätzers immer kleiner. Dies ist natürlich auch intuitiv nachvollziehbar und eine Alltagserfahrung. Mehr Information bedeutet – bei unabhängigen Beobachtungen – auch mehr Präzision. Durchschnitte sind also präziser als Einzelmessungen. Große Studien, Erhebungen usw. können präziser schätzen als kleine Studien. Bei $n = 1$ ist der Standardfehler äquivalent mit der Standardabweichung von X . Wenn n 4-mal grösser wird, dann wird der Standardfehler 2-mal kleiner, die Schätzung also 2-mal präziser. Die Stichprobengrösse wächst also im Quadrat zur gewünschten Präzision.

7.1.4 Zusammenfassung

Die geschätzte Standardabweichung von \bar{X} heisst *Standardfehler* des Mittelwerts. Der Standardfehler stellt eine Grösse dar für die *Präzision* bei der Schätzung des unbekannten Parameters μ . Die Statistik \bar{X} ist eine *erwartungstreue* und *konsistente* Schätzfunktion für μ , \bar{x} die zugehörige Realisation. Zusammenfassend haben wir für die Punktschätzung $\hat{\mu}$ des Parameters μ und für den Standardfehler von $\hat{\mu}$

$$\hat{\mu} = \bar{X}, \quad \text{se}(\hat{\mu}) = \frac{s}{\sqrt{n}}. \quad (7.1.7)$$

Der jetzt folgende zentrale Grenzwertsatz sagt zusätzlich, dass \bar{X} approximativ normalverteilt ist, wenn n gross.

7.2 Verteilungsaussagen

Später wird es nützlich oder nötig sein, die Verteilung des Schätzers zu kennen. Exakte Aussagen gibt es nur wenige. Für die Normalverteilung folgt das weiter unten. Einen allgemeinem Zugang liefert der Zentrale Grenzwertsatz.

7.2.1 Zentraler Grenzwertsatz

Bei unserem grundlegenden Schätzproblem kennen wir nun also den Mittelwert und die Varianz von unserem Schätzer $T = \bar{X}$. Gibt es auch Anhaltspunkte, *wie* die Statistik \bar{X} verteilt ist? Das interessanteste Resultat aus der Wahrscheinlichkeitsrechnung ist der sogenannte *zentrale Grenzwertsatz*:

X sei eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2 , und X_1, X_2, \dots, X_n seien i.i.d. Kopien davon.

Dann ist \bar{X} – für grosses n – approximativ normalverteilt mit Mittelwert μ und Varianz σ^2/n , und zwar unabhängig von der Verteilung des Merkmals in der Population.

In der Praxis ist “gross” oft $n = 30$. Die exakte Normalverteilung gilt erst für $n \rightarrow \infty$. Für grosses n gilt eine *asymptotische* oder *approximative* Normalverteilung. Man notiert dann

$$\bar{X} \xrightarrow{a} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad (7.2.1)$$

oder, äquivalent

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{a} \mathcal{N}(0, 1). \quad (7.2.2)$$

Mit der Simulation <https://rstudio.zhaw.ch/rsconnect/content/91> kann man den zentralen Grenzwertsatz empirisch mit verschiedenen Verteilungen nachvollziehen.

Der Grenzwertsatz gilt auch für die Summe von i.i.d. Zufallsvariablen

$$S = X_1 + X_2 + \dots + X_n.$$

Da $S = n\bar{X}$ und $\text{Var}(S) = n^2 \text{Var}(\bar{X}) = n^2\sigma^2/n = n\sigma^2$ (siehe 4.8.1), ist folgende Aussage äquivalent zu 7.2.1,

$$S \xrightarrow{a} \mathcal{N}(n\mu, n\sigma^2). \quad (7.2.3)$$

Beispiel Normal-Approximation der Binomialverteilung. Die *Normal-Approximation* ist eine Methode der Wahrscheinlichkeitsrechnung, um die Binomialverteilung für grosse Stichproben durch die Normalverteilung anzunähern. Wir hatten das bereits bei der Einführung der Normalverteilung gesehen. Die Summe von n i.i.d. $\text{Bin}(\pi, 1)$ Bernoulli-verteilten Zufallsvariablen, also eine binomialverteilte Zufallsvariable, ist bei n gross approximativ normalverteilt mit Erwartungswert $\mu = n\pi$ und Varianz $\sigma^2 = n\pi(1 - \pi)$, also

$$S \xrightarrow{a} \mathcal{N}(n\pi, n\pi(1 - \pi)).$$

Für die Interessierten*. Folgender Code zeigt das empirisch auf

```
nsim <- 1000
n <- 40 #n 'gross'
pi <- 0.3 #Eintretenswahrscheinlichkeit
X <- matrix(NA, ncol = nsim, nrow = n) #initialisieren Matrix mit nsim Kolonnen und n Zeilen
S <- rep(NA, nsim) #initialisieren Vektor mit nsim Elementen
for (i in 1:nsim) {
  X[, i] <- rbinom(n, size = 1, prob = pi) #Sample mit n iid Bernoulli
  S[i] <- sum(X[, i]) #Summe
}
hist(S)
mean(S) #empirischer Durchschnitt
n * pi #Erwartungswert
var(S) #empirische Varianz
n * pi * (1 - pi) #wahre Varianz
```

7.2.2 Exakte Aussagen bei Normalverteilung

Für normalverteilte Grössen hat man *exakte* Aussagen:

1. \bar{X} ist **normalverteilt**: $\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
2. $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ ist **standardnormalverteilt**: $\sim \mathcal{N}(0, 1)$.
3. $\frac{n-1}{\sigma^2}s^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$ ist **χ^2 -verteilt** mit $n - 1$ Freiheitsgraden: $\sim \chi^2_{n-1}$
4. \bar{X} und s^2 sind unabhängig.
5. $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ ist **t -verteilt** mit $n - 1$ Freiheitsgraden: $\sim t_{n-1}$

Diese Ergebnisse werden wir im Folgenden oft brauchen, insbesondere Aussagen 1, 2 und 5. Später werden wir auch auf Aussage 3 treffen.

Für die Interessierten*. Ohne exakte Beweise: Aussage 1 und 2 sind schon bekannt. Aussage 3 ist einigermassen plausibel aufgrund 4.4.1. Aussage 5 folgt ebenso aus 4.4.2

und der Umformung $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{s/\sigma} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{n-1}{\sigma^2}s^2}}$. Die Unabhängigkeit in Aussage 4 ist der überraschendste Teil.

7.3 Konstruktion von Schätzern*

Woher wissen wir denn, dass \bar{X} ein “guter” Schätzer ist für μ ? Oben haben wir nur begründet, dass \bar{X} ein *vernünftiger* Schätzer ist. Die zwei wichtigsten Methoden, um Schätzer zu konstruieren, sind

- Methode der kleinsten Quadrate
- Maximum-Likelihood-Methode

Wir schauen uns diese beiden Methoden ein bisschen genauer an. Es geht hier aber darum, das (philosophische) Prinzip dahinter zu verstehen, wir werden selber natürlich keine Schätzer “konstruieren” müssen!

Betrachten wir dazu wieder unser grundlegendes Schätzproblem: Seien X_1, X_2, \dots, X_n i.i.d. (unabhängig und gleichverteilte) Zufallsvariablen mit $E(X_i) = \mu$. Die Verteilung selber sei zunächst unbekannt. Ziel sei nun eine Schätzung von μ mit den Beobachtungen x_1, x_2, \dots, x_n .

7.3.1 Methode der kleinsten Quadrate

Der Kleinst-Quadrate-Schätzer (Least Squares, LS-Schätzer) von μ ist der Wert, der die Fehler-Quadratsumme

$$ss(\mu) = \sum_{i=1}^n (x_i - \mu)^2 \tag{7.3.1}$$

minimiert. Der LS-Schätzer von μ ist $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$, also $\hat{\mu}_{LS} = \frac{1}{n}(X_1 + \dots + X_n)$ ¹.

Beispiel. Seien die Daten z.B. $X = -1.679, 7.023, 6.885, 5.011, 2.485, 4.498, 3.366, 4.206, 3.577, 4.541, 3.386, 5.78, 4.776, -0.549, 2.9, 4.661, 2.804, 1.476, 5.527, 5.682$ mit $\bar{x} = 3.818$.

Der Schätzer von μ mit minimalen sum of squares ist $\hat{\mu}_{LS} = \bar{X}$ und der Schätzwert damit 3.818. Es gibt keine andere Statistik (keine andere Grösse, die wir aus den Daten berechnen können), für die die Summe der Abweichungsquadrate $ss(\mu)$ kleiner ist. Abbildung 7.1 zeigt die Summe der Abweichungsquadrate bei diesen Daten als Funktion von μ .

¹Beweis: Solche Probleme löst man in der Regel, indem man $ss(\mu)$ nach μ ableitet und nullsetzt: $-2(\sum_{i=1}^n x_i - n\mu) = 0$, daraus folgt für den LS-Schätzer: $\hat{\mu}_{LS} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$.

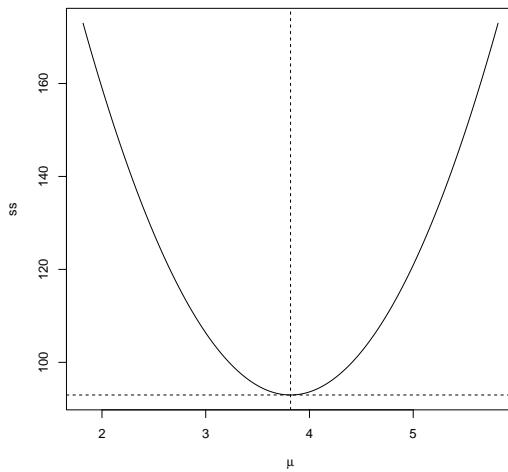


Abbildung 7.1: Summe der Abweichungsquadrate als Funktion von μ

Häufige Anwendung: Die Methode wird sehr häufig in der *linearen Regression* gebraucht. Wir haben dieses Prinzip bereits unter [11.2](#) angetroffen².

7.3.2 Maximum-Likelihood-Methode

Eine wichtige Grösse – für viele gar die fundamentalste Grösse in der Statistik – ist die *Likelihood*. Diese haben wir im Zusammenhang mit dem Bayes-Theorem schon kennengelernt.

Sei θ ein Parameter(vektor) in einem statistischen Modell. Bei Normalverteilung sind die Parameter, die das Modell spezifizieren, Erwartungswert μ und Varianz σ^2 , also $\theta = (\mu, \sigma^2)$. Die Maximum-Likelihood-Methode (ML) ist die wichtigste Methode für die Schätzung von unbekannten Grössen aus Daten. Dazu muss aber die Verteilung der Daten bekannt sein.

Seien die X_i zum Beispiel i.i.d. normalverteilte Zufallsvariablen, $X_i \sim \mathcal{N}(\mu, \sigma^2)$ mit bekannter Varianz, d.h. σ^2 ist bekannt. Ziel sei wieder die Schätzung von μ . Beobachtet wurden nun Daten x_1, \dots, x_n . Die Daten liegen also vor und uns interessiert, wie wahrscheinlich diese Daten unter verschiedenen Werten des Parameters

²Dort hängt der Erwartungswert $E(Y_i) = \mu_i$, von gesuchten Parametern ab, z.B. den Parametern α und β in der einfachen lineare Regression mit $Y_i = \alpha + \beta x_i + \epsilon_i$ und $E(Y_i) = \mu_i = \alpha + \beta x_i$. Minimiert wird hier

$$\sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \mu_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2,$$

um Schätzwerte für α und β zu finden.

sind. Die Wahrscheinlichkeitsdichte wird nun als Funktion von μ aufgefasst werden. Diese Funktion heisst dann *Likelihood-Funktion* $L(\mu)$,

$$L(\mu) = \prod_{i=1}^n p(x_i). \quad (7.3.2)$$

Oft braucht man auch die *Log-Likelihood-Funktion* $l(\mu)$ (siehe Rechenregeln für Logarithmus),

$$l(\mu) = \log L(\mu) = \sum_{i=1}^n \log p(x_i). \quad (7.3.3)$$

Die Likelihood misst die Plausibilität der *vorhandenen* Daten unter verschiedenen μ . Ziel ist zu wissen, welche Werte von μ durch die vorhandenen Daten am meisten gestützt werden. Da wir i.i.d. haben, ist die Likelihood von μ , also die Wahrscheinlichkeit der Daten gegeben μ , gleich dem Produkt der n Likelihoods $p(x_i)$ (mit $p(x_i)$ als den Dichten unter Normalverteilung³).

Der ML-Schätzer von μ ist $\hat{\mu}_{ML} = \bar{x}$, da \bar{x} die Likelihood maximiert⁴. Das ist – in diesem Fall – identisch mit $\hat{\mu}_{LS}$. Der Vorteil der ML-Methode liegt in den besseren Eigenschaften der Schätzfunktion (Konsistenz und Effizienz). Der Nachteil der Methode ist, dass die Verteilung bekannt sein muss. Wichtig: Bei Normalverteilung geben die LS-Schätzung und die ML-Schätzung dieselben Resultate. Im Allgemeinen ist das aber nicht der Fall.

Beispiel Binomialverteilung. Bei $n = 10$ Patienten wurde das (dichotome) Ansprechen auf eine Therapie gemessen. Die Daten sind:

$$y_1 = 0, y_2 = 1, y_3 = 0, y_4 = 0, y_5 = 0, y_6 = 1, y_7 = 0, y_8 = 1, y_9 = 0, y_{10} = 0$$

mit 1 for “Erfolg” und 0 for “Misserfolg”. π sei die unbekannte Erfolgswahrscheinlichkeit. Wir haben $k = 3$ Erfolge und $n - k = 7$ Misserfolge. Die Likelihood (unter dem Binomial-Modell) ist⁵

$$L(\pi) = p(y_1, \dots, y_{10} | \pi) = \pi^3(1 - \pi)^{10-3} = \pi^3(1 - \pi)^7. \quad (7.3.4)$$

mit log-Likelihood

$$l(\pi) = \log L(\pi) = 3 \log \pi + 7 \log(1 - \pi). \quad (7.3.5)$$

Abbildung 7.2 zeigt die Likelihood- und die log-Likelihood-Funktion. Man kann zeigen, dass der MLE für die Erfolgswahrscheinlichkeit π in diesem Modell $\hat{\pi} = k/n$ ist, 0.3 in diesem Fall. Für die Interessierten:

Beweis. Um den MLE für π zu finden, setzen wir die Ableitung (Steigung) von $l(\pi)$ auf 0 (Zur Erinnerung:

³Dichte Normalverteilung $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$.

⁴Die log-Likelihood ist $l(\mu) = \sum_{i=1}^n \log p(x_i) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$. Ableiten und Nullsetzen von $l(\mu)$ gibt $\frac{1}{\sigma^2} (\sum_{i=1}^n x_i - n\mu) = 0$ und für den Maximum-Likelihood-Schätzer $\hat{\mu}_{ML} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$.

⁵Wir können $\binom{n}{x}$ vernachlässigen, da diese Grösse nicht von θ abhängt.

$$\frac{d}{dx} \log(x) = 1/x.$$

$$\frac{3}{\pi} - \frac{7}{1-\pi} = 0, \quad (7.3.6)$$

daraus folgt $\hat{\pi} = 3/10$. □

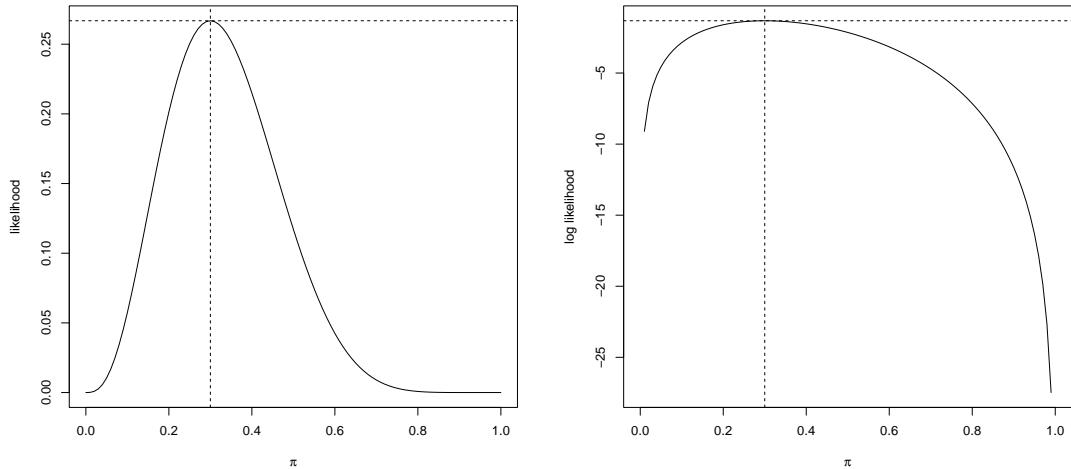


Abbildung 7.2: Likelihood und log-Likelihood für Erfolgswahrscheinlichkeit π , nach Beobachtung von 3 Erfolgen in 10 Versuchen.

7.4 Intervallschätzung

Bei unserem grundlegenden Schätzproblem kennen wir jetzt die Verteilung von \bar{X} , die gleichzeitig die Schätzfunktion, die Statistik ist für μ . Diese Verteilung hat Erwartungswert μ und Standardabweichung σ/\sqrt{n} . Die Punktschätzung liefert uns den Schätzwert $\hat{\mu} = \bar{x}$, der im Regelfall nicht mit dem wahren Wert μ identisch ist. Wir wollen nun die Präzision dieser Schätzung einbauen. Ein Mass für die Präzision haben wir bereits unter dem Begriff des *Standardfehlers* kennengelernt. Ein zusätzlicher Weg ist die Intervallschätzung.

7.4.1 Konfidenzintervall

Bei der *Intervallschätzung* ist das Ziel die Konstruktion von einem *Konfidenzintervall* (KI) für den unbekannten Parameter μ . Man versucht, die Wahrscheinlichkeit, mit der das Verfahren ein Intervall liefert, das den wahren Wert *nicht* enthält, durch eine vorgegebene *Irrtumswahrscheinlichkeit* α zu kontrollieren. Übliche Werte für diese Irrtumswahrscheinlichkeit sind $\alpha = 0.10$, $\alpha = 0.05$ oder $\alpha = 0.01$. Entsprechend ist

$1 - \alpha$ die Wahrscheinlichkeit, dass das Verfahren ein Intervall liefert, das den wahren Wert μ enthält. $1 - \alpha$ nennt man die *Überdeckungswahrscheinlichkeit*. Diese ist entsprechend 0.9, 0.95 und 0.99.

Im Abschnitt 7.2.1 haben wir gesehen, dass \bar{X} approximativ normalverteilt ist mit Mittelwert μ und Standardabweichung σ/\sqrt{n} . Also ist die standardisierte Form, die Statistik

$$T = \frac{\bar{X} - \mu}{\text{se}(\bar{X})}, \quad (7.4.1)$$

approximativ z-verteilt, $T \stackrel{a}{\sim} \mathcal{N}(0, 1)$, siehe Abbildung 4.10. Der Wert der Statistik T hängt von den Daten und vom unbekannten Parameter μ ab, die Verteilung von T hängt hingegen *nicht* von dem zu schätzenden Parameter μ ab. Das ist sehr hilfreich, denn die Standardnormalverteilung haben wir schon “verinnerlicht”. Quantile der z-Verteilung kennen wir, siehe B.2. Wir haben dazu die `qnorm()`-Funktion kennengelernt, die solche Tabellen eigentlich überflüssig macht. Wichtige Quantile sind $z_{0.5} = 0$ (Median), $z_{0.025} = -1.96$ und $z_{0.975} = 1.96$. Die beiden letzteren Quantile werden oft benutzt, da sich zwischen ihnen 95% aller Werte befinden.

Als Wiederholung in R nochmals die Verteilungsfunktion und deren Umkehrfunktion:

```
qnorm(p = c(0.025, 0.975), mean = 0, sd = 1)
## [1] -1.96 1.96

pnorm(q = c(-1.96, 1.96), mean = 0, sd = 1)
## [1] 0.025 0.975
```

Allgemein ist die Wahrscheinlichkeit, dass T im Intervall zwischen $z_{\alpha/2}$ und $z_{1-\alpha/2}$ ist, gleich $1 - \alpha$.

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\text{se}(\bar{X})} \leq z_{1-\alpha/2}\right) = 1 - \alpha, \quad (7.4.2)$$

und somit folgt durch Umformung, dass

$$\bar{X} \pm z_{1-\alpha/2} \cdot \text{se}(\bar{X}) \quad (7.4.3)$$

ein $(1 - \alpha)$ -Konfidenzintervall für μ darstellt. Ein solches Intervall überdeckt μ mit $1 - \alpha$ frequentistischer Wahrscheinlichkeit. Dazu mehr im folgenden Beispiel.

95%-KI: ±2-Regel. Wir wollen ein approximativeres 95%-KI für μ berechnen. Ein approximativeres 95%-KI für μ wird gemäss 7.4.3 konstruiert, mit $\alpha = 0.05$ und $z_{1-\alpha/2} = 1.96$. Das Intervall

$$\boxed{\bar{X} \pm 1.96 \cdot \text{se}(\bar{X})}, \quad \text{mit } \text{se}(\bar{X}) = \frac{s}{\sqrt{n}} \quad (7.4.4)$$

stellt also ein 95% KI für μ dar. Für \bar{X} wird dann die empirische Realisation \bar{x} eingesetzt. Da \bar{X} eine Zufallsgrösse ist, ist jedes Konfidenzintervall – bei hypothetisch neuer Stichprobennahme – verschieden. Der wahre Wert μ wird aber in 95% der Fälle (frequentistisch, “on the long run”) vom jeweiligen Konfidenzintervall umspannt. Wir hoffen beim Betrachten eines KI, dass wir nicht eines von zwanzig (entspricht $\alpha = 5\%$) erwischt haben, das den wahren Wert μ nicht überdeckt. Das ist der Preis, den wir zahlen für das Schliessen auf Unbekanntes! Das ist der Preis der Induktion.

Das Intervall stellt einen *Bereich von plausiblen Werten* für μ dar. Mit 95% frequentistischer Wahrscheinlichkeit wird μ vom Konfidenzintervall überdeckt. Alle Werte, die das Intervall enthält, sind plausibel. Alle Werte, die das Intervall nicht enthält, sind nicht plausibel.

Beispiel. Konfidenzintervall “by hand”. $n = 144$ Beobachtungen x_1, x_2, \dots, x_{144} einer Variable X wurden gemacht. Die aus den 144 Zahlen berechneten Kennwerte seien $\bar{x} = 100$ und $s = 24$. Gesucht ist ein approximatives 95%-Konfidenzintervall für μ . Ein solches Intervall ist dann gegeben durch $\bar{X} \pm 1.96 \times \text{se}(\bar{X})$. Durch Einsetzen bekommen wir folgendes 95%-Konfidenzintervall für μ :

$$100 \pm 1.96 \times \frac{24}{\sqrt{144}} = [100 - 1.96 \times 2, 100 + 1.96 \times 2] = [96.08, 103.92].$$

```
100 - qnorm(p = 0.975) * 24/sqrt(144)
100 + qnorm(p = 0.975) * 24/sqrt(144)
100 + c(-1, 1) * qnorm(p = 0.975) * 24/sqrt(144) ##als Einzeiler (aufgrund der Symmetrie von z)
```

Die (frequentistische) Wahrscheinlichkeit, dass dieses Intervall die unbekannte Grösse μ überdeckt, ist 95%. Die (frequentistische) Wahrscheinlichkeit, dass dieses Intervall den unbekannten Parameter nicht überdeckt, ist 5%, die a priori bestimmte Irrtumswahrscheinlichkeit α .

Wenn man “von Hand” Konfidenzintervalle berechnet, wird häufig die Zahl 1.96 durch Zahl 2 ersetzt (± 2 -Regel). In unserem Beispiel wäre das dann $[100 - 2 \times 2, 100 + 2 \times 2] = [96, 104]$, damit sind wir immer noch gut im Rennen.

Exakte Konfidenzintervalle bei Normalverteilung. Bei *beliebig verteiltem* Merkmal in der Population und $n > 30$ ist die Schätzstatistik 7.4.1 gemäss dem zentralen Grenzwertsatz approximativ standardnormalverteilt. Ist $n < 30$, so ist aber zu fordern, dass das Merkmal X in der Population normalverteilt ist. Gilt diese Voraussetzung, dann ist die Statistik T *exakt t-verteilt mit $n - 1$ Freiheitsgraden* (siehe 7.2.2 und Abbildung 4.14 für die Familie der t -Verteilungen). Wir ersetzen dann das 0.975-Quantil der z -Verteilung durch das 0.975-Quantil der t -Verteilung mit dem entsprechenden Freiheitsgrad $df = n - 1$ (df : degrees of freedom). Die entsprechenden Quantile waren früher in t -Tabellen zu konsultieren, heute brauchen wir dafür die inzwischen lieb gewonnenen Quantilfunktionen in R, für die t -Verteilung ist das `qt()`.

Für n gross nähern sich die Quantile der t -Verteilung denjenigen der z -Verteilung an. Abbildung 7.3 zeigt t -Verteilungen mit 5, 10 und 500 Freiheitsgraden, und die entsprechenden 0.975-Quantile.

```
qt(0.975, c(5, 10, 500))

## [1] 2.57 2.23 1.96
```

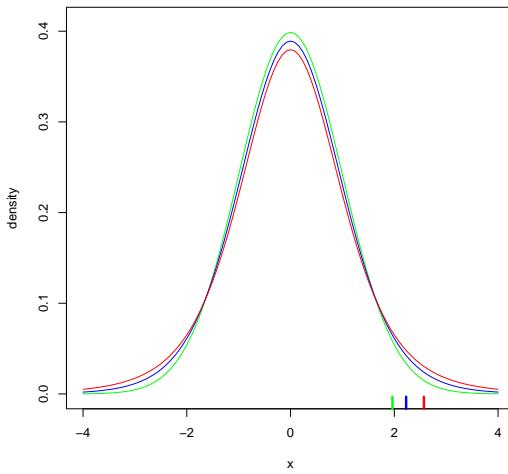


Abbildung 7.3: t -Verteilungen mit jeweiligem 0.975-Quantil. Die grüne Verteilung ist praktisch die Normalverteilung (t -Verteilung mit 500 Freiheitsgraden).

Das exakte 95%-Konfidenzintervall für μ ist also (also unter der Bedingung der Normalverteilung von X)

$$\bar{X} \pm t_{0.975, n-1} \cdot \frac{s}{\sqrt{n}}, \quad (7.4.5)$$

mit $t_{0.975, n-1}$ als dem 0.975-Quantil der t -Verteilung mit $n - 1$ Freiheitsgraden. Bei kleinem n wächst diese Grösse. Für $n = 10$ z.B. ist $t_{0.975, n-1} = 2.26$. Bei n gross ist $t_{0.975, n-1}$ praktisch identisch mit $z_{0.975}$, dem entsprechenden Quantil der z -Verteilung (1.96). Für grosses n benutzt man daher in der Praxis oft approximative Konfidenzintervalle, es sind dann keine Annahmen über die Verteilung von X nötig.

Beispiel 1. Wir haben dieselben Daten (Kennwerte) wie oben, aber wir wissen zusätzlich, dass die Daten normalverteilt sind. Ein exaktes 95%-Konfidenzintervall können wir mit folgendem Code bekommen. Das approximative Konfidenzintervall war leicht enger.

```
100 + c(-1, 1) * qt(p = 0.975, df = 144 - 1) * 24/sqrt(144)
```

Beispiel 2. In der Praxis werden wir Rohdaten direkt zur Verfügung haben, nicht nur die Kennwerte \bar{x} und s : Wir simulieren $n = 20$ Beobachtungen aus einer Normalverteilung mit $\mu = 3$ und $\sigma = 10$ (In einer Simulation kennen wir den Daten-generierenden Mechanismus). Dann beschreiben wir die simulierten Daten mit `psych::describe()`.

```
set.seed(5) ## damit wir alle dieselben Zufallszahlen haben
n <- 20 ## Stichprobengröße
mu <- 3 ## mu
sigma <- 10 ## sigma
X <- rnorm(n, mu, sigma) ## siehe ?rnorm
psych::describe(X)

##   vars n mean sd median trimmed mad min max range skew kurtosis    se
## X1    1 20  0.19 9.29 -0.79  -0.29 6.75 -18.8 20.1    39 0.43 -0.07 2.08
```

Ein exaktes 95% Konfidenzintervall können wir berechnen wie bisher. Wir haben Glück. Wir haben nicht eines von 20 KI erwischt, das $\mu = 3$ nicht überdeckt. Das wissen wir aber natürlich nur, weil wir in Simulationen die wahren Werte der Parameter kennen.

```
mean(X) + c(-1, 1) * qt(0.975, n - 1) * sd(X)/sqrt(n)

## [1] -4.16 4.54
```

Später werden wir Konfidenzintervalle direkt über den sogenannten t -Test konstruieren. Im folgenden Output interessiert uns im Moment nur das Konfidenzintervall. Wir kommen im Kapitel 8 auf den t -Test zurück.

```
t.test(X)
```

Wenn n klein ist und keine Normalverteilung vorausgesetzt werden kann, kommen sogenannte *robuste Methoden* zum Zuge. Ein nicht-parametrisches Analogon zum t -Test werden wir später unter dem Namen Wilcoxon-Test kennenlernen, der auch ein nicht-parametrisches KI herausgibt.

```
wilcox.test(X, conf.int = TRUE)
```

7.4.2 Frequentistische Wahrscheinlichkeit*

Anhand einer *Simulation* wollen wir das Prinzip vom Schätzen einmal “von hinten” aufrollen. In Simulationen kennen wir den Mechanismus, der die Daten generiert. Wir ziehen eine Stichprobe von $n = 30$ (Abbildung 7.4 links) respektive von $n = 300$

(rechts) einer beliebig verteilten Variable X mit Erwartungswert $\mu = 4$ und berechnen den empirischen Mittelwert \bar{x} (Obere Teilabbildung).

Dieses Verfahren wird nun 10'000-mal wiederholt, dabei “entsteht” gemäss zentralem Grenzwertsatz ein approximativ normalverteilter Mittelwert (die Stichprobenverteilung des Mittelwerts). Dies ist in der mittleren Teilabbildung ersichtlich.

Untere Teilabbildung: Berechnen wir für die ersten 100 Stichproben (der Übersicht halber, statt der 10'000) gemäss (7.4.5) 95%-Konfidenzintervalle für μ , so werden im Schnitt 5 von 100 Konfidenzintervallen den wahren Wert $\mu = 4$ nicht umspannen, dies entspricht der Irrtumswahrscheinlichkeit von $\alpha = 5\%$.

Die Präzision der Schätzung wird nun mit wachsendem n grösser; dies ist in der Unterschiedlichkeit der Breite der Konfidenzintervalle der beiden Abbildungen ersichtlich. Der Standardfehler wird mit grösserem n kleiner und die Konfidenzintervalle werden somit enger. Aber immer noch werden im Schnitt 5 von 100 Konfidenzintervallen den wahren Wert $\mu = 4$ nicht umspannen (die roten gestrichelten).

Nun, im Gegensatz zu dieser Simulation haben wir natürlich in einer Studie immer nur *eine* Stichprobe und kennen μ *nicht*. Wir haben nur *ein* Konfidenzintervall. Es bleibt dann nur zu hoffen, dass dieses zu den 95% “guten” gehört, dass man eines erwischt hat, das die wahre Grösse μ überdeckt. Das ist der *Preis*, den wir zahlen, um von der Stichprobe auf die Population verallgemeinern zu können, die oben eingeführte Irrtumswahrscheinlichkeit, das Risiko α .

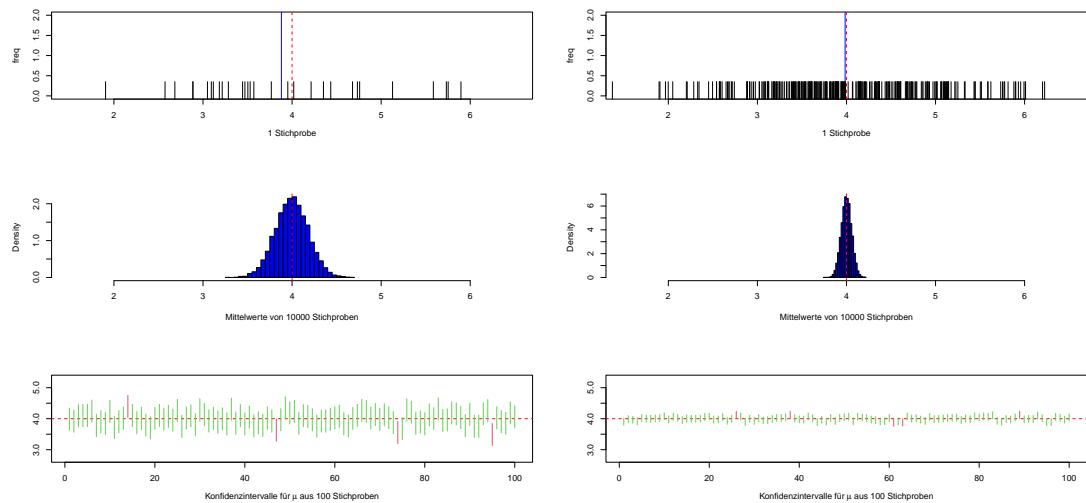


Abbildung 7.4: Frequentistische Wahrscheinlichkeit (siehe Text für Erklärung)

Long-run-Irrtumswahrscheinlichkeit. Simulation mit R*.

Folgender Code (für die Interessierten!) zieht $nsim = 1000$ mal aus einer Normalverteilung mit $\mu = 3$ und $\sigma = 10$ je $n = 144$ Werte. Dann wird aus den jeweiligen Daten ein 95%-KI konstruiert und geschaut, ob $\mu = 3$ im KI ist. Das Resultat ist die Proportion von Intervallen, die $\mu = 3$ nicht enthalten. Die empirische relative Häufigkeit sollte in der Nähe von $\alpha = 0.05$ sein.

```
set.seed(4)
n <- 144 #sample size
mu <- 3 #true mean
sigma <- 10 #true sd
nsim = 1000 #number of simulations
X <- matrix(0, nrow = n, ncol = nsim) #initialize matrix
ok <- rep(0, nsim) #initialize result
for (i in 1:nsim) {
  X[, i] <- rnorm(n, mu, sigma)
  ci <- mean(X[, i]) + c(-1, 1) * qt(0.975, n - 1) * sd(X[, i])/sqrt(n)
  ok[i] <- (mu < ci[2]) & (mu > ci[1])
}
prop.table(table(ok)) #proportion of intervals that do not cover the true value

## ok
##      0      1
## 0.053 0.947
```

7.5 Schätzen von anderen Größen

Wir haben in diesem Kapitel die Grundzüge des Schätzens kennengelernt. Wir haben das für die grundlegende Schätzung für den Erwartungswert μ einer Zufallsvariable X gemacht. Je nach Kontext können aber ganz andere Größen von Interesse sein, z.B. eine Korrelation ρ zwischen zwei Variablen X und Y , eine wahre Steigung β in einem Regressionsmodell, eine wahre Proportion π bei Abstimmungen, eine wahre Sensitivität Sn eines diagnostischen Tests, usw.⁶

In den nächsten Abschnitten werden wir beispielhaft das Schätzen 1) eines Zwischengruppeneffekts bei kontinuierlichem Outcome und 2) eines relativen Risikos oder Chancenverhältnisses bei dichotomer abhängiger Variable vertiefen.

7.5.1 Schätzen bei kontinuierlichen Daten und 2 Gruppen

Quantität von Interesse. In klinischen Studien mit zwei Gruppen Intervention und Control mit kontinuierlichem Outcome ist die Quantität von Interesse sehr oft der *wahre Zwischengruppenunterschied* δ ,

$$\delta = \mu_I - \mu_C. \quad (7.5.1)$$

⁶Der Autor stellt im Prozess von Abschlussarbeiten (MSc Physiotherapie) häufig fest, dass Studierende Mühe bekunden, diese *Quantity of interest* zu benennen.

Schätzer. Im Gegensatz zu μ ist jetzt δ die interessierende Grösse. Der Schätzer, die Statistik $\hat{\delta}$ für δ ist gerade der empirische Zwischengruppenunterschied, der auch hier eine Zufallsvariable ist,

$$\hat{\delta} = \hat{\mu}_I - \hat{\mu}_C = \bar{X}_I - \bar{X}_C. \quad (7.5.2)$$

Beliebige Verteilung und grosse Stichprobe. Bei *beliebiger Verteilung* (aber i.i.d.) des Merkmals in der Population und grossen Stichproben ($n_I + n_C > 60$) ist auch diese Statistik gemäss dem zentralen Grenzwertsatz approximativ normalverteilt,

$$\hat{\delta} \stackrel{a}{\sim} \mathcal{N}\left(\delta, \text{se}^2(\hat{\delta})\right), \quad \text{standardisiert: } \frac{\hat{\delta} - \delta}{\text{se}(\hat{\delta})} \stackrel{a}{\sim} \mathcal{N}(0, 1) \quad (7.5.3)$$

mit Standardfehler

$$\text{se}(\hat{\delta}) = \sqrt{\frac{\sigma_I^2}{n_I} + \frac{\sigma_C^2}{n_C}}. \quad (7.5.4)$$

Ein approximativer 95% Konfidenzintervall für den wahren Effekt δ ist dann

$$95\% \text{ KI} : \hat{\delta} \pm z_{0.975} \cdot \text{se}(\hat{\delta}). \quad (7.5.5)$$

σ_I^2 und σ_C^2 in 7.5.4 sind aber fast immer unbekannt. Es gibt dann zwei Fälle zu unterscheiden:

1. Sie werden sie als homogen angenommen ($\sigma_I^2 = \sigma_C^2$), dann ist der Standardfehler

$$\text{se}(\hat{\delta}) = \sqrt{\frac{1}{n_I} + \frac{1}{n_C}} \times s_{\text{pooled}} = \sqrt{\frac{1}{n_I} + \frac{1}{n_C}} \times \sqrt{\frac{(n_I - 1)s_I^2 + (n_C - 1)s_C^2}{n_I + n_C - 2}}. \quad (7.5.6)$$

2. Sie werden sie als heterogen angenommen ($\sigma_I^2 \neq \sigma_C^2$), dann ist der Standardfehler

$$\text{se}(\hat{\delta}) = \sqrt{\frac{s_I^2}{n_I} + \frac{s_C^2}{n_C}}. \quad (7.5.7)$$

Exakte Konfidenzintervalle bei Normalverteilung. Für ein exaktes Konfidenzintervall bei Normalverteilung des Merkmals wird in 7.5.5 $z_{0.975}$ ersetzt durch

- $t_{0.975, df}$ mit $df = n_I + n_C - 2$ bei homogenen Varianzen⁷
- $t_{0.975, df}$ mit $df = \frac{(s_I^2/n_I + s_C^2/n_C)^2}{(s_I^2/n_I)^2/(n_I-1) + (s_C^2/n_C)^2/(n_C-1)}$ bei heterogenen Varianzen⁸.

⁷ df ist dann Anzahl Beobachtungen minus Anzahl Parameter (μ_I und μ_C)

⁸Ohne Beweis, R wird das für uns im Hintergrund machen

Kleine Stichprobe und keine Normalverteilung. Bei $n_I + n_C < 60$ ist immer zu fordern, dass das Merkmal in der Population normalverteilt ist. Sonst werden *robuste Verfahren* angewandt.

Beispiel. Eine Studie will die obige Quantität von Interesse δ mit folgenden Daten schätzen

```
d.rct <- read.csv("https://raw.githubusercontent.com/mcdrl65/StatsRsource/master/Data/simpleRct.csv",
  stringsAsFactors = TRUE)
str(d.rct)

## 'data.frame': 60 obs. of 3 variables:
## $ id      : int 1 2 3 4 5 6 7 8 9 10 ...
## $ outcome: num 115.9 110.4 134.9 74.6 143.9 ...
## $ group   : Factor w/ 2 levels "Control","Intervention": 2 2 2 2 2 2 2 2 2 2 ...

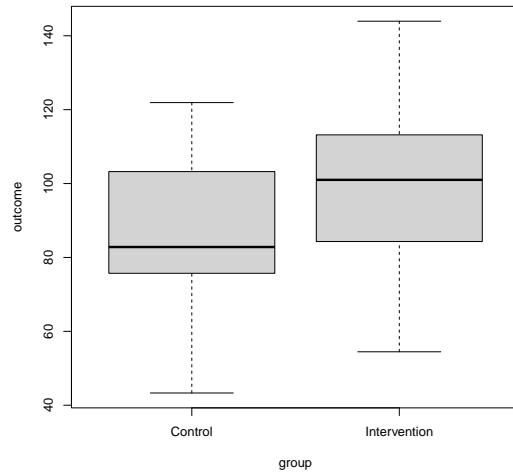
psych::headTail(d.rct)

##      id outcome      group
## 1     1 115.86 Intervention
## 2     2 110.45 Intervention
## 3     3 134.92 Intervention
## 4     4  74.57 Intervention
## ... ...
## 57    57  79.55     Control
## 58    58  61.94     Control
## 59    59  98.16     Control
## 60    60 103.24     Control

by(d.rct$outcome, d.rct$group, psych::describe)

## d.rct$group: Control
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1   1 30 87.20 20.6   82.8    87.8 22.1 43.3 122 78.6 -0.17 -0.79 3.75
## -----
## d.rct$group: Intervention
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1   1 30 99.4 20.6   101     99 21.5 54.5 144 89.5  0.03 -0.55 3.77

boxplot(outcome ~ group, data = d.rct)
```



Mittel und Standardabweichung in beiden Gruppe sind

```
int <- d.rct$outcome[d.rct$group == "Intervention"] #Vektor mit Outcomes der Interventionsgruppe
cont <- d.rct$outcome[d.rct$group == "Control"] #Vektor mit Outcomes der Kontrollgruppe
xbar1 <- mean(int) #mean Int
xbar2 <- mean(cont) #mean Cont
n1 <- n2 <- 30 #n1 = n2
s1 <- sd(int) #sd Int
s2 <- sd(cont) #sd Cont
```

Wir nehmen an, dass σ_I^2 und σ_C^2 homogen sind, aber unbekannt. Der beobachtete Zwischengruppenunterschied ist $\hat{\delta} = \bar{x}_I - \bar{x}_C = 12.475$ und der Standardfehler von $\hat{\delta}$ (7.5.6) ist $se(\hat{\delta}) = 5.32$:

```
spooled <- sqrt((s1^2 * (n1 - 1) + s2^2 * (n2 - 1))/(n1 + n2 - 2))
se <- sqrt(1/n1 + 1/n2) * spooled
se

## [1] 5.32
```

Ein (approximatives) 95%-Konfidenzintervall für δ ist (7.5.5):

```
(xbar1 - xbar2) + c(-1, 1) * qnorm(0.975) * se

## [1] 2.05 22.90
```

Ein exaktes 95%-Konfidenzintervall (bei Normalverteilung) für δ ist

```
(xbar1 - xbar2) + c(-1, 1) * qt(0.975, n1 + n2 - 2) * se

## [1] 1.83 23.12
```

Später werden wir auch für diese Analysen Konfidenzintervalle direkt über den sogenannten t -Test konstruieren. Im nachfolgenden Output interessiert im Moment nur das Konfidenzintervall für $\mu_I - \mu_C$. Beachte die Vorzeichen, im t -Test wird `Control` minus `Intervention` gerechnet (alphabetische Reihenfolge der Stufen von `group`).

```
t.test(d.rct$outcome ~ d.rct$group, var.equal = TRUE) #für homogene Varianzen
t.test(d.rct$outcome ~ d.rct$group) #für heterogene Varianzen ('Welch-Test')
```

Interpretation. Das Konfidenzintervall, das nach (7.5.5) konstruiert wird, ist *ein Bereich von plausiblen wahren Therapieeffekten* δ . Mit 95% frequentistischer Wahrscheinlichkeit überdeckt ein solches Konfidenzintervall den wahren Effekt δ , mit 5% frequentistischer Wahrscheinlichkeit nicht. Der Nulleffekt $\delta = 0$ ist in diesem Beispiel *nicht* plausibel, weil dieser Wert nicht im Konfidenzintervall ist.

7.5.2 Schätzen bei dichotomen Outomes und zwei Gruppen

In klinischen Studien mit *dichotomer* abhängiger Variable (Ereignis eingetroffen oder nicht) ist die interessierende Quantität oft ein

- Chancenverhältnis: $OR = Odds_1/Odds_2$
- log Chancenverhältnis: $\log OR = \text{logit}_1 - \text{logit}_2$
- ein relatives Risiko: $RR = R_1/R_2$
- log relatives Risiko: $\log RR = \log R_1 - \log R_2$
- eine Risikodifferenz: $RD = R_1 - R_2$

Das erste und zweite Effektmaß braucht man eher in Fall-Kontroll-Studien, das dritte und vierte eher in Kohortenstudien (inklusive RCT's). Der Wertebereich von Chancenverhältnissen und relativen Risiken ist $(0, \infty)$, der Wertebereich von $\log OR$ und $\log RR$ ist $(-\infty, +\infty)$, der Wertebereich von RD ist $(-1, +1)$.

$\log OR = 0$ (äquivalent mit $OR = 1$), $\log RR = 0$ (äquivalent mit $RR = 1$) oder $RD = 0$ bedeuten jeweils, dass beide Gruppen die gleiche Chance, dasselbe Risiko haben, entsprechen also dem “Nulleffekt”.

Beispieldaten aus Studie. Eine randomisierte kontrollierte Studie untersuchte das Risiko für selbst eingeschätzte Inkontinenz nach zwei verschiedenen Therapien I (Beckenbodentraining) versus C (Kontroll), siehe [33].

Zuerst wollen wir ein paar wichtige R-Funktionen im Umgang/Beschreibung mit/von kategorialen Daten anschauen. Wir kommen in Kapitel 10 nochmals auf entsprechende Funktionen zurück. Wir reproduzieren die Daten der Studie (Tabelle 2, S. 316, Zeile 36 wk) mit

```
morkved.mat <- matrix(c(79, 74, 100, 48), byrow = TRUE, nrow = 2, dimnames = list(Intervention = c("Control", "Training"), Outcome = c("negativ", "positiv")))
morkved.mat

##           Outcome
## Intervention negativ positiv
##   Control      79      74
##   Training     100      48
```

Mit `marginSums()` können wir die Randsummen berechnen⁹. Da es sich um eine prospektive Kohortenstudie handelt (mit randomisierten Gruppen), sind die Zeilen (Behandlungsart als unabhängige Variable) „fixiert über das Design“:

```
## marginSums(morkved.mat)
marginSums(morkved.mat, margin = 1)  ##gegeben Zeilen

## Intervention
## Control Training
##      153      148

## marginSums(morkved.mat,margin=2) ##gegeben Kolonnen
```

Mit `addmargins()` können wir die Summen anhängen (`margin=1` für Zeilen anhängen und `margin=2` für Kolonnen anhängen)

```
morkved.tabm <- addmargins(morkved.mat, margin = 2)
morkved.tabm

##           Outcome
## Intervention negativ positiv Sum
##   Control      79      74 153
##   Training     100      48 148
```

Wenn wir die Proportionen wollen, können wir das mit `proportions()` berechnen¹⁰

```
morkved.prop <- proportions(morkved.mat, margin = 1)  ##gegeben Zeilen
morkved.prop

##           Outcome
## Intervention negativ positiv
##   Control      0.516    0.484
##   Training     0.676    0.324
```

Damit haben wir die Proportionen von Inkontinenz in der Trainings- und Kontrollgruppe.

⁹Der Befehl heisst auch `margin.table()`

¹⁰Der Befehl heisst auch `prop.table()`

Normalapproximation des logarithmierten Odds Ratio. Mit dem zentralen Grenzwertsatz lässt sich zeigen, dass die Zufallsgrösse $\log \widehat{OR}$ asymptotisch normalverteilt ist gemäss

$$\log \widehat{OR} \stackrel{a}{\sim} \mathcal{N}(\log OR, se^2(\log \widehat{OR})). \quad (7.5.8)$$

Man kann zeigen, dass der Standardfehler dieses Schätzers

$$se(\log \widehat{OR}) = \sqrt{1/x_I + 1/x_C + 1/(n_I - x_I) + 1/(n_C - x_C)} \quad (7.5.9)$$

ist¹¹, mit x_I als der Anzahl Events in der Trainingsgruppe und x_C als der Anzahl Events in der Kontrollgruppe. Wir machen nun eine Punkt- und Intervallschätzung für das wahre OR :

```
(OR <- ((48/100)/(74/79))) ## beob. OR
## [1] 0.512

logOR <- log(OR) ## beob. log OR
se.logOR <- sqrt(1/48 + 1/100 + 1/74 + 1/79) ## Standardfehler
CI.logOR <- logOR + c(-1, 1) * qnorm(0.975) * se.logOR ## plus minus 2-Regel
(CI.OR <- exp(CI.logOR)) ## zurück auf Originalskala mit der Exponentialfunktion
## [1] 0.321 0.818
```

Der Grund für den Umweg über die log-Transformation ist folgender: Wenn man direkt mit der OR arbeitet und zwei Standardfehler (jetzt von \widehat{OR}) Unsicherheit um das beobachtete Chancenverhältnis aufspannt, kann es sein, dass man ein Konfidenzintervall bekommt, bei dem die untere Grenze des Intervalls negativ ist. OR sind aber immer positiv.

Somit ist $[0.32, 0.82]$ ein 95% KI für das wahre OR .

Normalapproximation des logarithmierten RR. Auch die Zufallsgrösse $\log \widehat{RR}$ ist approximativ normalverteilt,

$$\log \widehat{RR} \stackrel{a}{\sim} \mathcal{N}(\log RR, se^2(\log \widehat{RR})). \quad (7.5.10)$$

Der Standardfehler für diesen Schätzer ist

$$se(\log \widehat{RR}) = \sqrt{1/x_I + 1/x_C - 1/n_I - 1/n_C}. \quad (7.5.11)$$

```
(RR <- ((48/148)/(74/153)))
```

¹¹mit der sogenannten Delta-Methode

```
## [1] 0.671

logRR <- log(RR)
se.logRR <- sqrt(1/48 + 1/74 - 1/148 - 1/153)
CI.logRR <- logRR + c(-1, 1) * qnorm(0.975) * se.logRR
(CI.RR <- exp(CI.logRR))

## [1] 0.505 0.891
```

Somit ist $[0.5, 0.89]$ ein 95% KI für das wahre RR . Vergleiche dies mit den Resultaten in der Publikation [33] (Tabelle 2, S. 316, Zeile 36 wk). Eine Punktschätzung ergab dort $\widehat{RR} = 0.67$, d.h., das Risiko einer Inkontinenz war für die Interventionsgruppe ungefähr 2/3 vom Risiko der Kontrollgruppe. Die Interventionsgruppe hatte 1/3 weniger Risiko für Inkontinenz als die Kontrollgruppe. Diese Grösse nennt man die relative Risikoreduktion $\widehat{RRR} = \frac{\widehat{R}_I - \widehat{R}_c}{\widehat{R}_c} = \widehat{RR} - 1 = 0.67 - 1 = -1/3$. Ein 95%-Konfidenzintervall für das wahre relative Risiko ergab $[0.50, 0.89]$. Die (frequentistische) Wahrscheinlichkeit, dass ein solches Intervall den wahren Effekt überdeckt, ist 95 Prozent.

Der Nulleffekt entspricht bei diesem Effektmass dem Wert 1 (bei logarithmiertem Effektmass 0). Dieser Wert ist *nicht* im Konfidenzintervall enthalten und demnach nicht plausibel.

Man sagt dann: Der Nulleffekt wird mit einem *statistischen Test* auf α -Niveau verworfen. Das führt uns im nächsten Abschnitt zur wichtigen Dualität zwischen Konfidenzintervallen und statistischen Tests.

OR versus RR.* Man kann zeigen, dass

$$RR = OR / (1 - R_c + OR \times R_c), \quad (7.5.12)$$

mit R_c als dem Risiko in der Kontrollgruppe. Wenn $OR < 1$, dann $RR > OR$, wenn $OR > 1$, dann $RR < OR$. Wenn $OR = 1$, dann $OR = RR$. Das RR ist also immer näher beim neutralen Wert 1 als das OR . In unserem Beispiel war $OR = 0.512$ und $RR = 0.671$. Anbei ein Beispiel einer Funktion in R, um (7.5.12) zu veranschaulichen:

```
ORtoRR <- function(or, rc) {
  rr <- or/(1 - rc + or * rc)
  return(rr)
}

ORtoRR(or = 0.512, rc = 74/153)

## [1] 0.67
```

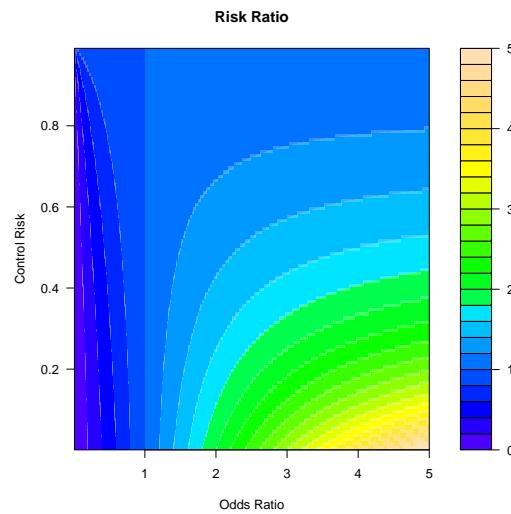


Abbildung 7.5: Transformation von OR zu RR

Implementation: package epiR. Am Anfang ist es immer sinnvoll, als Training Größen “von Hand” zur berechnen, wie wir das oben gemacht haben. Es gibt dann im Universum der R-packages aber meistens eine Funktion in einer package, wo die Berechnungen implementiert sind. So sind z.B. in der Funktion `epi.2by2` aus der package `epiR` obige Berechnungen implementiert. In der Funktion muss die Kreuztabelle mit den Daten übergeben werden:

```

library(epiR)
epi.2by2(morkved.mat)

##          Outcome +    Outcome -    Total     Inc risk *      Odds
## Exposed +       79        74     153      51.6      1.07
## Exposed -      100        48     148      67.6      2.08
## Total          179       122     301      59.5      1.47
##
## Point estimates and 95% CIs:
## -----
## Inc risk ratio                  0.76 (0.63, 0.92)
## Odds ratio                     0.51 (0.32, 0.82)
## Attrib risk *                 -15.93 (-26.87, -5.00)
## Attrib risk in population *   -8.10 (-17.46, 1.26)
## Attrib fraction in exposed (%) -30.86 (-58.19, -8.25)
## Attrib fraction in population (%) -13.62 (-23.71, -4.35)
## -----
## Test that OR = 1: chi2(1) = 7.924 Pr>chi2 = 0.00
## Wald confidence limits
## CI: confidence interval
## * Outcomes per 100 population units

```

Die Hilfunktion dieser Funktion, `help(epi.2by2)`, ist eine gute Wiederholung der Grundlagen der Epidemiologie.

7.6 Dualität mit Hypothesentests

Konfidenzintervalle stellen Bereiche dar von plausiblen Werten für einen unbekannte Grösse (oft ein Parameter eines parametrischen Modells). Wir können also, indem wir Konfidenzintervalle betrachten, gleichzeitig auch *Hypothesen testen*.

Eine zentrale Dualität ist folgende: **Ein $(1 - \alpha)$ -Konfidenzintervall für einen Parameter θ beinhaltet alle Werte des unbekannten Parameters, den ein entsprechender statistischer Test auf α -Niveau nicht verwirft.**

Eine Intervallschätzung für den Zwischengruppenunterschied δ ergab ein 95%-Konfidenzintervall [15.37, 24.62]. Daraus folgt:

- Die Hypothese “ $\delta = 16$ ” ist plausibel. Ein Test auf 5%-Niveau würde diese Hypothese nicht verwerfen.
- Die Hypothese “ $\delta = 63$ ” ist nicht plausibel. Ein Test auf 5%-Niveau würde diese Hypothese verwerfen.
- Die Hypothese “ $\delta = 0$ ” ist nicht plausibel. Ein Test auf 5%-Niveau würde diese Hypothese verwerfen.

Eine ganz spezifische zu testende Aussage wie “ $\delta = 0$ ” nennt man häufig die Nullhypothese über δ (hier “der Effekt ist Null”). Allgemein ist die Nullhypothese die Hypothese “to be nullified”, die zu verwerfende Hypothese; das *Null* steht also nicht notwendigerweise für den *Nulleffekt*.

Ein sogenannter *Hypothesentest* würde nun diese Nullhypothese $H_0 : \delta = 0$ gerade mit einer 5%-igen Irrtumswahrscheinlichkeit verwerfen. Man sagt dann, dass der beobachtete Effekt *statistisch signifikant* von Null verschieden ist.

Im Beispiel mit dichotomen Outcome ergab eine Intervallschätzung für das relative Risiko RR ein 95%-Konfidenzintervall $[0.50, 0.89]$. Daraus folgt:

- Die Hypothese “ $RR = 0.6$ ” ist plausibel. Ein Test auf 5%-Niveau würde diese Hypothese nicht verwerfen.
- Die Hypothese “ $RR = 0.9$ ” ist nicht plausibel. Ein Test auf 5%-Niveau würde diese Hypothese verwerfen.
- Die Hypothese “ $RR = 1$ ” ist nicht plausibel. Ein Test auf 5%-Niveau würde diese Hypothese verwerfen.

p-Werte. Alternativ wird die *statistische Signifikanz* in Studien sehr häufig über sogenannte *p*-Werte dargestellt. Wenn *p*-Werte kleiner sind als die festgesetzte Irrtumswahrscheinlichkeit α (meistens 5%), dann wird die Nullhypothese verworfen.

Ausblick. In den folgenden Kapiteln werden wir vertieft auf das *Testen von Hypothesen* eingehen. Es ist aber oft eleganter und sogar genügend, über das Betrachten des Bereichs in Konfidenzintervallen zu “testen”. Dabei können wir beliebige Hypothesen testen. Wie wir sehen werden, wird bei klassischen Hypothesentests häufig “nur” die “Null”-Nullhypothese getestet, die Hypothese vom Null-Nulleffekt.

Kapitel 8

Hypothesentests

Hypothesentests sind Verfahren, mit denen bestimmte *Hypothesen* über Parameter oder Kennwerte einer Verteilung, z.B. den Erwartungswert μ , getestet werden. Man will jetzt ganz *spezifische* Eigenschaften von Merkmalen der Population (Theorie) durch stichprobenartig erhobene Daten (Empirie) bestätigen oder widerlegen. Beim Schätzen haben wir nach einem *Bereich von plausiblen Werten* des Parameters gefragt. Jetzt fragen wir, ob ein *bestimmter Wert* des Parameters plausibel ist.

Im Gegensatz zum Schätzen werden also *a priori* festgelegte Hypothesen oder *Aussagen* anhand von Daten getestet. Hypothesen sind nichts anderes als Aussagen, die man empirisch *überprüfen* kann, wie z.B. " $\mu = \mu_0$ " mit μ_0 als dem postulierten Wert für den Parameter μ .

Konkrete Daten x_1, \dots, x_n betrachten wir als Realisationen von i.i.d. Zufallsvariablen X_1, \dots, X_n . In der *parametrischen* Statistik wird zusätzlich a priori angenommen, dass die Zufallsvariablen aus einer Familie vorgegebener Wahrscheinlichkeitsverteilungen stammen, die bis auf *endlich* viele Parameter eindeutig *bestimmt* sind. Einige Verteilungen haben wir in Kapitel 4 kennengelernt. So hatte z.B. die Normalverteilung die Parameter μ und σ^2 .

Die bekanntesten statistischen Analyseverfahren sind parametrische Verfahren. Das wohl berühmteste davon ist der *t*-Test.

8.1 Prinzipien von statistischen Tests

Für einen statistischen Test wird die Vermutung über einen Parameter formal in einer *Nullhypothese* festgehalten, die wir mit H_0 bezeichnen. Es kann zum Beispiel die Gültigkeit von

$$H_0 : \theta = \theta_0$$

im Vergleich zur *Alternative*

$$H_1 : \theta \neq \theta_0$$

interessieren. Es ist wichtig, nochmals zu betonen, dass die Nullhypothese *nicht* die Hypothese vom Nulleffekt sein muss (weil man häufig $\theta_0 = 0$ setzt), sondern die

Hypothese, die verworfen, getestet werden soll, die Hypothese “to be nullified”. Die Nullhypothese ist häufig die Hypothese, die der Skeptiker vertritt, der Status quo. *Null-Nullhypotesen* sind Spezialfälle von Nullhypotesen.

Mit einer Zufallsstichprobe X_1, \dots, X_n können wir über einen Hypothesentest die Nullhypothese beibehalten oder verwerfen. Beim Entscheiden über die Richtigkeit oder Falschheit einer Nullhypothese können wir zwei Arten von *Fehlern* machen. Diese sind in Tabelle 8.1 dargestellt.

Zum einen einen *Fehler 1. Art*, wenn wir H_0 zu Unrecht ablehnen, zum anderen einen *Fehler 2. Art*, wenn H_0 irrtümlicherweise beibehalten wird. Die relative Häufigkeit des Fehler bei sehr häufiger Anwendung ist dann die Wahrscheinlichkeit des Fehlers.

Die Wahrscheinlichkeit des Auftretens eines Fehlers 1. Art nennt man *Irrtumswahrscheinlichkeit* oder *Signifikanzniveau* des Tests. Diese Grösse haben wir schon bei der Einführung von Konfidenzintervallen kennengelernt und wird auch hier mit α bezeichnet. Der Fehler 2. Art β macht erst bei *spezifischen* Alternativhypotesen Sinn, wir gehen darauf in Kapitel 8.6 ein.

		Wahrheit	
		Alternativ $H_1(+)$	Null $H_0(-)$
Test+	TPR	α	
	β	TNR	

Tabelle 8.1: Entscheidungen aufgrund eines Testresultats: vier Arten von Entscheidungen: richtig positiv Rate (TPR), richtig negativ Rate (TNR), falsch negativ (β) und falsch positiv (α).

Für einen statistischen Test braucht man natürlich empirische Daten und dann eine *Statistik* T , also wieder eine Grösse, die wir aus den Daten berechnen können und die sensibel ist für das Testproblem. Wir haben diesen Begriff in (7.1.1) eingeführt¹.

Für einen parametrischen Test müssen wir die Verteilung der Statistik (oder des Schätzers) kennen. Ein allgemeiner Zugang lieferte der Zentrale Grenzwertsatz (7.2.1). Viele (standardisierte) Teststatistiken folgen (asymptotisch) einer (Standard)-Normalverteilung. Für die Normalverteilung haben wir exakte Aussagen (7.2.2) kennengelernt.

Beim grundlegenden Schätzproblem (Schätzung von μ) haben wir die Statistik $\hat{\mu} = \bar{X}$ eingeführt. Die standardisierte Version dieser Statistik (siehe 7.4.1) war $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$. Der Wert einer solchen Statistik hängt ab vom unbekannten μ , die Verteilung hingegen ist unabhängig von μ . Bei Normalverteilung der X_i ist diese Statistik t -verteilt mit $n - 1$ Freiheitsgraden, bei beliebiger Verteilung der X_i und n gross ist diese Statistik approximativ normalverteilt. Letzteres nehmen wir im Folgenden an. Der t -Test folgt später.

¹Zur Erinnerung: Alle Grössen, die wir aus Daten berechnen können, wird Statistik genannt.

Wir können jetzt unsere Hypothese $H_0 : \mu = \mu_0$ zusammen mit den Daten in die Statistik (eine Zufallsvariable) einsetzen und dann Wert der Statistik $T = t$ berechnen.

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (8.1.1)$$

Dieser Wert wird dann mit dem *kritischen* Wert der Statistik, $T = t_{krit}$, verglichen. t_{krit} stellt beim einseitigen Testen das $z_{1-\alpha}$ -Quantil, beim zweiseitigen Testen das $z_{1-\alpha/2}$ -Quantil dar. Wenn nun (beim zweiseitigen Test) T betragsmäßig grösser ist als der kritische Wert t_{krit} , wird H_0 verworfen. Diesen *Entscheid* des Verwerfens macht man dann mit einem α -*Risiko*. Das Risiko α kommt daher, da ja unter H_0 die Statistik – die Zufallsgrösse T – gerade mit Wahrscheinlichkeit α in diesen Bereich fällt.

p -Wert. Alternativ wird auch über sogenannte p -*Werte* entschieden:

Der p -Wert ist die Wahrscheinlichkeit – gegeben Gültigkeit des Modells der Nullhypothese – den beobachteten Wert der Statistik oder einen in Richtung der Alternative “extremeren” Wert zu erhalten.

Für die zweiseitige Alternative ist der p -Wert definiert als

$$p\text{-Wert} = \Pr(|T| \geq |t| \mid H_0). \quad (8.1.2)$$

H_0 wird dann verworfen, wenn der p -Wert kleiner ist als α , was gleichbedeutend ist mit $|t| > |t_{krit}|$; die Irrtumswahrscheinlichkeit α war ja definiert als die Wahrscheinlichkeit, dass die Statistik in die Verwerfungszone fällt (dass man H_0 verwirft), wenn H_0 wahr ist, also

$$\alpha = \Pr(|T| \geq |t_{krit}| \mid H_0). \quad (8.1.3)$$

Abbildung 8.1 zeigt die Verteilung einer standardnormalverteilten Teststatistik $T = Z$ unter dem Modell $H_0 : \mu = \mu_0$. Die schraffierte Fläche ist α , diese Grösse muss *a priori* festgelegt werden und wird im Umfeld der Gesundheitswissenschaften oft 5% gesetzt. Diesen 5% entsprechen die schon bekannten kritischen Quantile $z_{0.025} = -1.96$ und $z_{0.975} = 1.96$. Die schwarze Fläche ist der p -Wert, der *a posteriori*, nachdem man Daten hat, bestimmt werden kann, je nachdem wie gross der Wert der Statistik $T = t$ ist. Die aus den Daten berechnete Statistik wäre hier extremer als 1.96, liegt also im Verwerfungsbereich, im Bereich $T > t_{krit}$. Die Nullhypothese wird verworfen, weil $p < \alpha$ ist.

8.2 z -Test

Der z -Test (oder Gauss-Test) ist ein grundlegender Test für den Erwartungswert einer Zufallsgrösse. Dazu haben wir eine konkrete Stichprobe, x_1, \dots, x_n , aus n

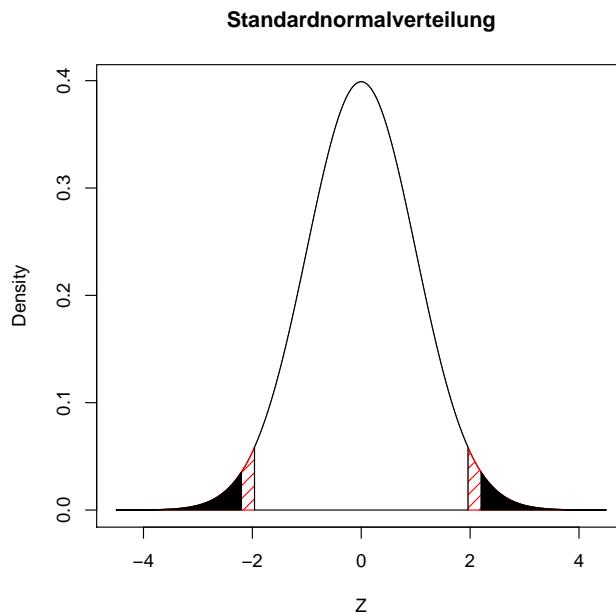


Abbildung 8.1: Verteilung einer Teststatistik, zweiseitiger Test, α (a priori, schraffiert) und p -Wert (a posteriori, schwarz).

i.i.d. X_1, \dots, X_n Stichprobenvariablen (n sei ‘gross’). Wir nehmen zudem an, dass die wahre Streuung σ bekannt ist, diese muss also nicht über die empirische Standardabweichung s geschätzt werden.

- Wir spezifizieren die Alternativ- und die Nullhypothese. Es gibt drei Möglichkeiten, eine mit *zweiseitiger* Alternative und zwei mit *einseitigen* Alternativen:

$$\begin{aligned} H_0 : \mu &= \mu_0, & H_1 : \mu &\neq \mu_0 \\ H_0 : \mu &\leq \mu_0, & H_1 : \mu &> \mu_0 \\ H_0 : \mu &\geq \mu_0, & H_1 : \mu &< \mu_0 \end{aligned}$$

- Wir legen dann das *Signifikanzniveau* fest, z.B. $\alpha = 0.05$, und betrachten die *Teststatistik*, die standardisierte Version von \bar{X} , die sensibel ist für das Testproblem,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}. \quad (8.2.1)$$

- Unter der Nullhypothese $H_0 : \mu = \mu_0$ gilt: Z ist approximativ standardnormalverteilt, $Z \stackrel{a}{\sim} \mathcal{N}(0, 1)$. Die Annahme der Normalverteilung der Daten X_i wurde oben nicht gemacht. Wenn aber die X_i zusätzlich normalverteilt

sind, dann gilt die exakte Standardnormalverteilung².

- Der *Verwerfungsbereich* für die Teststatistik Z für die *zweiseitige* Alternative $H_1 : \mu \neq \mu_0$ ist dann $Z \leq z_{\alpha/2}$ und $Z \geq z_{1-\alpha/2}$. Wir verwerfen H_0 , falls der realisierte Wert z in diesen Verwerfungsbereich fällt. Bei *einseitigen* Alternativen ist der Verwerfungsbereich auch einseitig, man benutzt dann das $(1 - \alpha)$ -Quantil anstatt dem $(1 - \alpha/2)$ -Quantil. Meistens wird $\alpha = 0.05$ gesetzt, dann haben wir wieder $z_{1-\alpha/2} = -z_{\alpha/2} = 1.96$ für die zweiseitige Alternative und $z_{1-\alpha} = -z_\alpha = 1.64$ für die einseitige Alternative.
- Wird H_0 verworfen, nennt man das Resultat *statistisch signifikant auf α -Niveau*.

8.3 t-Test

Beim z -Test brauchten wir für die z -Statistik (8.2.1) die *wahre Standardabweichung* σ . In der Praxis ist aber σ meistens nicht bekannt und muss aus den Daten (mit der empirischen Standardabweichung) geschätzt werden (siehe (5.5.6)):

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (8.3.1)$$

Darum macht man auch meistens einen t -Test anstatt obigem Gauss-Test.

Hypothesen. Wir spezifizieren die Null- und die Alternativhypothese. Es gibt wieder drei Möglichkeiten:

$$\begin{aligned} H_0 : \mu &= \mu_0, & H_1 : \mu &\neq \mu_0. \\ H_0 : \mu &\leq \mu_0, & H_1 : \mu &> \mu_0 \\ H_0 : \mu &\geq \mu_0, & H_1 : \mu &< \mu_0 \end{aligned}$$

Signifikanzniveau. Wir legen dann das *Signifikanzniveau* fest, z.B. $\alpha = 0.05$.

Statistik. Wir betrachten die *Teststatistik*, eine standardisierte Version von \bar{X} , die sensibel ist für das Testproblem,

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}. \quad (8.3.2)$$

Statt der unbekannten wahren Standardabweichung σ steht jetzt die empirische Standardabweichung s im Nenner.

²Wir erinnern daran, dass Z nichts anderes darstellt als eine standardisierte Version von \bar{X} . Da \bar{X} normalverteilt ist mit Mittelwert μ und Streuung σ/\sqrt{n} , ist Z standardnormalverteilt.

Verteilung der Statistik. Die Statistik T ist wieder Zufallsvariable und ist identisch mit 7.4.1. Die Verteilung von T ist bekannt und sie ist unabhängig vom Parameter. Unter der Nullhypothese $H_0 : \mu = \mu_0$ gilt:

- Bei beliebiger Verteilung und n “gross”: T ist approximativ standardnormalverteilt,

$$T \xrightarrow{a} \mathcal{N}(0, 1).$$

- Bei Normalverteilung von X : T ist exakt t -verteilt mit $n - 1$ Freiheitsgraden,

$$T \sim t_{n-1}.$$

Die t -Verteilung ist wie die Normalverteilung eine *symmetrische* Verteilung um 0. Sie ist aber *langschwänziger* als die Standardnormalverteilung. Bei grossem n kann man die t -Verteilung durch eine $\mathcal{N}(0, 1)$ -Verteilung approximieren (siehe Abbildung 4.14).

- Bei n klein und ohne Normalverteilung müssen *robuste* Methoden eingesetzt werden (dazu mehr später).

Entscheid. H_0 wird verworfen, wenn $|t| > t_{0.975, n-1}$ für die zweiseitige Alternative ($t_{0.975, n-1}$ entspricht 1.96, wenn n gross) und $|t| > t_{0.95, n-1}$ für die einseitige Alternative ($t_{0.95, n-1}$ entspricht 1.64, wenn n gross). Diese kritischen Werte kann man mit `qt(p=, df=)` bestimmen. Wird H_0 verworfen, nennt man das Resultat *statistisch signifikant auf α -Niveau*.

Beispiel: t -Test “by hand”. Gegeben sei eine beliebig verteilte i.i.d. Zufallsstichprobe X_1, \dots, X_{144} mit Kennwerten $\bar{x} = 100$, $s = 20$, $n = 144$. Es soll die Nullhypothese $H_0 : \mu = 95$ getestet werden.

Die Statistik $T = \frac{\bar{X} - 95}{\text{se}(\bar{X})}$ ist unter $H_0 : \mu = 95$ approximativ standardnormalverteilt (da n “gross”). Einsetzen ergibt $t = \frac{100 - 95}{20/12} = 3$. Der Wert der Statistik ist extremer als der kritische Wert 1.96. Also können wir die Hypothese $H_0 : \mu = 95$ verwerfen. Die Alternativhypothese $H_1 : \mu \neq 95$ wird angenommen.

Dasselbe Ergebnis erhalten wir – eleganter – über die bekannte Konstruktion eines 95%-Konfidenzintervalls für μ . Ein solches Intervall ist gegeben durch $100 \pm 1.96 \times 5/3 = [96.73, 103.26]$.

```
100 + c(-1, 1) * 1.96 * 20/12
## [1] 96.7 103.3
```

Der H_0 -Wert $\mu = 95$ ist nicht im Konfidenzintervall enthalten, er ist nicht plausibel und wird daher verworfen. Aufgrund der Dualität zwischen dem Zugang über das Schätzen und über das Testen führen beide Methoden zum selben Resultat.

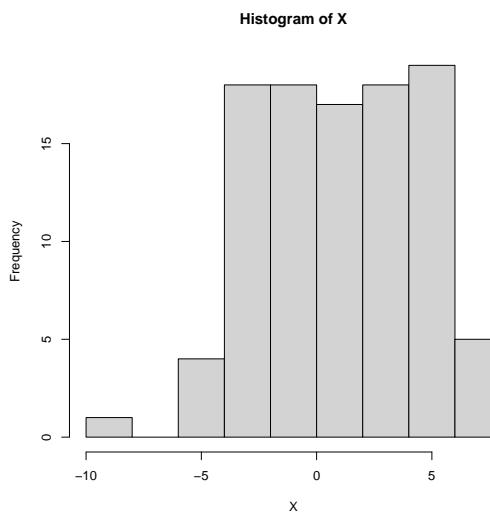
Implementation in R. Die Implementation für den Einstichproben-*t*-Test ist `t.test()`. Konsultiere dazu `help(t.test)` und studiere die Argumente der Funktion:

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Bei einem Einstichproben-*t*-Test muss man (mindestens) zwei Argumente übergeben, den Vektor mit den Daten und die Hypothese bezüglich μ . (Defaultmäßig führt diese Funktion einen zweiseitigen Test aus mit $\mu_0 = 0$).

Wir machen im Folgenden einen Einstichproben-*t*-Test mit simulierten normalverteilten Daten (mit $\mu = 1$ und $\sigma = 4$):

```
set.seed(3)
X <- rnorm(100, 1, 4) ##Simulation aus Normalverteilung mit mu=1 und sigma=4 (n=100)
hist(X)
```



```
psych::describe(X) ##Beschreiben

##   vars   n mean    sd median trimmed  mad   min   max range skew kurtosis    se
## X1     1 100  1.04  3.42    1.14    1.09  4.34 -8.06  7.94    16 -0.13   -0.77  0.34
```

Wir konstruieren noch einmal vorab “von Hand” ein exaktes 95% Konfidenzintervall für μ .

```
mean(X) + c(-1, 1) * qt(p = 0.975, df = length(X) - 1) * sd(X)/sqrt(length(X)) ## 95%KI

## [1] 0.365 1.724
```

Wir testen nun $H_0 : \mu = 0$ zweiseitig mit $\alpha = 0.05$

```
t.test(X) ##Testen von H_0: mu=0: DEFAULT-EINSTELLUNG ist mu=0!
```

```
## 
##  One Sample t-test
##
## data: X
## t = 3.05, df = 99, p-value = 0.0029
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.36469 1.72360
## sample estimates:
## mean of x
##      1.0441
```

Wir reproduzieren die Quantitäten im Output von Hand.

```
(se <- sd(X)/sqrt(length(X))) ## StandardError se, sehen wir auch im describe() Output
## [1] 0.342

(t <- (mean(X) - 0)/se) ##t-Statistik
## [1] 3.05

(df <- length(X) - 1) ##Freiheitsgrad
## [1] 99
```

Da die Statistik grösser ist als 0, ist die Wahrscheinlichkeit von Werten grösser als t

```
(1 - pt(q = t, df = df))
## [1] 0.00147
```

Die Wahrscheinlichkeit von Werten der Statistik, die *betragsmässig* grösser sind als t (zweiseitiger Test!)

```
(1 - pt(q = t, df = df)) * 2 #mal 2, da zweiseitig
## [1] 0.00294
```

Wir können natürlich auch eine andere Hypothese testen, z.B. $H_0 : \mu = 1$:

```
t.test(X, mu = 1)
```

```
## 
## One Sample t-test
##
## data: X
## t = 0.129, df = 99, p-value = 0.9
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
## 0.36469 1.72360
## sample estimates:
## mean of x
## 1.0441
```

Reproduziere die Quantitäten im Output von Hand (Übung).

Siehe <https://rstudio.zhaw.ch/rsconnect/content/85> für Simulation eines Samples und Analyse über einen t -Test. Dort kann man n Beobachtungen simulieren aus einer $\mathcal{N}(\mu, \sigma^2)$ -Verteilung mit anschliessendem t -Test für $H_0 : \mu = 0$. Der t -Test ist dort schon in der Terminologie der linearen Modelle aufgeschrieben (β steht für μ , siehe 13.1.5), auf diese Notation kommen wir später zurück.

8.4 Zwei-Stichproben Problem. t -Test für unabhängige Stichproben

Der bekannteste statistische Test ist der t -Test für zwei unabhängige Stichproben, z.B. für das Testen eines *Zwischengruppeneffekts*

$$\delta = \mu_X - \mu_Y$$

in einer klinischen Studie, wobei μ_X und μ_Y die unbekannten Erwartungswerte in den beiden Populationen darstellen. Es geht dabei also um den Vergleich der beiden Erwartungswerte von zwei Merkmalen X und Y (analog zu 7.5.1).

Es gibt wieder drei Möglichkeiten, wie wir die Hypothesen aufstellen können. Bei zweiseitiger Alternative schreiben wir

$$H_0 : \delta = \delta_0, \quad H_1 : \delta \neq \delta_0. \quad (8.4.1)$$

Bei einer einseitigen Alternative hingegen

$$H_0 : \delta \geq \delta_0, \quad H_1 : \delta < \delta_0 \quad (8.4.2)$$

oder

$$H_0 : \delta \leq \delta_0, \quad H_1 : \delta > \delta_0, \quad (8.4.3)$$

mit δ_0 als dem postulierten Effekt unter der Nullhypothese. Vielfach wird $\delta_0 = 0$ gesetzt, d.h. man testet die Hypothese, dass der Unterschied 0 ist. Wir betrachten im Folgenden den Fall (8.4.1). Bezüglich der einseitigen Alternative sei aber noch erwähnt, dass man

z.B. im Fall von (8.4.3) $H_0 : \delta \leq \delta_0$ verwerfen kann, sobald man $H_0 : \delta = \delta_0$ verwerfen kann.

Wir wissen, dass die Zwischengruppendifferenz $\hat{\delta} = \bar{X} - \bar{Y}$ der beste Schätzer für den Unterschied in den Populationen, für unsere **Quantität von Interesse** $\delta = \mu_X - \mu_Y$, ist.

Unsere Schätzstatistik ist dann wieder eine normierte Version vom Schätzer, nämlich

$$T = \frac{\hat{\delta} - \delta_0}{\text{se}(\hat{\delta})}. \quad (8.4.4)$$

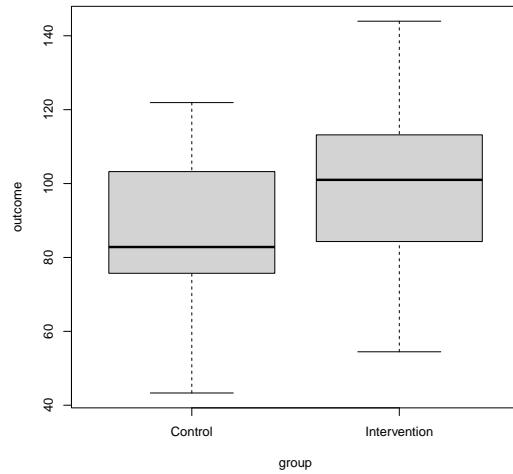
Analog zu 7.5.1 unterscheiden wir bezüglich dem Standardfehler der Zwischengruppendifferenz (der Grösse $\text{se}(\hat{\delta})$) zwischen homogenen und heterogenen Varianzen. Wenn X und Y normalverteilt sind, gilt (unter H_0) die exakte t -Verteilung mit entsprechendem Freiheitsgrad. Wenn wir aber Stichproben mit $n_X + n_Y > 60$ haben, dürfen die Variablen X und Y beliebig verteilt sein, die Teststatistik ist dann nach dem zentralen Grenzwertsatz immer noch approximativ standardnormalverteilt.

Der Ablehnungsbereich ist dann im zweiseitigen Test (bei Normalverteilung) $|T| > t_{1-\alpha/2}$ und beim einseitigen Test $T > t_{1-\alpha}$, respektive $T < t_\alpha$.

Wir wollen aber hier nochmals betonen, dass das “Null” von *Nullhypothese* nicht notwendigerweise den Nulleffekt meint, sondern die zu verworfene Hypothese, *the hypothesis to be nullified*. Wir werden in 8.7 auf das philosophische Problem von *Strohmann-Nullhypotesen* eingehen.

Beispiel t-Test für Zwischengruppenunterschied. Wir kehren zurück zum Beispiel 7.5.1 zur Schätzung eines Zwischengruppenunterschiedes aus dem vorigen Kapitel.

```
d.rct <- read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/simpleRct.csv",
  stringsAsFactors = TRUE)
boxplot(outcome ~ group, data = d.rct)
```



Zuerst berechnen wir noch einmal (wie in Kapitel 7.5.1):

```
int <- d.rct$outcome[d.rct$group == "Intervention"] #Vektor mit Outcomes der Interventionsgruppe
cont <- d.rct$outcome[d.rct$group == "Control"] #Vektor mit Outcomes der Kontrollgruppe
xbar1 <- mean(int)
xbar2 <- mean(cont)
n1 <- n2 <- 30
s1 <- sd(int)
s2 <- sd(cont)
spooled <- sqrt((s1^2 * (n1 - 1) + s2^2 * (n2 - 1))/(n1 + n2 - 2))
se <- sqrt(1/n1 + 1/n2) * spooled
df <- n1 + n2 - 2
(conf <- (xbar1 - xbar2) + c(-1, 1) * qt(0.975, df) * se)

## [1] 1.83 23.12
```

Unsere Nullhypothese sei, dass der Unterschied der Erwartungswerte beider Gruppen gleich 0 ist, also $H_0 : \delta = 0$. Die (standardisierte) Teststatistik ist dann

```
(t <- ((xbar1 - xbar2) - 0)/se)

## [1] 2.34
```

Die t -Statistik ist extremer als das 0.975-Quantil einer t -Verteilung mit $df = 58$ Freiheitsgraden,

```
(crit <- qt(p = 0.975, df = df))

## [1] 2
```

Der p -Wert für den zweiseitigen Test ist über `pt()` zu haben. Da die Statistik grösser ist als 0, ist der p -Wert die Wahrscheinlichkeit von Werten grösser gleich t (und dann mal zwei, weil zweiseitig), also

```
(pValue <- (1 - pt(q = t, df = df)) * 2)

## [1] 0.0225
```

Der p -Wert ist in folgender Abbildung als Fläche unter der Dichte dargestellt. Es ist sehr unwahrscheinlich ($p = 0.022$), Daten wie diese (zusammengefasst in der Statistik t) oder extremer (in Richtung der Alternative, $|T| > |t|$) zu beobachten, wenn das H_0 -Modell gilt (d.h. wenn $\delta = 0$).

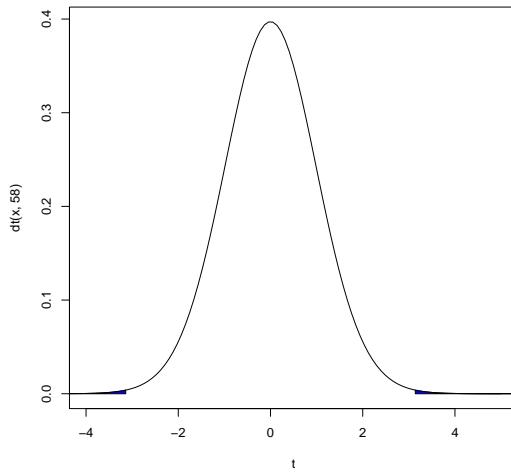


Abbildung 8.2: t -Verteilung mit $df = 58$ mit p -Wert bei beobachteter Statistik $t = 3.14$

Implementation in R. Dazu brauchen wir wieder die Funktion `t.test()`. Konsultiere wieder `help(t.test)`. Hier wurden die Default-Werte für `alternative`, `mu`, `paired`, `conf.level` übernommen. Da wir homogene Varianzen annehmen, muss dieses Argument gesetzt werden. Defaultmäßig ist `var.equal=FALSE` gesetzt.

In der *formula*-Syntax schreibt man

```
t.test(d.rct$outcome ~ d.rct$group, var.equal = TRUE)

##
##  Two Sample t-test
##
## data: d.rct$outcome by d.rct$group
## t = -2, df = 58, p-value = 0.02
## alternative hypothesis: true difference in means between group Control and group Intervention is not equal to 0
## 95 percent confidence interval:
## -23.12 -1.83
## sample estimates:
##      mean in group Control mean in group Intervention
##                  87.0                  99.4
```

Beachte die Vorzeichen, im *t*-Test wird `Control` minus `Intervention` gerechnet (es gilt die alphabetische Reihenfolge der Stufen von `group`).

In der *default*-Syntax für den Test braucht man die Daten pro Gruppe (als Vektoren) als Argumente. Gerechnet wird dann wieder erstes Argument minus zweites Argument.

```
t.test(x = int, y = cont, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  int and cont
## t = 2, df = 58, p-value = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    1.83 23.12
## sample estimates:
## mean of x mean of y
##      99.4      87.0
```

Annahmen vom *t*-Test. Es gibt Tests auf Normalverteilung (`shapiro.test()`) und Tests auf homogene Varianzen (`var.test()` bei zwei Stichproben).

```
shapiro.test(int) #H0: Normalverteilt
shapiro.test(cont)
var.test(d.rct$outcome ~ d.rct$group) #H0: Gleiche Varianzen
```

In diesem Fall zeigen die Tests nicht an. Normalverteilung und homogene Varianzen werden also *nicht* verworfen. Solche Tests sind aber **umstritten**.

Tests für Annahmen sind umstritten. Viele Autoren empfehlen, keine solchen Tests zu machen, um über die Strategie (z.B. *t*-Test mit homogenen Varianzen versus *t*-Test mit heterogenem Varianzen (“Welch-Test”) versus nicht-parametrische Alternative) zu entscheiden. Das würde nämlich voraussetzen, dass diese Tests ungefähr eine Teststärke von 1 haben *für alle* Stichprobengrößen (mehr zur Teststärke unten), dass das *Vor*-testen auf Normalität den α -Fehler der Testprozedur nicht beeinflusst und dass nicht-parametrische Tests klar weniger effizient sind. Alle diese Aussagen sind aber falsch.

Residuenanalyse. Wir werden später im Rahmen von *linearen Modellen* zurückkommen auf die Beurteilung von Modellannahmen. Wir werden dort den *t*-Test als einen Spezialfall eines solchen Modells einführen und im Rahmen von *Residuenanalysen* die Modellannahmen “testen”. Sinnvoller ist es also zunächst, die Verteilungen in beiden Gruppen visuell zu beurteilen (Boxplot oben) oder über einen Quantil-Quantil-Plot (Abbildung 8.3).

```
qqnorm(int)
qqnorm(cont)
```

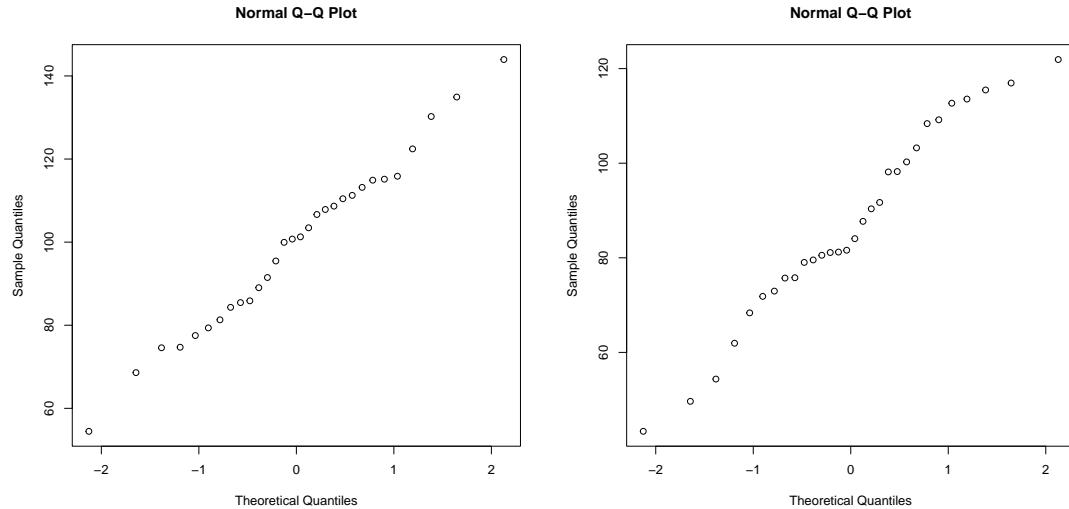


Abbildung 8.3: QQ-Plots für Daten in beiden Gruppen

Bei Ablehnung von homogenen Varianzen würden wir den Default-Wert `var.equal=FALSE` belassen. Den *t*-Test mit heterogenen Varianzen nennt man auch Welch-Test. Die Resultate sind in diesem Fall praktisch identisch.

```
t.test(d.rct$outcome ~ d.rct$group) #Test für heterogene Varianzen (Default: 'Welch'-Test)

##
## Welch Two Sample t-test
##
## data: d.rct$outcome by d.rct$group
## t = -2, df = 58, p-value = 0.02
## alternative hypothesis: true difference in means between group Control and group Intervention is not equal to 0
## 95 percent confidence interval:
## -23.12 -1.83
## sample estimates:
## mean in group Control mean in group Intervention
## 87.0 99.4
```

Robuste Verfahren*. *Verteilungsfreie, nicht-parametrische* Testverfahren machen keine Annahmen über die Wahrscheinlichkeitsverteilung der untersuchten Variablen und sind deswegen auch anwendbar, wenn die je nach Kontext Voraussetzungen an die Verteilungen nicht erfüllt sind, wie z.B. die Annahme einer Normalverteilung für einen *z*-Test oder einen *t*-Test, oder wenn die Daten nicht mindestens intervallskaliert sind. Eine robuste Alternative bei Verletzung der Annahmen für den Zwei-Stichproben

t-Test ist der Rangsummentest (auch Mann-Whitney Test genannt), implementiert mit `wilcox.test()`. Wir kommen auf nicht-parametrische Tests später zurück. Wenn vor allem *p*-Werte von Interesse sind, macht es oft Sinn, direkt nicht-parametrische Tests zu wählen.

```
wilcox.test(d.rct$outcome ~ d.rct$group)

##
## Wilcoxon rank sum exact test
##
## data: d.rct$outcome by d.rct$group
## W = 307, p-value = 0.03
## alternative hypothesis: true location shift is not equal to 0
```

Siehe <https://rstudio.zhaw.ch/rsconnect/content/83> für Simulation von kontinuierlichen Daten für zwei Stichproben und Analyse über einen *t*-Test für zwei unabhängige Gruppen. Dort kann man je n Beobachtungen simulieren aus zwei Gruppen. β_1 (Intercept) steht dort für μ_1 und β_2 für den Zwischengruppeneffekt $\mu_2 - \mu_1$. Das ist die Terminologie der linearen Modelle (siehe 13.1.5); wir kommen auf diese Notation später noch einmal zurück.

8.5 *t*-Test für gepaarte Daten

Beim *t*-Test für *gepaarte Daten* werden zwei Variablen miteinander verglichen, die nicht voneinander unabhängig sind. Wir betrachten zwei voneinander *abhängige* Variablen X_1 und X_2 , z.B. einen Score *vor* und *nach* einer Therapie in einer Kohortenstudie. Es ist klar, dass die *zweite Messung an demselben Patienten nicht unabhängig* ist von der ersten Messung, kurz $X_2 | X_1 \neq X_2$. Wir testen jetzt analog zu oben zweiseitig. Die Hypothesen lauten $H_0 : \mu_2 - \mu_1 = \delta_0$ und $H_1 : \mu_2 - \mu_1 \neq \delta_0$.

Dazu bilden wir die *Differenzen* $D_i = X_{i2} - X_{i1}, i = 1, \dots, n$ und machen mit dieser neuen Variable D , also mit der Variable der *individuellen Veränderungen*, den schon bekannten *t*-Test für eine Stichprobe. Die Teststatistik ist also

$$T = \frac{\bar{D} - \delta_0}{\text{se}(\bar{D})} = \frac{\bar{D} - \delta_0}{s_D / \sqrt{n}} \quad (8.5.1)$$

und der Test ist damit *äquivalent zu einem t-Test für eine Stichprobe*.

Implementation in R. Konsultiere zuerst den in R als data frame verfügbaren Datensatz `sleep`.

```
help(sleep)
```

Achtung: Die Variable `group` steht für Messwiederholung.

Long-Format versus wide-Format. Diese Daten liegen hier im sogenannten *long*-Format vor, d.h. wir haben pro statistische Einheit (ID) *zwei* Zeilen. $n = 10$ Probanden wurden zweimal gemessen, das data frame hat also 20 Zeilen.

```
sleep

##   extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

Ins *wide*-Format kann man diese Daten mit der `reshape()`-Funktion bringen. Im *wide*-Format hat man eine Zeile pro statistischer Einheit, also ein data frame mit 10 Zeilen.

```
sleep.wide <- reshape(sleep, v.names = "extra", timevar = "group", idvar = "ID", direction = "wide")
sleep.wide

##   ID extra.1 extra.2
## 1  1    0.7    1.9
## 2  2   -1.6    0.8
## 3  3   -0.2    1.1
## 4  4   -1.2    0.1
## 5  5   -0.1   -0.1
## 6  6    3.4    4.4
## 7  7    3.7    5.5
## 8  8    0.8    1.6
## 9  9    0.0    4.6
## 10 10   2.0    3.4
```

Beschreiben wir zuerst die beiden Messzeitpunkte.

```
by(sleep$extra, sleep$group, psych::describe)

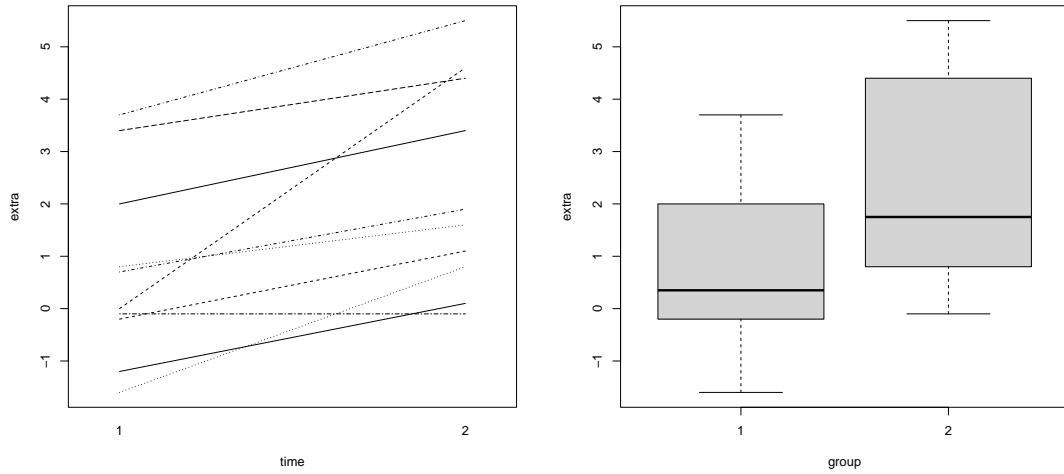
## sleep$group: 1
##   vars n mean sd median trimmed mad min max range skew kurtosis   se
## X1    1 10 0.75 1.79    0.35    0.68 1.56 -1.6 3.7    5.3 0.42    -1.3 0.57
## -----
## sleep$group: 2
##   vars n mean sd median trimmed mad min max range skew kurtosis   se
## X1    1 10 2.33 2   1.75    2.24 2.45 -0.1 5.5    5.6 0.28    -1.66 0.63
```

```
psych::describe(sleep.wide[, -1]) #Alternative

##           vars n mean   sd median trimmed mad min max range skew kurtosis    se
## extra.1     1 10 0.75 1.79   0.35    0.68 1.56 -1.6 3.7   5.3 0.42    -1.30 0.57
## extra.2     2 10 2.33 2.00   1.75    2.24 2.45 -0.1 5.5   5.6 0.28    -1.66 0.63
```

Um gepaarte Daten darzustellen, ist `interaction.plot()` sinnvoll, damit man die Paare als solche erkennt (*Spaghetti-Plot*). Boxplots sind hier weniger geeignet:

```
interaction.plot(x.factor = sleep$group, trace.factor = sleep$ID, response = sleep$extra, xlab = "time",
                  ylab = "extra", legend = FALSE)
boxplot(extra ~ group, sleep)
```



In R ist der Befehl für einen Test für gepaarte Daten (bei Nullhypothese $\delta_0 = 0$): `t.test(X, Y, paired=TRUE)`. Da wir gepaarte Pre-Post Daten haben, müssen wir das Argument `paired=TRUE` nicht vergessen. Da die Daten im *long*-Format vorliegen, müssen wir die Beobachtungen für beiden Messzeitpunkte über Indexierung `[]` extrahieren:

```
t.test(sleep$extra[sleep$group == 2], sleep$extra[sleep$group == 1], paired = TRUE) #mit long-
format

##
##  Paired t-test
##
##  data:  sleep$extra[sleep$group == 2] and sleep$extra[sleep$group == 1]
##  t = 4, df = 9, p-value = 0.003
##  alternative hypothesis: true mean difference is not equal to 0
##  95 percent confidence interval:
##  0.70 2.46
##  sample estimates:
##  mean difference
##                      1.58
```

Wir könnten auch direkt mit dem data.frame im wide-Format arbeiten:

```
t.test(x = sleep.wide$extra.2, y = sleep.wide$extra.1, paired = TRUE) #mit wide-
format, äquivalent
```

Wir können auch einen Einstichproben-*t*-Test der Veränderungen machen. Dazu berechnen wir zuerst die Variable `change` und machen damit den Einstichproben-*t*-Test.

```
change <- sleep$extra[sleep$group == 2] - sleep$extra[sleep$group == 1]
summary(change)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu. Max.
##   0.00   1.05   1.30   1.58   1.70   4.60
```

```
t.test(change)

##
## One Sample t-test
##
## data: change
## t = 4, df = 9, p-value = 0.003
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.70 2.46
## sample estimates:
## mean of x
##             1.58
```

Robuste Verfahren*. Auch für dein Einstichproben *t*-Test gibt es eine robuste, nicht-parametrische Variante, den Wilcoxon-Vorzeichen-Rang Test, den wir später einführen:

```
wilcox.test(change)

## Warning in wilcox.test.default(change): cannot compute exact p-value with ties
## Warning in wilcox.test.default(change): cannot compute exact p-value with zeroes

##
## Wilcoxon signed rank test with continuity correction
##
## data: change
## V = 45, p-value = 0.009
## alternative hypothesis: true location is not equal to 0
```

`wilcox.test()` kann hier keine exakten *p*-Werte berechnen, weil es Bindungen (gleiche Ränge) gibt. Es gibt dafür eine alternative Funktion im Packet `exactRankTests`

```
## install.packages('exactRankTests')
exactRankTests::wilcox.exact(change)

##
##  Exact Wilcoxon signed rank test
##
## data: change
## V = 45, p-value = 0.004
## alternative hypothesis: true mu is not equal to 0
```

Statistische Signifikanz und Relevanz. Über eine Verkleinerung des Standardfehlers – indem man n vergrössert – kann man jeden noch so kleinen Unterschied statistisch “signifikant” machen. Auch irrelevante Unterschiede sind häufig “statistisch signifikant”. Schauen wir uns hierzu die T -Statistik (8.3.2) nochmal genau an. Je grösser n , umso grösser wird die Statistik, und man kann immer erreichen, dass man $\mu = \mu_0$ zugunsten der zweiseitigen – unspezifischen – Alternative verwirft, egal, wie klein – oder eben unbedeutend – der Zähler ($\bar{X} - \mu_0$) ist. Statistische Signifikanz hat also zuerst einmal wenig zu tun mit klinischer Relevanz, sondern mehr mit Präzision.

8.6 Teststärke

Wir haben bisher nur *unspezifische Alternativhypotesen* betrachtet, die als *Komplement* zur Nullhypothese formuliert wurden. Das Komplement $\mu \neq \mu_0$ ist aber eine *zusammengesetzte* Hypothese und besteht aus vielen einzelnen Punkthypothesen.

Um auch den oben eingeführten β -Fehler zu kontrollieren (Tabelle 8.1), muss die Alternative μ_A , und damit der Abstand zur Nullhypothese, $\mu_A - \mu_0$, *spezifiziert* werden, und zwar auf den Wert, den der Test “entdecken” oder “signalisieren” soll.

Es gibt nun verschiedene Interpretationen zu dieser Grösse:

- eine (minimal) klinisch relevante Differenz
- eine erstrebenswerte Differenz
- eine realistische Einschätzung der Differenz

Diese Ideen vermischen z.T. die Anforderungen und Erwartungen bezüglich der Alternative, daher wird die spezifizierte Alternative oft dargestellt als “realistisch *und* relevant”. Das führt zwar zu philosophischen Schwierigkeiten, auf die wir hier nicht näher eingehen wollen³.

Die *Teststärke* oder *Power* des statistischen Tests ist nun die Wahrscheinlichkeit, dass der Test gegen H_0 entscheidet, *wenn* die spezifische Alternative μ_A wahr ist; diese

³In der Bayesianischen Statistik sind diese beiden Konzepte klar getrennt.

Wahrscheinlichkeit ist also eine Funktion von μ_A ,

$$\text{Power}(\mu_A) = 1 - \beta = \Pr(H_0 \text{ ablehnen} \mid \mu_A). \quad (8.6.1)$$

Abbildung 8.4 zeigt die Teststärke als Funktion von μ für verschiedene n und für verschiedene Standardabweichungen σ , für fixiertes $\alpha = 0.05$ und für einen einseitigen z -Test von $H_0 : \mu \leq 10$ gegen die Alternative $H_1 : \mu > 10$. Die Teststärke steigt mit Stichprobenumfang n und mit zunehmenden μ_A , sie steigt auch mit abnehmender Streuung der Daten.

Bei einer Stichprobe von $n = 20$, einer Standardabweichung der Daten von $\sigma = 1$ sowie einem angenommenen $\mu = 10.5$ ist die Wahrscheinlichkeit, die Nullhypothese $H_0 : \mu \leq 10$ auf 5%-Niveau zu verwerfen, 0.72 (roter Punkt).

Power einseitiger Test*. Die Wahrscheinlichkeit der Ablehnung von $H_0 : \mu \leq \mu_0$ auf α -Niveau, gegeben ein $\mu \geq \mu_0$, ist eine Funktion von μ, μ_0, σ und n :

$$F_\mu(\alpha) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha}\right), \quad (8.6.2)$$

mit $\Phi()$ als der kumulativen Dichte der Standardnormalverteilung. In R können wir dies mit den bekannten `pnorm()` und `qnorm()` “von Hand” berechnen:

```
pnorm((10.5 - 10)/(1/sqrt(20)) - qnorm(0.95))
## [1] 0.723
```

Die Teststärke wurde zusätzlich auch in Abbildung 8.5 dargestellt. Wir sehen auch, dass die Wahrscheinlichkeit, H_0 zu verwerfen, wenn H_0 wahr ist, gerade $\alpha = 0.05$ ist, α ist also gerade die Power unter $\mu_A = \mu_0 = 10$. In der Abbildung erkennt man, dass

$$\mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = \mu - z_{1-\beta} \frac{\sigma}{\sqrt{n}}. \quad (8.6.3)$$

Auflösen nach $1 - \beta$ ergibt dann die Power-Funktion (8.6.2) (Übung).

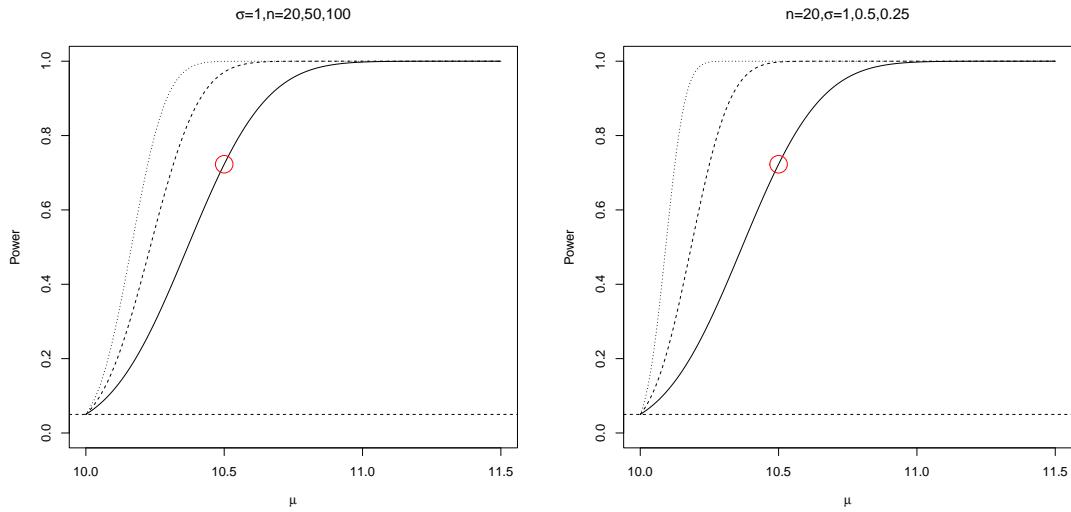


Abbildung 8.4: Power-Funktion eines einseitigen z -Tests, mit $H_0 : \mu \leq 10$. Links: $n = 20(—), 50(--)$, $100(\cdots)$ und $\sigma = 1$. Rechts: $n = 20$ und $\sigma = 1(—), 0.5(--)$, $0.25(\cdots)$.

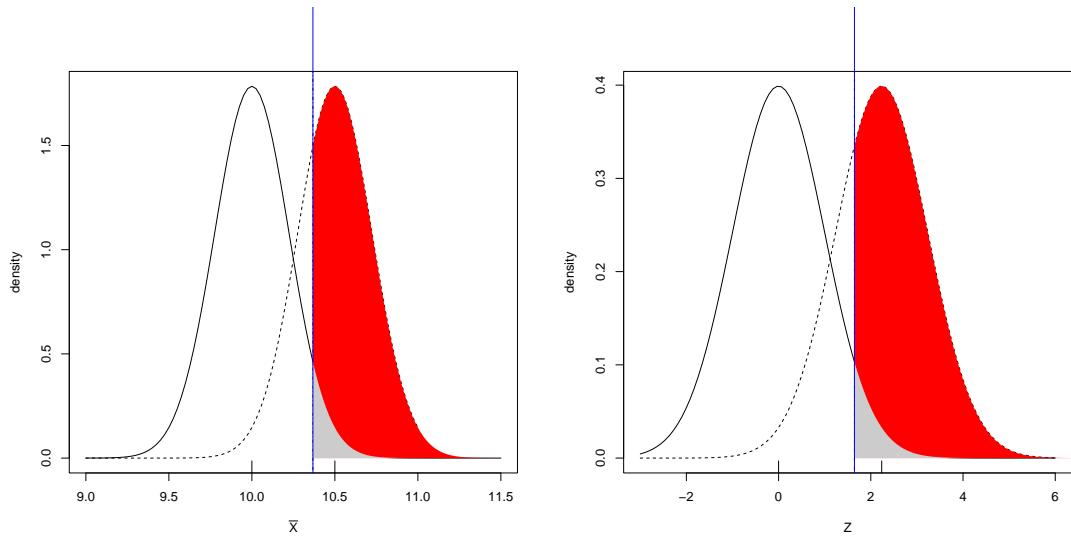


Abbildung 8.5: Links: Verteilung von \bar{X} unter $H_0 : \mu \leq 10$ und unter spezifizierter Alternative $\mu = 10.5$. α (hell) und Power (rot). Die blaue Linie markiert $\mu_0 + z_{0.95} \times \text{se}(\bar{X})$. Rechts: Standardisierte Version.

Power in R. In R ist die Berechnung der Power oder Teststärke implementiert über die t -Verteilung (statt mit der z -Verteilung) mit `power.t.test()`. Wenn wir die

Teststärke für obige geplante Testprozedur wollen, müssen wir als Argumente `n`, `delta`, `sd`, `type` und `alternative` spezifizieren. Die Power wird dann berechnet. Sie ist hier leicht verschieden als unter der Standardnormalverteilung. Abbildung 8.6 zeigt die Verteilung der Teststatistik $T = \frac{\bar{X} - 10}{s/\sqrt{n}}$ unter verschiedenen Alternativen. In R wird dafür die *nicht-zentrale t*-Verteilung gebraucht.

```
power.t.test(n = 20, delta = 0.5, sd = 1, type = "one.sample", alternative = "one.sided")

##
##      One-sample t test power calculation
##
##      n = 20
##      delta = 0.5
##      sd = 1
##      sig.level = 0.05
##      power = 0.695
##      alternative = one.sided
```

Nicht-zentrale t-Verteilung*. Nicht-zentrale *t*-Verteilungen haben noch einen zusätzlichen Parameter, den *non-centrality*-Parameter (`ncp`). Was stellt diese Grösse dar? Die Statistik $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ kann man schreiben als

$$T = \frac{\bar{X} - \mu + \mu - \mu_0}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} + \frac{\mu - \mu_0}{s/\sqrt{n}}.$$

Diese Statistik ist dann

- unter $\mu = \mu_1$ nicht-zentral *t*-verteilt mit $n - 1$ Freiheitsgraden und $ncp = \frac{\mu_1 - \mu_0}{s/\sqrt{n}}$
- unter $\mu = \mu_0$ (zentral) *t*-verteilt mit $n - 1$ Freiheitsgraden).

Obige Funktion (`power.t.test()`) berechnet dann die Wahrscheinlichkeit, dass T grösser ist als der kritische Wert, wenn $\mu = 10.5$ wahr ist, also folgende Grösse:

```
1 - pt(qt(0.95, 19), 19, ncp = 0.5/(1/sqrt(20)))

## [1] 0.695
```

Wie gross ist die Teststärke unter $H_0 : \mu = \mu_0$? Wir haben oben bereits gesehen, dass das gerade α ist.

```
1 - pt(qt(0.95, 19), 19, ncp = 0/(1/sqrt(20)))
power.t.test(n = 20, delta = 0, sd = 1, type = "one.sample", alternative = "one.sided")
```

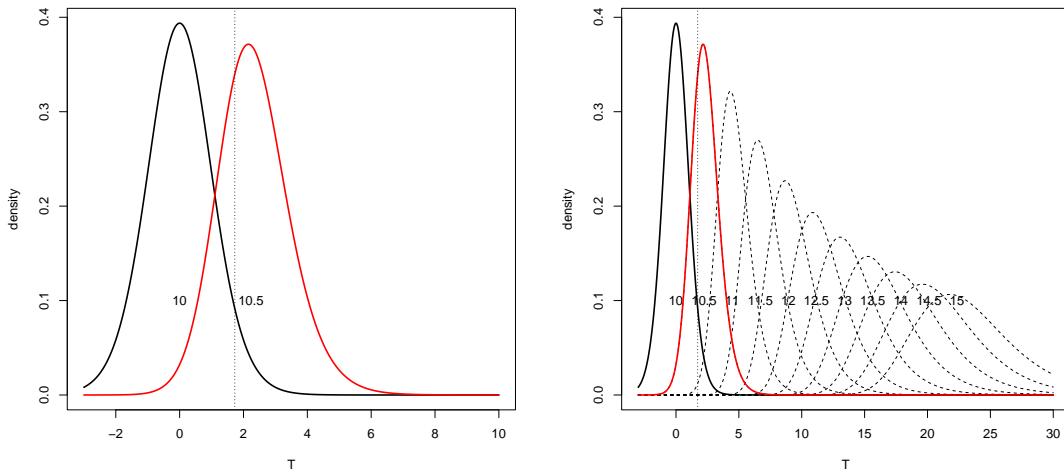


Abbildung 8.6: Dichte der Statistik $T = \frac{\bar{X} - 10}{s/\sqrt{n}}$ unter $\mu = 10$, unter $\mu = 10.5$ (links) und anderen (rechts).

Fallzahlberechnung. Bei vorgegebenem α , β und μ (und Standardabweichung σ) kann man die optimale Stichprobengröße n berechnen, die man benötigt, damit man H_0 mit einer Power von $1 - \beta$ auf α -Niveau verwerfen kann, *gegeben* die spezifische Alternative. Man nennt diese Prozedur *Fallzahlberechnung*.

Im Sinne der Bescheidenheit wird der α -Fehler meistens als “schlimmer”, als – mit mehr Kosten behaftet – betrachtet als ein β -Fehler. Viele Studien berechnen n daher *a priori* aufgrund einer Teststärke von $1 - \beta = 0.80$, also mit einem $\beta = 0.2$ und einem $\alpha = 0.05$, also mit einem Verhältnis von α zu β -Fehler von $1 : 4$.

Die Prozedur ist also folgende: Nehmen wir an, wir wollen z.B. $H_0 : \mu = \mu_0$ zugunsten von $H_1 : \mu \neq \mu_0$ mit einer gewissen Teststärke verwerfen:

1. Die Teststärke ist eine *Funktion* der Alternative. Damit wir also überhaupt von *einer* Teststärke sprechen können, muss anhand von inhaltlichen Kriterien *a priori* eine relevante und realistische Alternative (siehe oben) sowie eine Standardabweichung σ angegeben werden. Oft werden diese beiden Parameter als *Effektgröße*

$$d = \frac{\mu - \mu_0}{\sigma} \quad (8.6.4)$$

zusammengefasst. Diese Größe wird auch *Cohen's d* genannt. Häufig gelten dann $d = 0.8$ als gross, $d = 0.5$ als mittel und $d = 0.2$ als kleiner Effekt.

2. Es wird *a priori* α und β (und damit $Power = 1 - \beta$) bestimmt, sehr oft wählt man $\alpha = 0.05$ und $\beta = 0.2$. Sie sind aber – logisch gesehen – frei wählbar, je nach

Kosten, die man bereit ist zu akzeptieren bezüglich beiden Fehlerarten.

3. Der *optimale Stichprobenumfang* n wird berechnet: Oben war Power eine Funktion von $n, \alpha, \sigma, \mu - \mu_0$. Jetzt ist n eine Funktion von α, β und der Effektstärke, also

$$n = f(\alpha, \beta, \sigma, \mu - \mu_0). \quad (8.6.5)$$

Die Funktion $n = f(\alpha, \beta, \sigma, \mu - \mu_0)$ ist wieder implementiert mit `power.t.test()`. Wenn wir die Fallzahl wollen, müssen wir aber jetzt als Argumente `power`, `delta`, `sd`, `type` und `alternative` übergeben. Wollen wir z.B. eine Wahrscheinlichkeit von 0.8, die Nullhypothese $H_0 : \mu \leq 10$ auf 5%-Niveau zu verwerfen, gegeben eine Effektgrösse von $\frac{\mu - \mu_0}{\sigma} = \frac{0.5}{1}$, dann gibt folgender Code die nötige Fallzahl.

```
power.t.test(delta = 0.5, sd = 1, power = 0.8, type = "one.sample", alternative = "one.sided")

##
##      One-sample t test power calculation
##
##      n = 26.1
##      delta = 0.5
##      sd = 1
##      sig.level = 0.05
##      power = 0.8
##      alternative = one.sided
```

Die Stichprobengrösse n wird also *a priori* bestimmt, dies entspricht einer *ethischen Grundhaltung*: Wie viele Personen brauche ich unter der Annahme einer realistischen Effektgrösse, um das Studienziel – die Verwerfung der Nullhypothese, mit einer hohen Wahrscheinlichkeit zu erreichen.

8.7 Das Testen vom Strohmann*

Es soll hier nochmals betont werden, dass die Nullhypothese H_0 , die man mit einer gewissen Power versucht zu verwerfen, nicht die Hypothese vom Nulleffekt sein muss. $H_0 : \mu = 0$ zu verwerfen (oder z.B. $H_0 : RR = 1$ bei einem relativen Risiko oder Chancenverhältnis) ist als Ziel nicht immer nobel und manchmal ethisch unschön. Wir sollten diejenige H_0 zu verwerfen versuchen, deren logisches Komplement (Gegenteil) etwas klinisch Relevantes darstellt, z.B. Nullhypotesen wie $\mu < \mu_0$ mit μ_0 als die Grenze zwischen *irrelevant* und *relevant*. Wir sollten also nicht nur – wie man das leider oft sieht – gegen “Strohmann-Nullhypotesen” arbeiten. Ein sehr lesenswerter und auch immer noch sehr aktueller Artikel hierzu ist [29].

Mit dem folgenden Beispiel soll ein zentrales Problem der Wissenschaftsphilosophie dargestellt werden, das – aus meiner und vieler Sicht – ungenügend Beachtung findet. Dabei ist es wichtig zu betonen, dass es hier um Aspekte der Wissenschaftstheorie oder -logik geht und *nicht* um Wissenschaftssoziologie. Das Beispiel ist aus der Physik.

Die Suche nach dem Äther. Gegen Ende des Ende 19. Jahrhundert gab es eine *Hypothese* in der Physik: Der *Äther* existiert als *Medium*, in dem sich Lichtwellen (analog zu Schallwellen in der Luft) ausbreiten. Die Lichtgeschwindigkeit ist daher *nicht unabhängig* von der Richtung, mit der die Erde sich im Äther bewegt. Der hypothetisierte Wert für diesen *Ätherwind* ist ungefähr 30 km/s, gleich der Orbitalgeschwindigkeit der Erde.

Der *Interferometer* wurde von Michelson erfunden, um die Lichtgeschwindigkeit zu messen. Dieser wurde von Michelson und Morley 1881 verfeinert, um die Existenz vom Äther (Name von Aristoteles, dt. [blauer] Himmel) als Medium für das Licht aufzuzeigen. Die Lichtgeschwindigkeit wurde in verschiedenen Richtungen relativ zum Äther gemessen.

Abbildung 8.7 zeigt die Daten des Experiments (hier stark vereinfacht dargestellt), als der Verteilung der Lichtgeschwindigkeiten in verschiedene Richtungen relativ zum postulierten Äther.

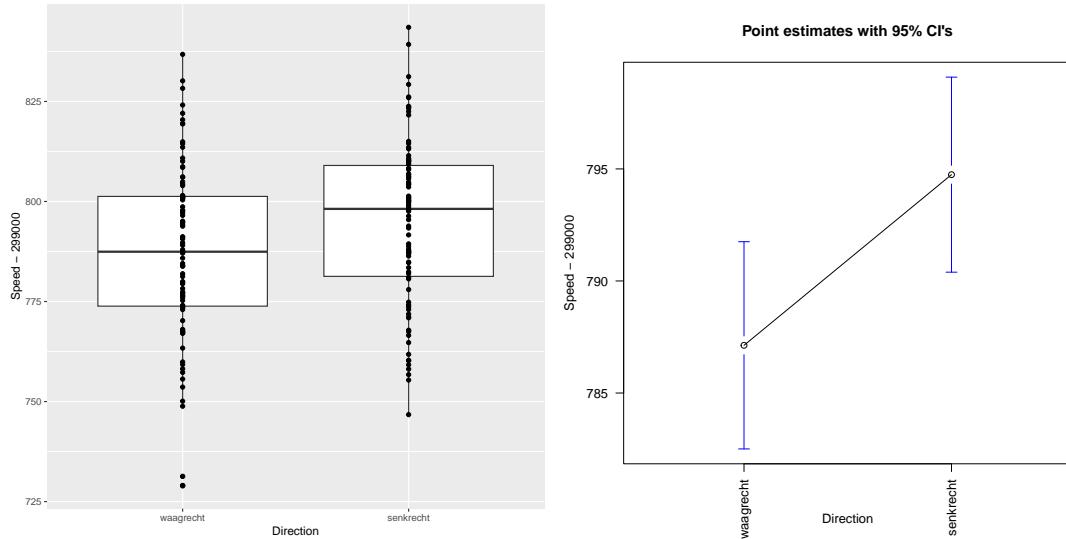


Abbildung 8.7: Lichtgeschwindigkeit (-299'000 km/s) versus Richtung

Der Kampf gegen den Strohmann. Wie wäre die Geschichte wohl gelaufen, wären die Physiker damals ähnlich *geschult* oder *sozialisiert* worden wie wir in den Gesundheitswissenschaften (aber auch in der Soziologie, Psychologie, usw.)?

Sie hätten wahrscheinlich – wie viele in unserem Umfeld – zuerst auch die *Strohmann-Hypothese*, die *Null-Nullhypothese* getestet. Hier ein typischer Strohmann-Test, wie er unzählige Male jeden Tag auf dieser Welt gemacht wird: Der *p*-Wert für den Test, dass

die Lichtgeschwindigkeit nicht von der Richtung abhängt, also vom Test von

$$H_0 : \delta = 0$$

ist $p = 0.011$. Das Ergebnis wäre also ein sogenannter *signifikanter Effekt* der Richtung auf die Lichtgeschwindigkeit gewesen. Die Physiker hätten dann vielleicht kommuniziert: Es gibt *Evidenz für den Äther* als Medium für das Licht. So jedenfalls hören wir das oft.

Klare Beschreibung der Daten. Zum Glück kannten die Physiker aber noch keine *Signifikanztests*; sie haben vor allem die Daten *deskriptiv* genau beschrieben (was wir hier jetzt nicht machen). Das Hauptresultat war: Die beobachtete durchschnittliche Differenz in der Lichtgeschwindigkeit (Ätherwind) war 7.615 km/s.

It seems fair to conclude....the ether is probably less than one sixth the earths orbital velocity, and certainly less than one fourth (Michelson and Morley, 1887a, p. 333).

Kämpfe nicht gegen den Strohmann. Für kritische Rationalisten wie KARL POPPER sollten wir *unsere* Hypothesen testen, nicht den Strohmann. Vorläufig wahr ist, was auch bei *bester Anstrengung nicht zur Falsifikation führt*⁴.

In unseren Wissenschaften führt grössere Präzision⁵ (z.B. mehr n , mehr Information, weniger Lärm in den Daten, mehr Genauigkeit in der Durchführung des Experiments) zu *mehr Risiko* für die Null-Nullhypothese, das heisst, *Evidenz für eine Theorie* hängt nur von der Genauigkeit der Durchführung ab, ein doch sehr unschönes Unterfangen.

Viele Physiker sind kritische Rationalisten. Diese zeichnen sich häufig dadurch aus, dass sie *Lust* und *Spass* daran haben, *ihre* Hypothesen zu testen, gegen *ihre* Theorien zu kämpfen⁶. Sie wollen, dass sich ihre Theorien in *strengen Tests bewähren*. Sie hätten daher damals eher die Forschungshypothese anstatt *Strohmann*-Nullhypothese getestet.

⁴Die deduktive Logik ist die Theorie der Übertragung der Wahrheit von Prämisse auf die Konklusion, aber auch die Übertragung der Falschheit von der Konklusion auf die Prämisse. Sei T ein Theorensystem und B eine Beobachtung, dann gilt nach dem Modus tollens: $(\neg B \cap (T \rightarrow B)) \rightarrow \neg T$. Das Problem der Verifikation: Modus ponens: $(T \cap (T \rightarrow B)) \rightarrow B$. Will man nun folgende Beobachtung B erklären: Peters Lendenwirbelsäule ist steif. Das Theorensystem T: T1: Alle Patienten haben eine steife Lendenwirbelsäule und T2: Peter ist ein Patient wird durch die Beobachtung Peters Lendenwirbelsäule ist steif verifiziert, die Beobachtung ist aus den Prämissen logisch gültig ableitbar, diese Prämisse – T1 – ist aber falsch. Aus einem logisch gültigen Schluss und einer Beobachtung folgt nicht die Richtigkeit der Prämisse, der Theorie oder der Hypothese.

⁵Im Gegensatz aber folgt aus der Falschheit der Konklusion die Falschheit der Prämissen.

⁶Für Anhänger von NEYMAN-PEARSON Hypothesen-Tests kann die *Präzision* von RONALD FISHER mit *Power* übersetzt werden.

⁶Kritischen Rationalisten wird oft vorgeworfen, dass sie nur zerstören wollen. Natürlich ist das Umgekehrte der Fall: Sie wollen wissenschaftliche Behauptungen auf den Prüfstand stellen, um deren Bewährung aufzuzeigen.

Wenn wir *unsere* Hypothesen testen, dann riskieren wir mit mehr Präzision, Evidenz *gegen* sie zu erhalten. Die Rolle von (statistischen) Tests ist also in unseren Wissenschaften sehr oft gerade *umgekehrt* als in den Naturwissenschaften. Unser Vorgehen kann häufig logisch nicht begründet werden. Wir haben oft nur *schwache Evidenz für unsere Hypothesen*, weil wir *schwache Tests (Weak Tests)* machen.

Strenge Tests. Jetzt der Test der damaligen Forschungshypothese, dass der Ätherwind (und damit der Unterschied in der Lichtgeschwindigkeit) 30 km/h beträgt, der *Strong Test*: Der p -Wert für die Hypothese

$$H_0 : \delta = 30,$$

ist $p = 6.004 \times 10^{-14}$. Dieser p -Wert ist sehr, sehr klein. Die Daten sind unter dieser Hypothese ($\delta = 30$) viel unwahrscheinlicher als unter der Hypothese, dass der Äther nicht existiert ($\delta = 0$). Die entsprechende *Likelihood-Ratio*⁷ ist.

$$\frac{\Pr(\text{data} \mid \delta = 30)}{\Pr(\text{data} \mid \delta = 0)} = \frac{\Pr(7.615 \mid \delta = 30)}{\Pr(7.615 \mid \delta = 0)} = 1.5 \times 10^{-11}. \quad (8.7.1)$$

Die Likelihood von $\delta = 30$ (Ätherwind von 30 km/h) ist also um den Faktor 6.665×10^{10} kleiner als die Likelihood von $\delta = 0$, also 66.649 Milliarden mal kleiner! Es gibt praktisch keine Evidenz für die Hypothese vom Äther. Abbildung 8.8 zeigt die relative Likelihood für verschiedene Hypothesen, gegeben die beobachteten Daten.

⁷Die Likelihood-Ratio ist das Verhältnis der Likelihoods von beiden Hypothesen. Die Likelihood einer Hypothese θ ist die Wahrscheinlichkeit der beobachteten Daten x , gegeben θ , als Funktion von θ : $L(\theta \mid x) = \Pr(x \mid \theta)$.

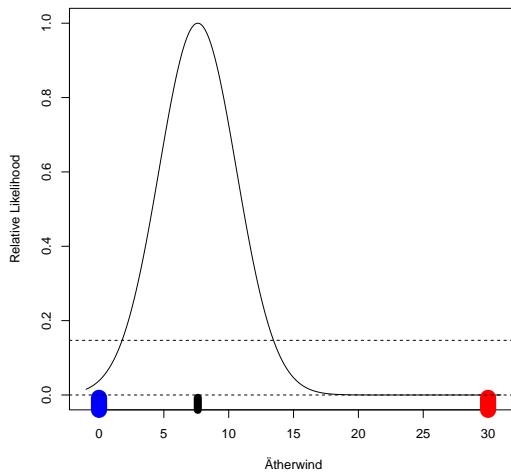


Abbildung 8.8: Relative Likelihood, gegeben die Daten. Schwarz: Der beobachtete Wert ist $= 7.62$ und hat maximale relative Likelihood. Der Schwellenwert von 0.145 für die relative Likelihood entspricht einem approximativen 95% -CI von $(1.77, 13.46)$. Die obere Grenze des CI ist klar weiter weg von $\delta = 30$ als die untere Grenze von $\delta = 0$.

Relevanz. Das Michelson-Morley-Experiment gilt bis heute als das *berühmteste Experiment mit negativem Ergebnis*.

Die Nicht-Existenz des Äthers führte in Einsteins *Wunderjahr* 1905 zum Durchbruch der speziellen Relativitätstheorie, die das Resultat des Michelson-Morley-Experiments erklären konnte. Die Lichtgeschwindigkeit c ist *unabhängig vom Bezugssystem* und eine Naturkonstante. Eine Konsequenz daraus ist, dass Längen und Zeitdauern vom Bewegungszustand des Betrachters abhängen und es keinen absoluten Raum und keine absolute Zeit gibt. Eine weitere Konsequenz ist die Äquivalenz von Masse und Energie, $E = mc^2$.

Konsequenzen. Auch in unseren Wissenschaften könnten wir – streng logisch betrachtet⁸ – mehr Fortschritte machen, wenn wir weniger gegen den Strohmann kämpfen.

⁸Natürlich gibt es neben der Logik viele andere Aspekte von Wissenschaft, wie der Wissenschaftsbetrieb und die Wissenschaftssoziologie.

8.8 Äquivalenztests*

Oft ist es nicht sinnvoll, Nullhypotesen wie $H_0 : \mu_1 - \mu_2 = \delta_0$ gegen $H_1 : \mu_1 - \mu_2 \neq \delta_0$ zu testen (Oft ist zudem $\delta_0 = 0$) und wir können mit einem Test “nur” etwas verwerfen, was rein logisch (a priori!) schon falsch ist, denn das wahre δ wird nie genau δ_0 sein (ausser bei perfekt randomisierten Experimenten, dort gehen wir z.B. davon aus, dass der wahre Zwischengruppenunterschied Null ist. Bei Beobachtungsstudien ist das aber nicht der Fall). Mit grosser Power, sprich grossem n ist es immer möglich, solche “Strohmann-Nullhypotesen” zu verwerfen.

Wenn eine Theorie aber nun einen Bereich für $\delta = \mu_1 - \mu_2$ vorhersagt, d.h. dass δ in einem Bereich liegt, der durch Grenzen $[-\epsilon, +\epsilon]$ festgelegt ist, haben wir folgende – viel stärkere – Testsituation:

$$H_0 : \delta \leq -\epsilon \text{ oder } \delta \geq +\epsilon \quad H_1 : -\epsilon < \delta < \epsilon. \quad (8.8.1)$$

- Die Nullhypothese besagt, dass der wahre Parameter ausserhalb einer *Toleranzregion* liegt oder in der Region der *Irrelevanz*.
- Die Alternative besagt, dass der wahre Parameter im durch die Theorie vorhergesagten Region liegt, im Bereich von *Relevanz*.

H_0 zu verwerfen heisst jetzt Irrelevanz zu verwerfen (definiert durch die Grenzen ϵ). Dieses Problem kann man lösen mit sogenannten *TOST*-Testverfahren (“Two One-sided *t*-Tests”). Man macht zwei einseitige Tests auf α -Niveau:

$$H_{0a} : \delta \leq -\epsilon \quad H_{1a} : -\epsilon < \delta \quad (8.8.2)$$

$$H_{0b} : \delta \geq +\epsilon \quad H_{1b} : \delta < +\epsilon. \quad (8.8.3)$$

Ablehnen von H_{0a} **und** H_{0b} bedeutet dann, $-\epsilon < \delta < \epsilon$. Die einzelnen Tests sind normale *t*-Tests, alle Varianten möglich (Unverbundene Stichproben (gleiche oder ungleiche Varianz), Verbundene Stichprobe, Eine Stichprobe).

Multiples Testen. Bei TOST-Testverfahren müssen wir keine *Korrektur für multiples Testen* durchführen, beide Tests werden zum Niveau α gemacht. Das hat zur Konsequenz, dass wir bei $\alpha = 0.05$ den Test auch mit einem 90% KI (statt einem 95% KI) (mit L, U also untere und obere Grenze) durchführen können.

- H_{0a} wird genau dann abgelehnt, wenn das einseitige $(1 - \alpha)$ -KI $[L, \infty)$ komplett rechts von $-\epsilon$ liegt
- H_{0b} wird genau dann abgelehnt, wenn das einseitige $(1 - \alpha)$ -KI $(-\infty, U)$ komplett links von $+\epsilon$ liegt
- Die Schnittmenge $[L, U] = [L, \infty) \cap (-\infty, U]$ ist aber genau das $(1 - 2\alpha)$ -KI für

die Differenz der Erwartungswerte.

Dualität von Testen und Schätzen. Wir können den TOST zum Niveau α also auf zwei Arten durchführen:

- Zwei einseitige t-Tests, H_0 ablehnen, wenn beide p -Werte kleiner als α
- $(1 - 2\alpha)$ -Konfidenzintervall für die Mittelwertdifferenz berechnen. Wir schliessen auf Äquivalenz, wenn das KI komplett in $(-\epsilon, +\epsilon)$ enthalten ist.

Implementation in R. diese sind im Packet TOSTER enthalten. (Siehe auch <https://cran.rstudio.com/web/packages/TOSTER/vignettes/IntroductionToTOSTER.html>.)

```
## install.packages('TOSTER') #auskommentieren zum Installieren
library(TOSTER)
```

Für den Einstichprobenfall braucht man `TOSTOne.raw()`, für den Zweistichprobenfall `TOSTtwo.raw()`.

Beispiel. Anbei ein Beispiel aus der Hilfefunktion: Eskine (2013) showed that participants who had been exposed to organic food were substantially harsher in their moral judgments relative to those exposed to control ($n_1 = n_2 = 21$, $d = 0.81$, 95% CI: [0.19, 1.45])⁹

A replication by Moery & Calin-Jageman (2016, Study 2) did not observe a significant effect (Control: $n = 95$, $M = 5.25$, $SD = 0.91$, Organic Food: $n = 89$, $M = 5.22$, $SD = 0.83$). Following Simonsohns (2015) recommendation the equivalence bound was set to the effect size the original study had 33% power to detect. With $n = 21$ in each condition, this means the equivalence bound is $d = 0.48$, which equals a difference of 0.419 on a 7-point scale given the sample sizes and a pooled standard deviation of 0.871.

```
## determine equivalence bounds (effect size) following Simonsohn's(2005) with n1=n2=21
(d <- power.t.test(n = 21, sd = 1, power = 0.33)$delta)

## [1] 0.481

## pooled standard deviation
spooled <- sqrt((0.91^2 * 20 + 0.83^2 * 20)/(21 + 21 - 2))
spooled

## [1] 0.871

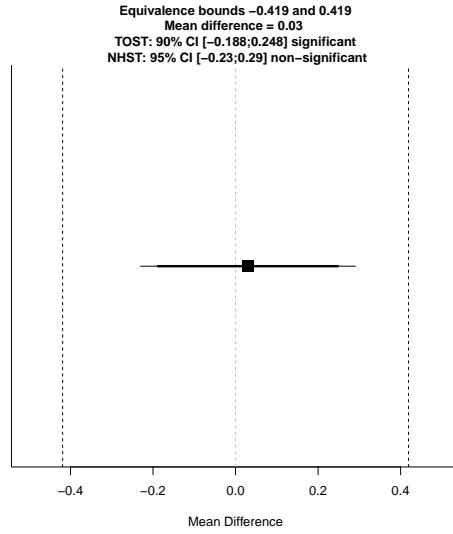
## delta=ES x SD
d * spooled

## [1] 0.419
```

⁹ d steht für die Effektgrösse nach Cohen, $d = \frac{\hat{\delta}}{s}$

Using a TOST equivalence test with $\alpha = 0.05$, assuming equal variances, and equivalence bounds of $d = -0.48$ and $d = 0.48$ is significant, $t(182) = -3.03$, $p = 0.001$. We can reject effects more extreme than $d = 0.48$ (or a raw difference of 0.419 scalepoints).

```
TOSTtwo.raw(m1 = 5.25, m2 = 5.22, sd1 = 0.95, sd2 = 0.83, n1 = 95, n2 = 89, low_eqbound = -0.419, high_eqbound = 0.419,
alpha = 0.05, var.equal = TRUE)
```



```
## TOST results:
## t-value lower bound: 3.40 p-value lower bound: 0.0004
## t-value upper bound: -2.95 p-value upper bound: 0.002
## degrees of freedom : 182
##
## Equivalence bounds (raw scores):
## low eqbound: -0.419
## high eqbound: 0.419
##
## TOST confidence interval:
## lower bound 90% CI: -0.188
## upper bound 90% CI: 0.248
##
## NHST confidence interval:
## lower bound 95% CI: -0.23
## upper bound 95% CI: 0.29
##
## Equivalence Test Result:
## The equivalence test was significant, t(182) = -
2.950, p = 0.0018, given equivalence bounds of -
0.419 and 0.419 (on a raw scale) and an alpha of 0.05.
##
## Null Hypothesis Test Result:
## The null hypothesis test was non-
significant, t(182) = 0.227, p = 0.820, given an alpha of 0.05.
```

Kapitel 9

Verteilungsfreie Testverfahren

Oft wird die Modellstruktur – damit meint man die Parameter eines Modells – nicht a priori festgelegt wie bei parametrischen Modellen, sondern aus den Daten bestimmt. Die Art und Anzahl der Parameter ist dann flexibel und wird nicht von vornherein festgelegt.

Verteilungsfreie, nicht-parametrische Testverfahren machen keine Annahmen über die Wahrscheinlichkeitsverteilung der untersuchten Variablen und sind deswegen auch anwendbar, wenn die Voraussetzungen an die Verteilungen nicht erfüllt sind, wie z.B. die Annahme einer Normalverteilung für einen z -Test oder einen t -Test, oder wenn die Daten nicht mindestens intervallskaliert sind.

In nicht-parametrischen Modellen stehen also nicht Parameter im Vordergrund, sondern generelle Aspekte der Verteilung wie *Median* oder ganz allgemein *Quantile*. Verteilungsfrei bedeutet, dass die Verteilung der Teststatistik unter H_0 *nicht* von der Verteilung der Daten abhängt. Die Statistik selber hat natürlich schon eine Verteilung.

Analog zum t -Test für unabhängige und abhängige Stichproben gibt es den *Mann-Whitney U-Test* (oder *Rangsummen-Test*) für unabhängige Stichproben und den *Wilcoxon-Test* (oder *Vorzeichen-Rang-Test*) für eine Stichprobe oder abhängige Stichproben.

9.1 Wilcoxon-Vorzeichen-Rang-Test

Der *Wilcoxon-Vorzeichen-Rang-Test* ist das nicht-parametrische Analogon zum t -Test für eine Stichprobe oder für gepaarte Stichproben. Sind $D_i = X_{i2} - X_{i1}$ i.i.d. (unabhängig und identisch verteilt), dann können wir mit dem Wilcoxon-Test folgende Hypothesen bezüglich den Medianen \tilde{x}_1 und \tilde{x}_2 testen:

$$\begin{aligned} H_0 : \tilde{x}_2 - \tilde{x}_1 &= \delta_0, & H_1 : \tilde{x}_2 - \tilde{x}_1 &\neq \delta_0 \\ H_0 : \tilde{x}_2 - \tilde{x}_1 &\leq \delta_0, & H_1 : \tilde{x}_2 - \tilde{x}_1 &> \delta_0 \\ H_0 : \tilde{x}_2 - \tilde{x}_1 &\geq \delta_0, & H_1 : \tilde{x}_2 - \tilde{x}_1 &< \delta_0 \end{aligned}$$

Wir konstruieren dann die Teststatistik folgendermassen:

- Wir berechnen die Differenzen $D_i = (X_{i2} - X_{i1}) - \delta_0$, $i = 1, \dots, n$.
- Wir berechnen den Rang R_i der absoluten Differenzen $R_i = \text{rang}(|D_i|)$
- Die Teststatistik W ist dann das Minimum der negativen und der positiven Rangsummen,

$$W^+ = \sum_{i=1}^n I_{D_i > 0} R_i \quad (9.1.1)$$

$$W^- = \sum_{i=1}^n I_{D_i < 0} R_i \quad (9.1.2)$$

$$W = \min(W^+, W^-) \quad (9.1.3)$$

mit I als der Indikatorfunktion.

- Man kann zeigen, dass $E(W) = \frac{1}{4}n(n + 1)$ und $\text{Var}(W) = \frac{n(n+1)(2n+1)}{24}$.
- Für n gross ist diese Teststatistik dann gemäss Grenzwertsatz approximativ normalverteilt,

$$\frac{W - \frac{1}{4}n(n + 1)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{a} \mathcal{N}(0, 1). \quad (9.1.4)$$

- Bei n klein werden *exakte* Verteilungen berechnet. Das macht der Computer, wir gehen auf diesbezügliche Methoden hier nicht ein.

Ist der wahre Median der Differenz δ_0 , erwarten wir $E(W) = E(W^-) = E(W^+) = n(n + 1)/4$, gerade die Hälfte der totalen Rangsumme $1 + \dots + n = n(n + 1)/2$ ¹. Je weiter die beobachtete Statistik von $E(W)$ entfernt ist, umso unwahrscheinlicher sind die Beobachtungen (oder extremere) unter H_0 .

Beispiel. Gegeben seien die Daten einer kleinen Vorher-Nachher-Studie:

```
X1 <- c(13.22, 6.81, 10.22, 14.03, 8.04, 10.16, 9.43, 13.07, 13.63, 5.05, 11.63)
X2 <- c(15.44, 6.69, 11.89, 16.25, 9.27, 10.74, 10.67, 13.52, 14.13, 7.21, 14.79)
Change <- X2 - X1
d.wilcox <- data.frame(X1 = X1, X2 = X2, Change = X2 - X1)
```

Zuerst eine deskriptive Übersicht:

```
summary(d.wilcox)

##          X1             X2            Change
##  Min.   : 5.05   Min.   : 6.69   Min.   :-0.12
##  1st Qu.: 8.73   1st Qu.: 9.97   1st Qu.: 0.54
##  Median :10.22   Median :11.89   Median : 1.24
##  Mean   :10.48   Mean   :11.87   Mean   : 1.39
```

¹Für die Summe der ersten n Zahlen aus \mathbb{N} gilt: $1 + 2 + \dots + n = n(n + 1)/2$.

```
## 3rd Qu.:13.14 3rd Qu.:14.46 3rd Qu.: 2.19
## Max. :14.03 Max. :16.25 Max. : 3.16
```

Wir wollen nun folgende Hypothese testen:

$$H_0 : \tilde{x}_2 - \tilde{x}_1 = 0, \quad H_1 : \tilde{x}_2 - \tilde{x}_1 \neq 0 \quad (9.1.5)$$

Implementation. Den Wilcoxon-Vorzeichen-Rang-Test machen wir mit `wilcox.test()`. Defaultmäßig wird ein exakter p -Wert berechnet für $n < 50$. Für $n \geq 50$ wird die Normalapproximation gebraucht. Da wir gepaarte Pre-Post Daten haben, müssen wir das Argument `paired=TRUE` setzen.

```
wilcox.test(X2, X1, paired = TRUE)

##
## Wilcoxon signed rank exact test
##
## data: X2 and X1
## V = 65, p-value = 0.002
## alternative hypothesis: true location shift is not equal to 0

# wilcox.test(X2-X1) ## äquivalent (ohne 'paired', da nur ein Vektor)
```

Wenn wir nur den p -Wert wollen:

```
wilcox.test(X2, X1, paired = TRUE)$p.value

## [1] 0.00195
```

Hier wird W als V bezeichnet und ist gleich 65. Die Nullhypothese ist also zu verwerfen.

Von Hand*. Wir können das Resultat “von Hand” gemäss obigem “Rezept” nachrechnen:

```
D <- (X2 - X1) #Differenzen
n <- length(D) #n
(R <- rank(abs(D))) #geordnete Differenzen
(Wplus <- sum(R[D > 0])) #stat1
(Wminus <- sum(R[D < 0])) #stat2
## approx Normalverteilung
(EW <- n * (n + 1)/4) #Erwartungswert unter H0
(VW <- (n * (n + 1) * (2 * n + 1))/24) #Varianz unter H0
(z <- (Wplus - EW)/sqrt(VW)) #z-Wert
(1 - pnorm(z)) * 2 #p-Wert approximative Normalverteilung
## exakt
(C1 <- qsignrank(p = 0.025, n) - 1) #kritischer Wert 1 (-1, weil diskret)
(C2 <- qsignrank(p = 0.975, n)) #kritischer Wert 2
(1 - psignrank(Wplus - 1, n)) * 2 #exakter p-Wert (Wplus-1, weil diskret)
psignrank(Wminus, n) * 2 #Alternative
```

Konfidenzintervall. Über das Argument `conf.int=TRUE` wird ein nicht-parametrisches Konfidenzintervall berechnet für den Pseudomedian².

```
wilcox.test(X2, X1, paired = TRUE, conf.int = TRUE)

##
## Wilcoxon signed rank exact test
##
## data: X2 and X1
## V = 65, p-value = 0.002
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## 0.56 2.19
## sample estimates:
## (pseudo)median
## 1.36
```

9.2 Wilcoxon-Rangsummentest

Der Wilcoxon-Test ist das nicht-parametrische Analogon zum gepaarten oder Einstichproben- *t*-Test. Der *Mann-Whitney-U-Test* oder *Wilcoxon-Rangsummentest* ist ein verteilungsfreier Test analog zum *t*-Test für zwei *unabhängige* Stichproben.

Er testet, ob es bei Betrachtung zweier Populationen gleich wahrscheinlich ist, dass ein zufällig aus der einen Population ausgewählter Wert grösser oder kleiner ist als ein zufällig ausgewählter Wert aus der anderen Population.

Sie testen nur unter folgender Annahme auch Gleichheit zweier Mediane: Die Zufallsvariable X und Y haben Verteilungen, die sich nur um eine *Verschiebung* a voneinander unterscheiden.

Wir haben unabhängige $X_1, \dots, X_n, Y_1, \dots, Y_m$. Wir wollen folgende Hypothesen testen:

$$H_0 : a = 0, \quad H_1 : a \neq 0.$$

Mann-Whitney-U-Statistik. Die Mann-Whitney-U-Statistik ist Anzahl, wie oft ein Element von X von einem Element von Y unterschritten wird,

$$U = \sum_{i=1}^n \sum_{j=1}^m I_{X_i > Y_j},$$

mit $I_{X>Y} = 1$, wenn $X > Y$, $I_{X>Y} = 1/2$, wenn $X = Y$ und sonst $I_{X>Y} = 0$. Man erwartet bei Gültigkeit von H_0 , bei guter “Durchmischung” der Ränge von X und Y , $n \cdot m/2$ Rangplatzunterschreitungen, kurz $E(U) = \frac{n \cdot m}{2}$, die Hälfte aller möglichen

²Der Pseudomedian von einer Verteilung F ist der Median der Verteilung $(u + v)/2$, mit u und v unabhängig, jede mit Verteilung F . Wenn F symmetrisch, dann ist der Pseudomedian gleich dem Median.

$(n \cdot m)$ Vergleiche. Ist wiederum der empirische U -Wert sehr verschieden von diesem Erwartungswert, dann sind die Daten (oder noch extremere) unter H_0 unwahrscheinlich.

Für $n > 3, m > 3$ und $m + n > 19$ kann man die Verteilung von U durch die Normalverteilung approximieren. Die kritischen Werte ergeben sich dann aus den kritischen Werten der approximativen Normalverteilung

$$U \stackrel{a}{\sim} \mathcal{N}\left(\frac{nm}{2}, \frac{nm(n+m+1)}{12}\right). \quad (9.2.1)$$

Bei n klein werden wieder *exakte* Verteilungen berechnet. Das macht der Computer im Hintergrund für uns, wir gehen auf diesbezügliche Methoden hier nicht ein.

Wir lehnen H_0 ab, wenn $U \leq u_{\alpha/2}$ (zweiseitig) oder wenn $U \leq u_{\alpha}$ (einseitig).

Beispiel.

```
X <- c(4.448, 5.767, 6.77, 4.28, 6.217, 5.645, 5.353, 5.706, 2.188, 9.991)
Y <- c(1.718, 9.382, 2.654, 4.269, 6.698, 8.002, 2.193, 5.808, 4.648)
```

Die beiden Vektoren haben verschiedene Längen, wir können sie daher nicht in eine $n \times 2$ Matrix oder in ein sogenanntes data frame im *wide*-Format darstellen. `data.frame(X, Y)` gibt eine Fehlermeldung.

```
data.frame(X, Y)
```

Eine Alternative ist, solche Daten in einem data frame im *long*-Format darzustellen und mit diesem Objekt weiterzuarbeiten.

```
outcome <- c(X, Y)
group <- c(rep("X", length(X)), rep("Y", length(Y)))
d.long <- data.frame(outcome = outcome, group = group) ## Daten im Long Format
d.long

##      outcome group
## 1      4.45     X
## 2      5.77     X
## 3      6.77     X
## 4      4.28     X
## 5      6.22     X
## 6      5.64     X
## 7      5.35     X
## 8      5.71     X
## 9      2.19     X
## 10     9.99     X
## 11     1.72     Y
## 12     9.38     Y
## 13     2.65     Y
## 14     4.27     Y
## 15     6.70     Y
```

```

## 16    8.00    Y
## 17    2.19    Y
## 18    5.81    Y
## 19    4.65    Y

by(d.long$outcome, d.long$group, summary)

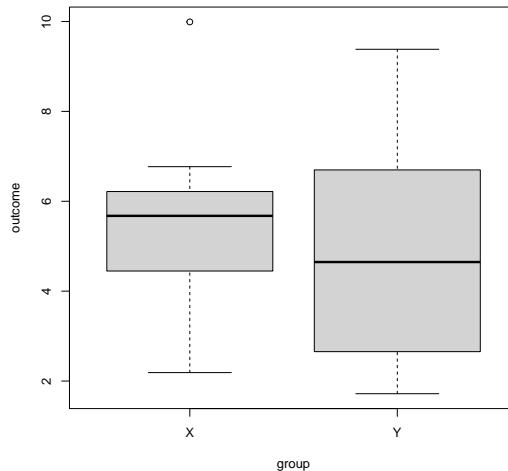
## d.long$group: X
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   2.19    4.67    5.68    5.64    6.10   9.99
## -----
## d.long$group: Y
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   1.72    2.65    4.65    5.04    6.70   9.38

by(d.long$outcome, d.long$group, psych::describe)

## d.long$group: X
##   vars n mean sd median trimmed mad min max range skew kurtosis se
##   X1   1 10 5.64 2 5.68    5.52 1.21 2.19 9.99    7.8 0.49    0.21 0.63
## -----
## d.long$group: Y
##   vars n mean sd median trimmed mad min max range skew kurtosis se
##   X1   1 9 5.04 2.66 4.65    5.04 3.04 1.72 9.38    7.66 0.23   -1.53 0.89

boxplot(outcome ~ group, d.long) ##formula version for boxplot

```



Implementation. Nun zum nicht-parametrischen Zwischengruppentest. Dazu brauchen wir wieder `wilcox.test()`, aber ohne `paired=TRUE`.

```
wilcox.test(X, Y) ## Default Version

##
## Wilcoxon rank sum exact test
##
## data: X and Y
## W = 51, p-value = 0.7
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(X, Y)$p.value

## [1] 0.661
```

Hier wird U als W bezeichnet und ist gleich 51. Die Nullhypothese kann nicht verworfen werden.

Von Hand*. Um ein bisschen hinter die Kulissen der nicht-parametrischen Statistik zu sehen, wollen wir dieses Resultat “von Hand” nachvollziehen (auch wenn wir später natürlich immer direkt obige Funktion brauchen): Der Erwartungswert bei $n = 10$ und $m = 9$ wäre $n \cdot m/2 = 45$. Wie oft wird ein Element von X durch ein Element von Y unterschritten? Dazu kann man die Funktion `outer()` brauchen. So können wir das U in obiger Funktion reproduzieren.

```
n <- length(X) #Länge X
m <- length(Y) #Länge Y
outer(X, Y, ">")

##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE
## [2,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE
## [3,] TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
## [4,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE
## [5,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE TRUE TRUE
## [6,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE
## [7,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE
## [8,] TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE
## [9,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [10,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

So wird z.B. X_{10} von Y_9 unterschritten, X_1 wird von Y_2 nicht unterschritten usw. Jetzt zählen wir die Anzahl Unterschreitungen (die TRUE's) zusammen. Das ist dann U :

```
U <- sum(outer(X, Y, ">"))
U

## [1] 51
```

Für die kritische(n) Werte gibt es wieder eine entsprechende Quantilfunktion.

```
(K1 <- qwilcox(0.025, n, m) - 1) ## kritischer Wert 1 (-1, weil diskret)

## [1] 20

(K2 <- qwilcox(0.975, n, m)) ## kritischer Wert 2

## [1] 69
```

Unsere Statistik ist nicht ausserhalb der kritischen Grenzen. Dasselbe Resultat sehen wir natürlich auch am p -Wert, dieser ist grösser als α .

```
(1 - pwilcox(U - 1, n, m)) * 2 ## (-1, weil U diskret)

## [1] 0.661
```

Formula-Schreibweise bei Data Frame im long-Format. Eine alternative Schreibweise für den Zwischengruppentest ist wieder die sogenannte Formula-Version, die man bei data frames im long-Format meistens braucht:

```
wilcox.test(outcome ~ group, data = d.long)
```

Konfidenzintervall. Über das Argument `conf.int=TRUE` wird wieder ein nicht-parametrisches Konfidenzintervall berechnet für den Unterschied im Lagemaß. Achtung. Es geht hier nicht um die Schätzung des Unterschieds der Mediane, sondern um die Schätzung des Medians der Differenz einer Stichprobe von X und einer Stichprobe von Y :

```
##
## Wilcoxon rank sum exact test
##
## data: outcome by group
## W = 51, p-value = 0.7
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -2.24 3.16
## sample estimates:
## difference in location
## 0.657
```

Güte von Klassifikatoren* Bei einem Zwischengruppenvergleich ist *Gruppe* die dichotome *unabhängige* Variable. Bei Problemen der *Klassifikation* kann man Gruppe aber auch als eine dichotome *abhängige* Variable betrachten, die man mit der anderen, einer mindestens ordinalskalierten unabhängigen Variablen, erklären möchte.

Es gibt jetzt einen schönen Zusammenhang zwischen der beschriebenen *U*-Statistik des Mann-Whitney Tests und einem Klassifikationsproblem. Die Mann-Whitney *U*-Statistik wird in der Diagnostik auch gebraucht als Schätzung für die sogenannte *Area under the curve* (AUC). Dies ist ursprünglich ein Mass aus der Signalentdeckungstheorie, das auch in der medizinischen Diagnostik oft benutzt wird für die *Güte* einer Klassifikation von Krankheit (die abhängige gruppierende Variable). AUC ist die Fläche unter der sogenannten *Reciever Operating Characteristics* (ROC) Kurve. In dieser wird die *Richtig-Positiv-Rate* (*Sensitivität*) gegen die *Falsch-Positiv-Rate* (*1-Spezifität*) für verschiedene Schwellenwerte aufgetragen (Abbildung 9.1). Auf die Begriffe der Sensitivität und Spezifität sind wir in Kapitel 3.5 eingegangen.

Dabei wird X_i als Testresultat bei *kranken*, Y_j als Testresultat bei *gesunden* Personen gesehen. Geschätzt wird hier $\Pr(X_i > Y_j)$, also die *Wahrscheinlichkeit, dass Testresultate bei Kranken mehr anzeigen als bei Gesunden*.

AUC ist dann einfach die relative Häufigkeit von Unterschreitungen in den mn Vergleichen, nämlich

$$AUC = \frac{U}{nm}.$$

Ist $AUC = 1$ (wenn die Kurve in 9.1 durch den Punkt $(0,1)$ gehen würde), geht das einher mit einer maximalen Richtig-Positiv-Rate ($Sn = 1$) und einer minimalen Falsch-Positiv-Rate ($1 - Sp = 0$) und damit mit einer *Likelihood-Ratio* eines positiven Tests von

$$LR+ = \frac{L(krank)}{L(gesund)} = \frac{\Pr(T+|krank)}{\Pr(T+|gesund)} = \frac{Sn}{1 - Sp} = \infty.$$

Das bedeutet auch, dass für alle mn Vergleiche $X_i > Y_j$ gilt, und dass der Test perfekt ist. Ist $AUC = 0.5$ (wenn die ROC-Kurve in 9.1 eine Gerade mit Winkel 45 Grad ist), geht das einher mit gleich grosser Sn und $1 - Sp$ und damit mit

$$LR+ = \frac{L(krank)}{L(gesund)} = \frac{\Pr(T+|krank)}{\Pr(T+|gesund)} = \frac{Sn}{1 - Sp} = 1.$$

Die Klassifikation also rein zufällig, das heisst, der *Klassifikator* wäre dann wertlos.

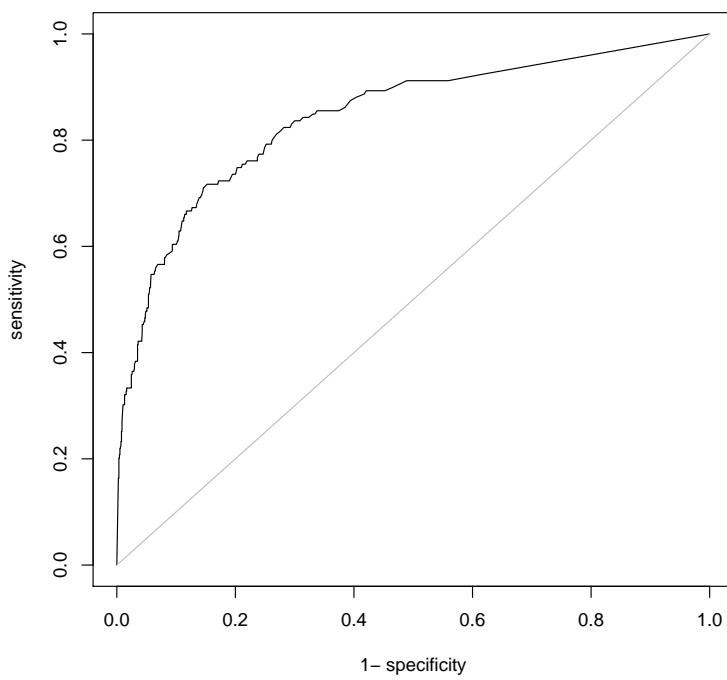


Abbildung 9.1: Reciever Operating Characteristics (ROC) Analyse mit Area under the curve $AUC = 0.84$. Wenn $AUC = 0.5$ (Gerade), ist die Klassifikation rein zufällig.

Kapitel 10

Verfahren für Häufigkeitsdaten

Bei den χ^2 -Verfahren handelt es sich um die Analyseverfahren für *nominalskalierte*, also kategoriale Variablen mit k Ausprägungsgraden (mit $k \geq 2$).

χ^2 -Verfahren kommen zur Anwendung bei

- Vergleich einer empirisch beobachteten Verteilung mit einer theoretischen Verteilung (“Verteilungstest” oder “Anpassungstest”)
- Prüfung auf Unabhängigkeit von nominalskalierten Variablen (“Unabhängigkeitstest”)
- Vergleich der Verteilung von zwei oder mehr Stichproben (“Homogenitätstest”)

Der Test zum Vergleich von unabhängigen Stichproben (“Homogenitätstest”) ist identisch mit dem Test auf Unabhängigkeit zweier Variablen. Daher behandeln wir im Folgenden Verteilungstests und Unabhängigkeitstests.

Bedingungen. Folgende Bedingungen müssen für einen χ^2 -Test erfüllt sein. Keine erwartete Häufigkeit darf 0 oder 1 sein. Es sollten höchstens 20% der erwarteten Häufigkeiten kleiner als 5 sein.

10.1 Eine kategoriale Variable

Wir erläutern das Prinzip von χ^2 -Tests anhand des Vergleichs der Verteilung eines Merkmals mit 2 oder mehr ($k > 2$) Kategorien mit einer theoretischen Verteilung.

Sei das Merkmal das *Geschlecht G* und vergleichen wir die Verteilung des Geschlechts in einer Stichprobe mit einer theoretischen oder einer postulierten Verteilung. Das heisst, dass wir eine Nullhypothese

$$H_0 : \pi_M = \pi_F = 0.5$$

gegen die Alternativhypothese

$$H_1 : \pi_M \neq \pi_F$$

testen.

Die empirischen Daten zeigen bei einer Stichprobe von 300 Personen folgende absolute Häufigkeiten: 159 Männern und 141 Frauen. Unter der Nullhypothese sind die erwarteten Häufigkeiten also je 50% von 300, also 150.

Wir können jetzt anhand der Daten eine Teststatistik entwickeln, die die Abweichung von der Nullhypothese bewertet. Man vergleicht hierzu für jede Zelle i die beobachtete Häufigkeit X_i mit der unter der Nullhypothese zu erwartende Häufigkeit $n\pi_i$. Dazu betrachtet man die *quadrierten Differenzen* zwischen den empirischen und theoretischen Werten und normiert diese Differenzen geeignet, um die Verteilung der Teststatistik bestimmen zu können. Die Teststatistik ist dann die *Pearson- χ^2 -Statistik*,

$$\boxed{\chi^2 = \sum_{i=1}^k \frac{(X_i - n\pi_i)^2}{n\pi_i}}. \quad (10.1.1)$$

Diese Statistik ist unter H_0 χ^2 -verteilt mit $k - 1$ Freiheitsgraden. Das wollen wir zeigen für den Spezialfall mit $k = 2$:

Beweis. Für den Spezialfall von zwei Zellen folgen die Anzahlen einer Binomialverteilung, $X \sim \text{Bin}(\pi, n)$ mit Erwartungswert $n\pi$ und Varianz $n\pi(1 - \pi)$ (siehe 4.2). Wenn n gross ist, kann man diese Verteilung aufgrund vom Grenzwertsatz durch eine Normalverteilung approximieren,

$$\text{Bin}(\pi, n) \xrightarrow{a} \mathcal{N}(n\pi, n\pi(1 - \pi)).$$

Sei nun X_1 die Anzahl an Beobachtungen des Samples in der ersten Zelle. Da wir zwei Zellen haben, ist damit auch die Anzahl X_2 durch das Total n gegeben. Die Pearson-Statistik ist dann

$$\frac{(X_1 - n\pi)^2}{n\pi} + \frac{(n - X_1 - n(1 - \pi))^2}{n(1 - \pi)}.$$

Durch Umformen zeigt man (Übung), dass diese Grösse gleich

$$\left(\frac{X_1 - n\pi}{\sqrt{n\pi(1 - \pi)}} \right)^2 = Z^2$$

ist und somit eine *quadrierte standardnormalverteilte* Zufallsvariable darstellt, und damit χ^2 -verteilt ist mit 1 Freiheitsgrad. (siehe 4.4.1). Für mehrwertige ($k > 2$) Zufallsvariable ist der Beweis analog, aber komplizierter. \square

Voraussetzung ist jedoch (damit wir den Grenzwertsatz brauchen dürfen), dass die erwarteten Häufigkeiten alle grösser als 1 sind und dass für mindestens 80% der Zellen die erwarteten Häufigkeiten $(n\pi_i) \geq 5$ sind.

Für unser Beispiel ergibt sich als empirischer χ^2

$$\chi^2 = \frac{(159 - 150)^2}{150} + \frac{(141 - 150)^2}{150} = 1.08. \quad (10.1.2)$$

Das 0.95-Quantil der χ^2 -Verteilung mit 1 Freiheitsgrad bestimmen und damit der kritische Wert ist $\chi^2 = 3.84$.

```
qchisq(p = 0.95, df = 1)
```

```
## [1] 3.84
```

Unsere Daten sind also unter H_0 nicht extrem genug, als dass man die Nullhypothese der Gleichverteilung verwerfen könnte. Die Daten sind mit der Nullhypothese gut vereinbar.

In R ist dieser Test implementiert mit `chisq.test()`. Wir müssen einen Vektor mit den Daten übergeben.

```
dataCat1 <- c(M = 159, W = 141)
chisq.test(dataCat1)

##
## Chi-squared test for given probabilities
##
## data: dataCat1
## X-squared = 1, df = 1, p-value = 0.3
```

Den p -Wert könnten wir reproduzieren mit

```
chisqValue <- chisq.test(dataCat1)$stat
1 - pchisq(q = chisqValue, df = 1)
```

Anstatt der Hypothese der Gleichverteilung $\pi_M = \pi_F$ könnte man auch eine andere Hypothese testen, z.B. $\pi_M = 0.45, \pi_F = 0.55$. Dies ergibt dann aufgrund der erwarteten absoluten Häufigkeit von $0.45 \cdot 300 = 135$ für Männer und $0.55 \cdot 300 = 165$ für Frauen den empirischen χ^2 von

$$\chi^2 = \frac{(159 - 135)^2}{135} + \frac{(141 - 165)^2}{165} = 7.76.$$

Anhand der Daten und aufgrund vom überschrittenen $\chi^2_{0.95}$ -Quantil von 3.84 wäre diese Nullhypothese zu verwerfen. Bei Nullhypotesen, die nicht der Default-Einstellung entsprechen (Gleichwahrscheinlichkeit für alle Zellen) muss das Argument `p=` in `chisq.test` spezifiziert werden.

```
data <- c(M = 159, W = 141)
chisq.test(data, p = c(0.45, 0.55))

##
## Chi-squared test for given probabilities
##
## data: data
## X-squared = 8, df = 1, p-value = 0.005
```

Die Nullhypothese bezüglich der theoretischen Verteilung $H_0 : \pi_M = 0.45, \pi_F = 0.55$ wird hier verworfen.

Für eine kategoriale Zufallsvariable mit mehr als zwei Ausprägungsgraden ist die Vorgehensweise ganz analog. Wir haben dann einfach in der χ^2 -Statistik k Summanden und $k - 1$ Freiheitsgrade.

Beispiel. Es wurde gemessen, in was für einer Schicht $n = 300$ Patienten mit Depression leben. Die dreiwertige Variable hatte die Kategorien Unterschicht (U), Mittelschicht (M) oder Oberschicht (O). Die Daten zeigten folgende absolute Häufigkeiten: $X_U = 62$, $X_M = 155$ und $X_O = 83$.

Die zu testende Nullhypothese sei nun $\pi_U = 0.3$, $\pi_M = 0.5$, $\pi_O = 0.2$. Dann ist die χ^2 -Statistik

$$\chi^2 = \sum_{i=1}^3 \frac{(X_i - 300\pi_i)^2}{n\pi_i} = \frac{(62 - 90)^2}{90} + \frac{(155 - 150)^2}{150} + \frac{(83 - 60)^2}{60} = 17.69. \quad (10.1.3)$$

Der Freiheitsgrad der Verteilung ist $k - 1 = 2$ und somit ist der kritische Wert, das 0.95-Quantil dieser Verteilung

```
qchisq(p = 0.95, df = 2)
## [1] 5.99
```

Die Nullhypothese kann also verworfen werden. Die Implementation in R ist hier analog zu oben.

```
data3cat <- c(62, 155, 83)
chisq.test(data3cat, p = c(0.3, 0.5, 0.2))

##
## Chi-squared test for given probabilities
##
## data: data3cat
## X-squared = 18, df = 2, p-value = 0.0001
```

Auch hier würde die Nullhypothese bezüglich der theoretischen Verteilung verworfen.

10.2 Zwei kategoriale Variablen

Beim χ^2 -Unabhängigkeitstest wird auf Unabhängigkeit zwischen zwei nominalskalierten Zufallsvariablen getestet. Die Hypothese H_0 : "Zwei Merkmale sind voneinander unabhängig" nimmt für zwei kategoriale Merkmale eine einfache Form an. Aufgrund (3.4.1) gilt nämlich bei Unabhängigkeit

$$H_0 : \Pr(X_1 = i, X_2 = j) = \Pr(X_1 = i) \times \Pr(X_2 = j), \quad \text{für alle } i, j. \quad (10.2.1)$$

Als Beispiel betrachten wir die $k \cdot l$ -Tafel 10.1 mit $k = 2$ und $l = 3$. Die Variable X_1 = Antwortverhalten hat zwei Kategorien und X_2 = Freundlichkeit hat drei Kategorien.

$n = 77$	freundlich	forsch	sachlich	total
ja	13	3	5	21
nein	15	22	19	56
total	28	25	24	77

Tabelle 10.1: Unabhängigkeitstest: Bivariate Häufigkeiten.

Beschreibende Statistik. Zuerst wollen wir hier wichtige und häufig benutzte R-Funktionen im Umgang/Beschreibung mit/von kategorialen Daten wiederholen.

Obige Daten können reproduziert werden, als matrix- oder table-Objekt mit

```
data23<-matrix(c(13,15,3,22,5,19),ncol=3,
                 dimnames=list(Antwort=c("ja","nein"),
                               Freundlichkeit=c("freundlich","forsch","sachlich")))
data23

##      Freundlichkeit
## Antwort freundlich forsch sachlich
##   ja          13     3    5
##   nein         15    22   19
```

Diese 2×3 Matrix kann man auch als `table` und als `data.frame` haben:

```
(dtable <- as.table(data23))

##      Freundlichkeit
## Antwort freundlich forsch sachlich
##   ja          13     3    5
##   nein         15    22   19

(dframe <- as.data.frame(dtable))

##   Antwort Freundlichkeit Freq
## 1      ja    freundlich   13
## 2    nein    freundlich   15
## 3      ja      forsch    3
## 4    nein      forsch   22
## 5      ja      sachlich   5
## 6    nein      sachlich  19
```

Hier wird dann jede Kombination auf den beiden kategorialen Variablen mit ihren Anzahlen dargestellt.

Kurzer Ausflug: Die ursprünglichen Rohdaten würden wie folgt aussehen, ein data frame mit $n = 77$ Beobachtungen (Zeilen) auf den beiden Variablen. (Code für die Interessierten.)

```
dfRaw <- dframe[rep(1:nrow(dframe), dframe$Freq), -3]
psych::headTail(dfRaw)

##      Antwort Freundlichkeit
## 1      ja    freundlich
## 1.1    ja    freundlich
## 1.2    ja    freundlich
## 1.3    ja    freundlich
## ...   <NA>      <NA>
## 6.15   nein   sachlich
## 6.16   nein   sachlich
## 6.17   nein   sachlich
## 6.18   nein   sachlich

## DescTools:::Untable(data23) #Alternative Funktion
```

Aus einem solchen ursprünglichen data frame würde man mit dann mit

```
table(dfRaw)
```

wieder zu obiger Tabelle kommen.

Zurück zu unserem Problem. Wenn wir nur die Randsummen wollen, können wir das mit `marginSums()` verlangen. Man muss sich dann entscheiden, auf was man bedingen will.

- Bedingte Häufigkeiten gegeben die Zeilen, mit `margin=1`.

```
marginSums(data23, margin = 1)

## Antwort
##   ja nein
## 21 56
```

- Bedingte Häufigkeiten gegeben die Spalten, mit `margin=2`

```
marginSums(data23, margin = 2)

## Freundlichkeit
## freundlich forsch sachlich
##     28     25     24
```

Mit `addmargins()` können wir die Randsummen hinzufügen.

```
addmargins(data23)

##      Freundlichkeit
## Antwort freundlich forsch sachlich Sum
##   ja           13     3     5  21
##   nein          15    22    19  56
##   Sum           28    25    24  77
```

Oft will man statt absolute eine Tabelle mit relativen Häufigkeiten.

```

proportions(data23) #unbedingt

##      Freundlichkeit
## Antwort freundlich forsch sachlich
##   ja      0.169  0.039  0.0649
##   nein     0.195  0.286  0.2468

proportions(data23, margin = 1) #bedingt, gegeben Zeilen

##      Freundlichkeit
## Antwort freundlich forsch sachlich
##   ja      0.619  0.143  0.238
##   nein     0.268  0.393  0.339

proportions(data23, margin = 2) #bedingt, gegeben Kolonnen

##      Freundlichkeit
## Antwort freundlich forsch sachlich
##   ja      0.464  0.12   0.208
##   nein     0.536  0.88   0.792

```

`summary()` angewandt auf eine Kreuztabelle (als `table`, nicht als `matrix`) macht auch den χ^2 -Test, auf den wir jetzt eingehen.

```

summary(dtable)

## Number of cases in table: 77
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9, df = 2, p-value = 0.01

```

Zurück zum Test. Wenn H_0 wahr ist, wenn also X_1 und X_2 voneinander unabhängig sind, dann gilt gemäss (10.2.1) für jede einzelne Zelle, dass die *gemeinsame Wahrscheinlichkeit oder Verbundwahrscheinlichkeit* gleich dem Produkt der Randwahrscheinlichkeiten ist:

$$\pi_{ij} = \pi_i \cdot \pi_j, \quad \text{für alle } i, j.$$

Dies folgt wieder aus (3.4.1) und aus (10.2.1). Betrachten wir die erste Zelle. Unter H_0 ist die *gemeinsame Wahrscheinlichkeit* "freundlich" und "Jasager" zu sein: $\pi_{11} = \frac{28}{77} \cdot \frac{21}{77}$. Diese Wahrscheinlichkeit multiplizieren wir mit $n = 77$ und erhalten so die für diese Zelle erwartete Häufigkeit. Dieselbe Prozedur machen wir für alle 6 Zellen in der Tabelle und erhalten so alle erwarteten Häufigkeiten unter H_0 in Tabelle 10.2.

Wir kreieren nun die Teststatistik gemäss (10.1.1). Diese ist

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(X_{ij} - 77 \cdot \pi_{ij})^2}{77 \cdot \pi_{ij}}. \quad (10.2.2)$$

$n = 77$	freundlich	forsch	sachlich	total
ja	7.64	6.83	6.54	21
nein	20.36	18.18	17.46	56
total	28	25	24	77

Tabelle 10.2: Unabhängigkeitstest: Erwartete Häufigkeiten.

Diese Statistik ist χ^2 -verteilt mit $(k-1) \cdot (l-1) = 1 \cdot 2 = 2$ Freiheitsgraden. Ausgeschrieben ergibt obige Statistik für die empirischen Daten

$$\chi^2 = \frac{(13 - 7.64)^2}{7.64} + \frac{(3 - 6.82)^2}{6.82} + \dots + \frac{(19 - 17.46)^2}{17.46} = 8.61.$$

Unabhängigkeit von Freundlichkeit und Antwortverhalten kann somit verworfen werden.

Approximativer Test auf Unabhängigkeit in R. Als χ^2 -Test auf Unabhängigkeit (approximative χ^2 -Verteilung) ist dieser Test wieder mit `chisq.test()` implementiert. Man übergibt jetzt aber eine Kreuztabelle.

```
chisq.test(data23)

##
##  Pearson's Chi-squared test
##
## data: data23
## X-squared = 9, df = 2, p-value = 0.01
```

Die Wahrscheinlichkeit der Daten (oder extremerer Daten) ist also unter der Unabhängigkeitsannahme kleiner als 5%. Diese Annahme wird also verworfen. Es besteht ein statistisch signifikanter Zusammenhang zwischen Freundlichkeit und Antwortverhalten.

Exakter Test auf Unabhängigkeit in R. *Fisher's exakter Test* basiert nicht auf der χ^2 -Verteilung, sondern auf *fixierten Randverteilungen* und auf der *hypergeometrischen Verteilung*.

Die Grundidee ist folgende:

- Man bildet dazu alle möglichen Kreuztabellen mit den *gleichen Randhäufigkeiten* wie die beobachtete Kreuztabelle.
- Dann berechnet man die Wahrscheinlichkeit, mit der eine solche Kreuztabelle unter der Nullhypothese entsteht (mit der hypergeometrischen Verteilung).
- Dann addiert man alle Wahrscheinlichkeiten, die nicht grösser sind als die von der beobachteten Kreuztabelle.

- Diese Summe ist dann der p -Wert.

Wir haben hier wieder ein schönes Beispiel, was der p -Wert darstellt: Die Wahrscheinlichkeit der beobachteten – oder extremeren Daten – gegeben die Nullhypothese. Fisher's exakter Test ist eher konservativ, d.h. das effektive α -Niveau ist eher kleiner und damit die effektive Falsch-Positiv Rate kleiner.

Obiger Vorgang wäre von Hand sehr mühsam, das ist inhärent bei nicht-parametrischen Verfahren. Zum Glück aber ist dieser Test implementiert mit der Funktion `fisher.test()`

```
fisher.test(data23)

##
## Fisher's Exact Test for Count Data
##
## data: data23
## p-value = 0.01
## alternative hypothesis: two.sided
```

Das Resultat ist hier also praktisch identisch wie beim approximativen Test. Dieser Test ist immer dann zu brauchen, wenn die Annahmen für die approximative χ^2 -Verteilung nicht erfüllt sind, wenn also die Häufigkeiten in den Zellen klein ist.

Wir kommen am Schluss zurück zu den Daten der Studie, die wir schon in [7.5.2](#) behandelt haben. Dort hatten wir ein Chancenverhältnis OR und ein Risikoverhältnis RR abgeschätzt. Die Daten waren

```
morkved.mat <- matrix(c(48, 100, 74, 79), byrow = TRUE, nrow = 2, dimnames = list(c("Training", "Control"),
  c("positiv", "negativ")))
morkved.mat

##           positiv negativ
## Training      48     100
## Control       74      79
```

Zusätzlich zur Intervallschätzung haben die Autoren einen “exact computation of the Pearson χ^2 -Test” gemacht, wie in den Methoden beschrieben. Siehe die Resultate in der Publikation [33] (Tabelle 2, S. 316, Zeile 36 wk). Die χ^2 -Statistik und den exakten p -Wert können wir aber reproduzieren, indem wir in `chisq.test()` das Argument `simulate.p.value=TRUE` setzen, was dann einen Fisher-Test macht.

```
chisq.test(morkved.mat, simulate.p.value = TRUE, B = 100000)

##
## Pearson's Chi-squared test with simulated p-value (based on 100000 replicates)
##
## data: morkved.mat
## X-squared = 8, df = NA, p-value = 0.007

fisher.test(morkved.mat)$p.value

## [1] 0.00681
```

Natürlich wissen wir bereits, dass das 95%-Konfidenzintervall (für OR oder RR) den Nulleffekt ($OR = RR = 1$) nicht beinhaltete.

Gepaarte Daten. Bei den χ^2 -Verfahren für nominalskalierte Daten gibt es wie beim t -Test für gepaarte Stichproben und wie beim Wilcoxon-Test einen Vorher-Nachher-Vergleich für *gepaarte* Daten. Der *McNemar-Test* macht das für eine zweiwertige Variable. Er prüft bei einer verbundenen Stichprobe, ob eine Veränderung eingetreten ist. Das ist ein Test auf Symmetrie der Zeilen und Kolonnen in der Kreuztabelle. Der *McNemarBowker Test* ist eine Erweiterung für gepaarte Variablen mit mehr als 2 Kategorien.

Bei zweiwertiger Variable sehen die Daten folgendermassen aus (Häuf):

```
##           Messung 2
## Messung 1 nein ja
##      nein a   b
##      ja    c   d
```

Unter der Hypothese keiner Veränderung müssten die Randsummen bezüglich Zeilen und Kolonnen gleich sein,

$$H_0 : a + b \approx a + c. \quad (10.2.3)$$

Daher der Name “Test auf Symmetrie”. Unter Gültigkeit der Nullhypothese sind die erwarteten Zellhäufigkeiten dann $(b + c)/2$, und es ergibt sich die Teststatistik

$$\chi^2 = \frac{(b - \frac{b+c}{2})^2}{\frac{b+c}{2}} + \frac{(c - \frac{b+c}{2})^2}{\frac{b+c}{2}} = \frac{(b - c)^2}{b + c}. \quad (10.2.4)$$

Anbei reproduzieren wir die Daten aus dem Hilfe-File der `mcnemar.test()`-Funktion.

```
## Agresti (1990), p. 350. Presidential Approval Ratings. Approval of the President's performance
## in office in two surveys, one month apart, for a random sample of 1600 voting-age Americans.
```

```
Performance <- matrix(c(794, 86, 150, 570),
                       nrow = 2, dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                                                 "2nd Survey" = c("Approve", "Disapprove")))
```

```
Performance

##           2nd Survey
## 1st Survey Approve Disapprove
##   Approve     794      150
##   Disapprove    86      570
```

Die Rohdaten wären ursprünglich ein data frame mit 1600 Beobachtungen. Das könnten wir haben mit

```
dframe <- as.data.frame(as.table(Performance))
dfRaw <- dframe[rep(1:nrow(dframe), dframe$Freq), -3]
str(dfRaw)

## 'data.frame': 1600 obs. of 2 variables:
## $ X1st.Survey: Factor w/ 2 levels "Approve","Disapprove": 1 1 1 1 1 1 1 1 1 ...
## $ X2nd.Survey: Factor w/ 2 levels "Approve","Disapprove": 1 1 1 1 1 1 1 1 1 ...

psych::headTail(dfRaw)

##          X1st.Survey X2nd.Survey
## 1           Approve     Approve
## 1.1         Approve     Approve
## 1.2         Approve     Approve
## 1.3         Approve     Approve
## ...           <NA>       <NA>
## 4.566    Disapprove   Disapprove
## 4.567    Disapprove   Disapprove
## 4.568    Disapprove   Disapprove
## 4.569    Disapprove   Disapprove
```

Mit `addmargins()` können wir die Randsummen anfügen.

```
addmargins(Performance)

##          2nd Survey
## 1st Survey Approve Disapprove Sum
##   Approve     794      150 944
## Disapprove     86      570 656
##   Sum        880      720 1600
```

Der Test ist implementiert mit `mcnemar.test()`, übergeben wird die Kreuztabelle (als Matrix oder Table):

```
mcnemar.test(Performance, correct = FALSE)

##
## McNemar's Chi-squared test
##
## data: Performance
## McNemar's chi-squared = 17, df = 1, p-value = 3e-05
```

Für kleine Stichproben macht R eine Stetigkeitskorrektur (Default ist `correct=TRUE`). Die χ^2 -Statistik und den p -Wert vom Output können wir reproduzieren mit

```
chi2 <- (Performance[1, 2] - Performance[2, 1])^2/(Performance[1, 2] + Performance[2, 1])
chi2

## [1] 17.4

1 - pchisq(chi2, df = 1)

## [1] 0.000031
```

Die Daten sind nicht kompatibel mit der Nullhypothese. Wir haben eine statistisch signifikante Verminderung der Zustimmung von der ersten zur zweiten Periode.

Zur Illustration ein Szenario, bei dem die Daten gut kompatibel sind mit der Nullhypothese (Dichotome Ratings von “Blockade Iliosakralgelenk”)

```
ISGtest <- matrix(c(101, 46, 42, 102), nrow = 2, dimnames = list(Rating1 = c("Ja", "Nein"), Rating2 = c("Ja", "Nein")))
addmargins(ISGtest)

##           Rating2
## Rating1   Ja Nein Sum
##     Ja    101   42 143
##     Nein  46   102 148
##     Sum    147   144 291

mcnemar.test(ISGtest)

##
## McNemar's Chi-squared test with continuity correction
##
## data: ISGtest
## McNemar's chi-squared = 0.1, df = 1, p-value = 0.7
```

Exakter Test*. Mit π als der Wahrscheinlichkeit, dass eine Beobachtung “links unten” landet, haben wir als Nullhypothese (der Symmetrie) $H_0 : \pi = 0.5$, und die Statistik B ist dann $B \sim \text{Bin}(b + c, 0.5)$.

Teil IV

Lineare Modelle

Ziel. Aufbau eines Verständnisses für statistische Modellierung mit linearen Zusammenhängen. Ziel ist es, Beziehungen zwischen Variablen zu quantifizieren und zu interpretieren.

Kapitel 11

Lineare Modelle

In diesem Kapitel führen wir die grundlegenden Begriffe ein bezüglich *linearen Modellen*. Dazu wiederholen wir die linearen Funktionen aus der Schulzeit und erweitern diese mit einer stochastischen Komponente.

11.1 Lineare Funktionen

Aus der Schule kennen wir die grundlegende *lineare Funktion* $f : \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) = a + b \cdot x, \quad a, b \in \mathbb{R}. \quad (11.1.1)$$

Der Graph dieser Funktion ist eine *Gerade*. Dabei wird $y = f(x)$ auf der Vertikalen und x auf der Horizontalen aufgezeichnet, a ist der “ y -Achsenabschnitt” der Geraden und b ist die *Steigung* der Geraden. **Diese lineare Funktion – und Verallgemeinerungen davon – spielen eine ganz zentrale Rolle in der Wissenschaft.** Abbildung 11.1 zeigt eine lineare Funktion $y = 4 + 5x$ und eine nichtlineare Funktion $y = \exp(x)$.

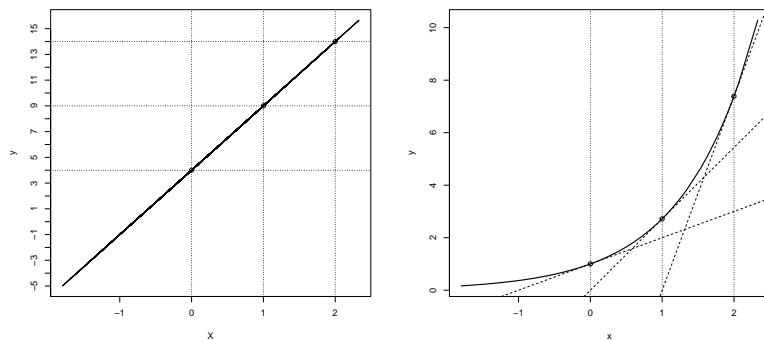


Abbildung 11.1: Links: Lineare Funktion $y = 4 + 5x$, $a = 4$ und $b = 5$ können abgelesen werden. Rechts: Exponentialfunktion $y = \exp(x)$ mit nicht konstanter Steigung.

Insbesondere ist bei der linearen Funktion die Steigung $b = 5$ eine zentrale Grösse, die wir später in linearen Modellen oft antreffen werden. Damit wir diese Grösse auch allgemein gut verstehen, brauchen wir den Begriff der Ableitung.

Definition. Die *Ableitung* $f'(x_0)$ einer Funktion f an einem Punkt x_0 ist

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}. \quad (11.1.2)$$

Abbildung 11.2 zeigt die Ableitung einer Funktion bei $x_0 = 0$ als Grenzwert von Steigungen von Sekanten, wenn $x \rightarrow x_0$ (Steigung der roten Tangente).

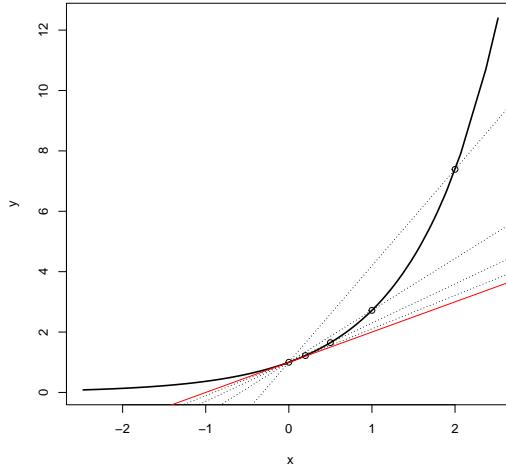


Abbildung 11.2: Ableitung einer Funktion an der Stelle $x_0 = 0$

Bei einer linearen Funktion $f(x) = a + bx$ ist die Ableitung oder Steigung $f'(x) = b$, also über alle x konstant.

$$\text{Beweis. } f'(x_0) = \lim_{x \rightarrow x_0} \frac{a + bx - (a + bx_0)}{x - x_0} = \lim_{x \rightarrow x_0} \frac{b(x - x_0)}{x - x_0} = \lim_{x \rightarrow x_0} b = b \quad \square$$

Diese Grösse ist dann zu interpretieren als die *Veränderung auf y pro Einheit Veränderung auf x*,

$$f'(x) = b = \frac{\Delta y}{\Delta x}. \quad (11.1.3)$$

In der Abbildung 11.1 ist sie auf der Y-Achse ablesbar. Bei einer Einheit Veränderung auf x verändert sich y um 5, bei zwei Einheiten um 10, usw. Ebenso ist a als der Wert der Funktion bei $x = 0$ erkennbar. Bei einer linearen Funktion ist die Steigung, der Zuwachs pro Einheit, überall gleich gross, nämlich b . Im Allgemeinen aber, wie

z.B. bei der Exponentialfunktion, ist das nicht der Fall. Jede Exponentialfunktion steigt irgendwann viel schneller als jede lineare Funktion mit noch so grosser Steigung. Die Ableitung der Exponentialfunktion ist $f'(x) = \exp(x)$. Der Zuwachs wächst also selber exponentiell. Lineare Funktionen nehmen also einen ganz speziellen Platz ein in der Menge aller Funktionen.

11.2 Lineare Regression

Eine (Pearson) Korrelation haben wir eingeführt als ein Mass für die Stärke des (linearen) Zusammenhangs zwischen zwei Variablen X und Y . Wenn wir wissen wollen, wie die *Art* des Zusammenhangs ist, müssen wir den Daten ein (lineares) *Modell* zugrunde legen, das uns erlaubt, Y durch X zu erklären.

Man sagt dann, man macht eine *Regression* von der *Zielgrösse* Y auf die *Eingangsgrösse*¹ X . Die Zielgrösse Y nennen wir manchmal auch *Kriterium*, *Outcome* oder *Response*, die Eingangsgrösse X manchmal auch *Prädiktor* oder *Kovariable*.

Einfache Lineare Regression. Das grundlegende Regressionsmodell ist das *einfache lineare Modell*, ein Modell mit *einer* Eingangsgrösse X . In diesem Modell ist die Zielgrösse Y die Summe aus einer linearen Funktion von x und einem *zufälligen* Messfehler ϵ . Die Zielgrösse Y_i wird modelliert mit

$$Y_i = \underbrace{\alpha + \beta \cdot x_i}_{\text{linearer Prädiktor}} + \underbrace{\epsilon_i}_{\text{Messfehler}}, \quad i = 1, \dots, n. \quad (11.2.1)$$

Wir unterscheiden in 11.2.1 einen systematischen und einen zufälligen Teil,

$$Y_i = \underbrace{\alpha + \beta x_i}_{\text{systematischer Teil}} + \underbrace{\epsilon_i}_{\text{zufälliger Teil}}.$$

Im Gegensatz zu einer linearen *Funktion* hat es also in einem linearen *Modell* zusätzlich eine stochastische, eine Zufallsgrösse. Oft wird angenommen, dass die Abweichungen, die stochastischen Fehler ϵ_i , $i = 1, \dots, n$, eine bestimmte Verteilung haben mit Erwartungswert 0, z.B. eine *Normalverteilung*, und dass sie *stochastisch unabhängig*, also nicht korreliert sind. Sie bilden dann eine i.i.d. Zufalls-Stichprobe. Bei Normalverteilung notieren wir dann $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, die ϵ_i sind i.i.d. (für *independent and identically distributed*). Da der Erwartungswert des Fehlers ϵ_i Null ist, ist

$$\mathbb{E}(Y_i | x_i) = \alpha + \beta x_i. \quad (11.2.2)$$

¹regredi=zurückgehen

Diese Grösse nennt man den *linearen Prädiktor*. Somit ist

$$\mathbb{E}(Y_i | x_i + 1) - \mathbb{E}(Y_i | x_i) = \alpha + \beta(x_i + 1) - (\alpha + \beta x_i) = \beta \quad (11.2.3)$$

die Veränderung im Erwartungswert pro Einheit Veränderung auf der Eingangsgrösse. Der Parameter β kann theoretisch Werte zwischen $-\infty$ und $+\infty$ einnehmen. Wir werden sehen, dass viele Quantitäten von Interesse (*Quantities of interest*) in der Wissenschaft solche “Steigungen” sind.

Die beiden *unbekannten Parameter* α und β heissen *Regressionskoeffizienten* und sind aus den Daten zu *schätzen*. Die geschätzten Grössen bezeichnen wir dann mit $\hat{\alpha}$ und $\hat{\beta}$. Dazu müssen wir eine Gerade so durch die Punktewolke legen, dass diese in einem gewissen Sinn *optimal* ist, dass diese Gerade “im Schnitt am nächsten zu allen Punkten ist” (Abbildung 11.3).

Kleinste-Quadrat Schätzer. Dies geschieht mit der sogenannten *Methode der kleinsten Quadrate* (Least Squares). Diese Methode wählt diejenigen Parameter α und β , die die *Summe der quadrierten Residuen* (“Reste”) des Modells minimiert: Das i -te Residuum r_i ist

$$r_i = y_i - \hat{y}_i \quad (11.2.4)$$

$$= y_i - (\hat{\alpha} + \hat{\beta}x_i) \quad (11.2.5)$$

mit $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ als dem *angepassten* Wert. Die angepassten Werte sind in Abbildung 6.8 durch Kreuze markiert. Die Residuen sind durch die gestrichelten Linien markiert.

Man bestimmt also diejenigen $\hat{\alpha}$ und $\hat{\beta}$, für die die Quadratsumme $\sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2$ minimal wird. Man hat damit ein *Optimierungsproblem* zu lösen, nämlich

$$\arg \min_{\alpha, \beta} \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2. \quad (11.2.6)$$

Wir werden später für den allgemeinen Fall der multiplen Regression eine allgemeine Herleitung machen für die Lösung dieses Problems. Für die einfache Regression ist die Lösung

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \quad (11.2.7)$$

Beweis* Der zu minimierende Ausdruck wird nach α und β abgeleitet und die entsprechenden Nullstellen werden gesucht. $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ nach α ableiten und nullsetzen gibt für $\hat{\alpha}$:

$$-2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$ nach Einsetzen von α nach β ableiten und nullsetzen gibt für $\hat{\beta}$:

$$\begin{aligned} \sum_{i=1}^n (y_i - (\bar{y} - \beta\bar{x}) - \beta x_i)^2 &= \sum_{i=1}^n [(y_i - \bar{y}) - \beta(x_i - \bar{x})]^2 \\ \Rightarrow -2 \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})](x_i - \bar{x}) &= 0 \\ \Rightarrow \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x})^2 &= 0 \\ \Rightarrow \hat{\beta} &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\widehat{\text{Cov}}(x, y)}{\widehat{\text{Var}}(x)} \end{aligned}$$

Lineare Regression mit R. Diese Berechnungen wird natürlich R für uns machen. In R werden lineare Modelle mit `lm()` (Linear Model) angepasst. `lm()` hat folgende Argumente:

```
args(lm)
## function (formula, data, subset, weights, na.action, method = "qr",
##           model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
##           contrasts = NULL, offset, ...)
## NULL
```

`fm<-lm(formula,data,...)`

- `lm()` : model-fitting function for linear models.
- `formula` : symbolic description of the model.
- `data` : data set containing the variables from the formula.
- `...` : further arguments, e.g., control parameters
- `fm` : name of the fitted-model object of class `lm`.

Folgender Code macht eine Regression von Reaktionszeit (abhängige Variable, vor dem ~ Zeichen) auf Alkoholkonzentration (unabhängige Variable, nach dem ~ Zeichen). Wir hatten diese Daten schon, aber hier generieren wir sie noch einmal:

```
A <- c(0, 0.2, 0.5, 0.7, 1, 1.4, 1.8, 2.25, 2.5)
R <- c(554, 581, 589, 628, 623, 687, 692, 734, 812)
ARdata <- data.frame(Alkohol = A, Reaktionszeit = R)
```

Angepasste Modelle werden in R auch als Objekte abgespeichert, hier im Objekt `myfirstmodel`:

```
myfirstmodel <- lm(R ~ A, data = ARdata)
myfirstmodel

##
## Call:
## lm(formula = R ~ A, data = ARdata)
##
```

```
## Coefficients:
## (Intercept)          A
##      551.7       90.3
```

Von Hand könnten wir die Steigung auch berechnen mit

```
cov(R, A)/var(A)

## [1] 90.3
```

Im Output – d.h. in der Default `print`-Methode (`print(myfirstmodel)` oder kurz `myfirstmodel`) – steht zuerst nur das wichtigste, nämlich die beiden Punktschätzungen der Parameter. `Intercept` steht für $\hat{\alpha}$, unter `A` steht der geschätzte Parameter für den Effekt von Alkoholkonzentration, also $\hat{\beta}$. In einem Modellobjekt wie `myfirstmodel` ist aber noch viel mehr Information enthalten, die wir dann später brauchen. Im Moment benötigen wir aber diese Information (noch) nicht:

```
str(myfirstmodel)
```

Mit `fitted()` kann man die angepassten Werte \hat{y}_i des Modells ausgeben lassen. Das Modellobjekt muss als Argument übergeben werden:

```
fitted(myfirstmodel)

##   1    2    3    4    5    6    7    8    9
## 552 570 597 615 642 678 714 755 777
```

Die Residuen des Modells erhält man mit `residuals()`:

```
residuals(myfirstmodel)

##   1    2    3    4    5    6    7    8    9
## 2.32 11.26 -7.84 13.09 -19.01  8.86 -22.27 -20.92 34.50
```

Das sind natürlich die $y_i - \hat{y}_i$:

```
R = fitted(myfirstmodel)

##   1    2    3    4    5    6    7    8    9
## 2.32 11.26 -7.84 13.09 -19.01  8.86 -22.27 -20.92 34.50
```

Schauen wir jetzt das Modell graphisch an, indem wir die Regressionsgerade zu den Daten im Streudiagramm von Reaktionszeit versus Alkoholkonzentration hinzufügen. In der Funktion `abline()` wird das Modellobjekt als Argument übergeben (`a,b` steht gerade für die Parameter des Modells). Mit `abline()` kann man aber auch vertikale und horizontale Linien in einer Graphik hinzufügen (`?abline`).

```
plot(A, R, xlab = "Alkoholkonzentration [Promille]", ylab = "Reaktionszeit [ms]")
abline(myfirstmodel) ## fügt die Regressionsgerade hinzu, siehe ?abline
points(A, fitted(myfirstmodel), pch = 3) ## plottet angepasste Werte als Kreuze
segments(x0 = A, y0 = fitted(myfirstmodel), x1 = A, y1 = R, lty = 2) ## Code für die Interessierten: plottet Residuen.
```

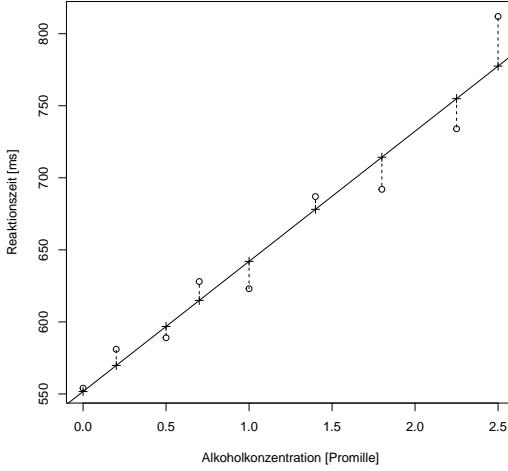


Abbildung 11.3: Regression von Reaktionszeit auf Alkoholkonzentration. Kreuze: Angepasste Werte, Punkte: Beobachtete Werte, Gestrichelt: Residuen.

In unserem Beispiel ist also $\hat{\beta} = 90.329$ die erwartete Zunahme an Reaktionszeit in Millisekunden pro Promille Zunahme in der Alkoholkonzentration. Mit den geschätzten Parametern kann man für *neue Werte* der Eingangsgröße die Zielgröße *vorhersagen*, der vorhergesagte Wert für ein x_{neu} (Alkoholkonzentration) wäre dann

$$\hat{Y}_{neu} = \hat{\alpha} + \hat{\beta}x_{neu} = 551.678 + 90.329x_{neu}.$$

Mit `predict()` werden Vorhersagen gemäss Modell (also gemäss den geschätzten Parametern) gemacht. Für die beobachteten x -Werte sind die Punktvorhersagen natürlich gerade die angepassten Werte:

```
predict(myfirstmodel)

##   1   2   3   4   5   6   7   8   9
## 552 570 597 615 642 678 714 755 777
```

Mit `predict()` kann man aber vor allem für *neue* nicht beobachtete Werte auf den Prädiktoren die Outcomes gemäss Modell vorhersagen, z.B. für *neue* Alkoholkonzentrationen, mit dem Argument `newdata=data.frame()`. Wir wollen z.B. eine Vorhersage gemäss Modell für neue Prädiktorwerte 0.3, 0.9 und 1.6:

```

new <- c(0.3, 0.9, 1.6)
pred.frame <- data.frame(A = new)
pred <- predict(myfirstmodel, newdata = pred.frame)
data.frame(new, pred)

##   new pred
## 1 0.3 579
## 2 0.9 633
## 3 1.6 696

```

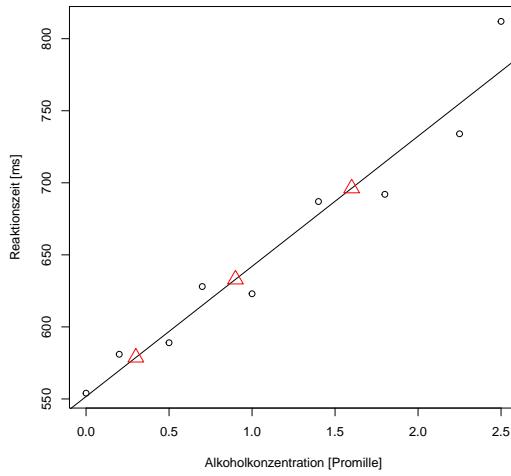


Abbildung 11.4: Regression von Reaktionszeit auf Alkoholkonzentration. Rote Dreiecke: Punktvorhersagen für neue Werte auf Alkoholkonzentration.

Unsicherheit der Vorhersage*. Später brauchen wir dann auch die *Unsicherheit* der Vorhersage einer neuen Beobachtung:

```

predict(myfirstmodel, newdata = pred.frame, interval = "prediction")

##   fit lwr upr
## 1 579 525 632
## 2 633 582 684
## 3 696 644 748

```

Überanpassung. Der *angepasste* Wert \hat{y}_i ist i.A. nicht identisch mit dem beobachteten Wert y_i , ausser unser Modell hat gleich viele Parameter wie Daten. Es ist immer möglich, eine komplexe Kurve genau durch die 9 Datenpunkte zu zeichnen, dieses *komplexere* Modell hätte dann aber gleich viele Parameter wie Daten. Abbildung 11.5 zeigt das

lineare Modell und zwei komplexere Modelle mit 4 respektive 9 Parametern (Polynome 3. und 8. Grades). Letzteres hat gleiche viele Parameter wie Daten und passt dann perfekt zu *diesen* Daten, ist aber für die *Vorhersage* schlecht. Solche Modelle haben eine *Überanpassung (overfit)*. Einfache Modelle sind oft besser, auch wenn sie nicht perfekt zu den verfügbaren Daten “passen”. Wir werden diese Problematik in einem späteren Kapitel vertiefen.

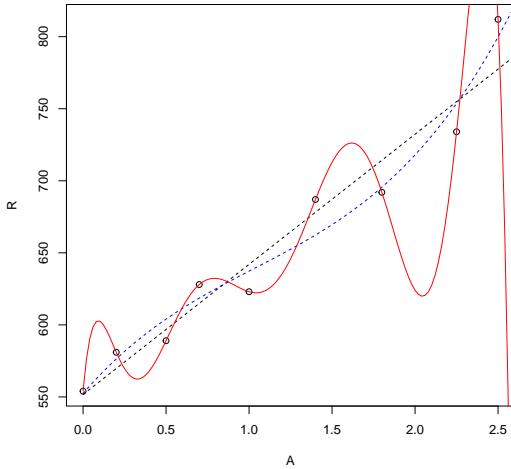


Abbildung 11.5: Einfaches lineares Modell (schwarz, 2 Parameter), Polynom 3. Grades (blau, 4 Parameter) und Polynom 8. Grades (rot, $n = 9$ Parameter)

Quadratsummen. Zurück zum linearen Modell. Wir betrachten die Abbildung 11.6. $y_i - \hat{y}_i$ nannten wir oben das i -te Residuum. Wir haben n Daten, n Residuen und n angepasste Werte, wir nennen dann

- $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ die **Residuen-Quadratsumme**. (Das wäre in Abbildung 11.6 die Summe aller quadrierten Längen der roten Strichlinien).
- $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ als der durch das Modell **erklärte (explained) Quadratsumme**. (analog, punktierte, blaue Linien).
- $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ die **totale Quadratsumme** (analog, grüne Strichpunktlinien)

Diese Quadratsummen werden später wichtig werden im Zusammenhang mit *Varianzanalysen* (ANOVA). Wir werden sie auch geometrisch deuten. Wir werden beim Allgemeinen Linearen Modell sehen, dass man die totale Quadratsumme gemäss PYTHAGORAS zerlegen kann, also

$$TSS = ESS + RSS. \quad (11.2.8)$$

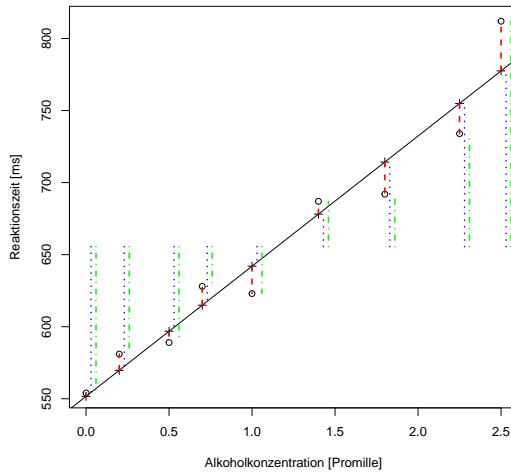


Abbildung 11.6: Residuen: $y_i - \hat{y}_i$ (rot gestrichelt). Durch Modell erklärt: $\hat{y}_i - \bar{y}$ (blau gepunktet). Total: $y_i - \bar{y}$ (grün strichpunkt)

Die Korrelation zwischen den Beobachtungen y_i und den angepassten Werten \hat{y}_i nennt man *multiple Korrelation*. Die quadrierte multiple Korrelation R^2 wird auch *Bestimmtheitsmaß* oder *Varianzaufklärung* genannt, da sie den Anteil der Streuung der y -Werte bestimmt, der durch die Regression *bestimmt* wird. R^2 wird berechnet als Verhältnis von erklärter zu totaler Streuung,

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS}. \quad (11.2.9)$$

Die angepassten Werte haben wir bereits mit `fitted()` extrahiert (die Kreuze in Abbildung 11.6)

Die durch das Modell erklärte Quadratsumme ist dann

```
sum((fitted(myfirstmodel) - mean(R))^2)

## [1] 52138
```

die Residuen-Quadratsumme ist

```
sum((R - fitted(myfirstmodel))^2)

## [1] 2929
```

und die totale Quadratsumme ist

```
sum((R - mean(R))^2)
```

```
## [1] 55066
```

Diese Quadratsummen werden wir mit der `anova()`-Funktion (Analysis of Variance, Varianzanalyse) direkt von R bekommen. Wir kommen darauf im allgemeinen Fall zurück. Im folgenden Output interessieren im Moment nur die Quadratsummen (`Sum Sq`):

```
anova(myfirstmodel)
```

```
## Analysis of Variance Table
##
## Response: R
##           Df Sum Sq Mean Sq F value Pr(>F)
## A          1 52138   52138     125  1e-05
## Residuals  7  2929     418
```

Also ist $R^2 = \frac{52137.69}{52137.69+2928.532} = 0.947$.

Unser lineares Modell (mit zwei Parametern α und β) erklärt also 94.6% der Streuung der beobachteten Werte. Man kann zeigen, dass bei der einfachen Regression (nur eine erklärende Variable) R^2 identisch ist dem Korrelationskoeffizienten im Quadrat:

```
cor(A, R)^2
```

```
## [1] 0.947
```

Einfache Regression und Kausalität. Wie bei der Korrelation gilt auch für die einfache Regression: Die Tatsache, dass es einen linearen Zusammenhang gibt, bedeutet i.A. nicht, dass eine Veränderung auf der Eingangsgrösse auch eine *Ursache* ist für eine Veränderung auf der Zielgrösse. Es könnte sein, dass die wahre Ursache gar keine Eingangsgrösse im Modell war. Das führt uns später zum allgemeinen Fall der *multiplen Regression*. Dort werden zusätzliche potentielle Ursachen, oder Störgrössen (*Confounder*) als Eingangsgrössen ins Modell aufgenommen. Mit dem einfachen linearen Modell haben wir dafür die Grundlage geschaffen.

Tests und Konfidenzintervalle. Für den allgemeinen Fall werden wir später die Standardfehler der geschätzten Parameter einführen, über die man dann Konfidenzintervalle konstruieren kann und Hypothesen testen kann. Mit `summary()` kommen wir auf diese Grössen:

```
summary(myfirstmodel)
```

```
##  
## Call:  
## lm(formula = R ~ A, data = ARdata)  
##  
## Residuals:  
##   Min     1Q Median     3Q    Max  
## -22.27 -19.01   2.32  11.26  34.50  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 551.68     11.54   47.8  4.6e-10  
## A           90.33      8.09   11.2  1.0e-05  
##  
## Residual standard error: 20.5 on 7 degrees of freedom  
## Multiple R-squared:  0.947, Adjusted R-squared:  0.939  
## F-statistic: 125 on 1 and 7 DF, p-value: 0.0000103
```

Die Punktschätzungen bezüglich der Parameter haben wir bereits gesehen. Die Standardfehler führen wir weiter unten mit der multiplen Regression ein, ebenso die t -Statistiken und die zugehörigen p -Werte. R^2 haben wir oben eingeführt. Auch die Bedeutung der F -Tests werden wir später einführen.

Bevor wir zum Allgemeinen Linearen Modell schreiten, müssen wir einen kurzen Ausflug machen in die *Matrixalgebra*.

Kapitel 12

Kurze Matrixalgebra

Lineare Modelle stellen einen Grossteil von Analysen im Bereich der Gesundheitswissenschaften dar. Für ein gutes Verständnis von linearen Modellen braucht es ein bisschen *Lineare Algebra*. Insbesondere wollen wir hier die wichtigsten Aspekte des Rechnens mit *Matrizen* anschauen.

Eine Matrix A ist ein Objekt mit m Zeilen und n Spalten mit Elementen a_{ij} , ($1 \leq i \leq m$, $1 \leq j \leq n$) :

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

Der erste Index steht für die Zeile und der zweite Index steht für die Spalte.

Wann sind zwei solche Objekte gleich? Wie können wir solche Objekte kombinieren? Dazu folgende Definition:

Definition 1

Zwei Matrizen A und B sind *gleich*, $A = B$, wenn $a_{ij} = b_{ij}$, ($1 \leq i \leq m$, $1 \leq j \leq n$).

Wir wissen, wie man Zahlen addiert, subtrahiert, multipliziert und dividiert (wenn möglich). Wie geht das mit Matrizen? Das ist möglich, aber zuerst schauen wir uns eine andere Multiplikation an.

Definition 2

Seien A und B mit Elementen a_{ij} und b_{ij} beide $m \times n$ Matrizen. Die *Summe*, $A + B$, und die *Differenz*, $A - B$, sind die Matrizen $(a_{ij} + b_{ij})$ und $(a_{ij} - b_{ij})$. Die *skalare Multiplikation* für jedes $r \in \mathbb{R}$, rA ist die Matrix (ra_{ij}) .

Es gilt Kommutativität: $rA = Ar$.

Beispiel

Seien $A = \begin{pmatrix} 2 & 3 \\ -1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} -1 & 2 \\ 6 & -2 \end{pmatrix}$, und $C = \begin{pmatrix} 1 & 2 & 3 \\ -1 & -2 & -3 \end{pmatrix}$.

```
A <- matrix(c(2, -1, 3, 2), nrow = 2) # matrix() füllt defaultmäßig zuerst die Spalten!
B <- matrix(c(-1, 6, 2, -2), nrow = 2)
C <- matrix(c(1, -1, 2, -2, 3, -3), nrow = 2)
```

$$1. A + B.$$

Da A und B beide 2×2 Matrizen, können wir sie addieren.

$$A + B = \begin{pmatrix} 2 & 3 \\ -1 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 2 \\ 6 & -2 \end{pmatrix} = \begin{pmatrix} 2 + (-1) & 3 + 2 \\ -1 + 6 & 2 + (-2) \end{pmatrix} = \begin{pmatrix} 1 & 5 \\ 5 & 0 \end{pmatrix}.$$

```
A + B
```

```
##      [,1] [,2]
## [1,]    1    5
## [2,]    5    0
```

$$2. B + C.$$

Nicht möglich

$$3. 4C.$$

Wir multiplizieren jeden Eintrag von C mit 4:

$$4C = 4 \begin{pmatrix} 1 & 2 & 3 \\ -1 & -2 & -3 \end{pmatrix} = \begin{pmatrix} 4(1) & 4(2) & 4(3) \\ 4(-1) & 4(-2) & 4(-3) \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 \\ -4 & -8 & -12 \end{pmatrix}.$$

```
4 * C
```

```
##      [,1] [,2] [,3]
## [1,]    4    8   12
## [2,]   -4   -8  -12
```

Eigenschaften

Seien A , B , und C alle $m \times n$ Matrizen und $r, s \in \mathbb{R}$.

1. $A + B = B + A$
2. $A + (B + C) = (A + B) + C$
3. $r(A + B) = rA + rB$
4. $(r + s)A = rA + sA$
5. $(rs)A = r(sA)$

Nun zur Multiplikation. Achtung: Wir werden **NIE** einfach die Einträge multiplizieren.

Definition 3

Sie $\mathbf{a} = (a_1 \ a_2 \ \cdots \ a_p)$ ein Zeilenvektor mit p Einträgen und $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$ ein
Kolonnenvektor mit p Einträgen. Das *Skalarprodukt* $\mathbf{ab} = (a_1 \ a_2 \ \cdots \ a_p) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$,
ist die reelle Zahl $\mathbf{ab} = a_1b_1 + a_2b_2 + \cdots + a_pb_p$.

Beispiel

$$1. (2 \ 3 \ 4) \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix} = 2(3) + 3(4) + 4(5) = 38.$$

In R werden Vektoren oder Matrizen multipliziert mit `%*%`

```
a <- matrix(c(2, 3, 4), nrow = 1)
b <- matrix(c(3, 4, 5), nrow = 3)
a

##      [,1] [,2] [,3]
## [1,]    2    3    4

b

##      [,1]
## [1,]    3
## [2,]    4
## [3,]    5

a %*% b ## Das ist ein Skalarprodukt (eine Zahl!!)
##      [,1]
## [1,]   38
```

$$2. (-1 \ 2 \ -2 \ 3) \begin{pmatrix} 2 \\ -2 \\ -1 \\ 2 \end{pmatrix} = -1(2) + 2(-2) + (-2)(-1) + 3(2) = 2.$$

```
a <- matrix(c(-1, 2, -2, 3), nrow = 1)
b <- matrix(c(2, -2, -1, 2), nrow = 4)
a %*% b ## Das ist ein Skalarprodukt (eine Zahl!!)
```

Jetzt multiplizieren wir eine Matrix mit einem Kolonnenvektor. Damit das geht, **muss die Anzahl Elemente in einer Zeile gleich der Anzahl Elemente in der Kolonne sein**. Dann multiplizieren wir jede Zeile der Matrix mit der Kolonne.

Definition 4

Sei A eine $m \times p$ Matrix und \mathbf{b} ein $p \times 1$ Kolonnenvektor. Das *Produkt* Ab ist ein $m \times 1$ Kolonnenvektor mit dem Produkt der i ten Zeile von A und \mathbf{b} als i tem Eintrag ($1 \leq i \leq m$).

Beispiel

$$1. \left(\begin{array}{ccc} 1 & 2 & 3 \\ -2 & 1 & 2 \end{array} \right) \left(\begin{array}{c} 1 \\ 2 \\ -3 \end{array} \right) = \left(\begin{array}{c} 1(1) + 2(2) + 3(-3) \\ -2(1) + 1(2) + 2(-3) \end{array} \right) = \left(\begin{array}{c} -4 \\ -6 \end{array} \right).$$

```
A <- matrix(c(1, -2, 2, 1, 3, 2), nrow = 2)
B <- matrix(c(1, 2, -3), nrow = 3)
A %*% B
```

$$2. \left(\begin{array}{cc} 2 & -2 \\ 0 & 3 \\ -1 & 4 \end{array} \right) \left(\begin{array}{c} 5 \\ -1 \end{array} \right) = \left(\begin{array}{c} 2(5) + (-2)(-1) \\ 0(5) + 3(-1) \\ -1(5) + 4(-1) \end{array} \right) = \left(\begin{array}{c} 12 \\ -3 \\ -9 \end{array} \right).$$

```
A <- matrix(c(2, 0, -1, -2, 3, 4), nrow = 3)
B <- matrix(c(5, -1), nrow = 2)
A %*% B
```

Nun kommen wir zur Erweiterung der Multiplikation auf beliebige Matrizen. Eine Matrix besteht aus mehreren Kolonnenvektoren. **Die Anzahl Kolonnen in der ersten Matrix muss gleich der Anzahl Zeilen der zweiten Matrix sein.**

Definition 5

Sei A eine $m \times p$ Matrix und B eine $p \times n$ Matrix. Das *Produkt* AB ist die $m \times n$ Matrix mit dem Produkt der i ten Zeile von A und der j ten Kolonne von B als ij ten Eintrag ($1 \leq i \leq m, 1 \leq j \leq n$)

Beispiel

Sei $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix}$, $C = \begin{pmatrix} 2 & 2 & 9 \\ -1 & 0 & 8 \end{pmatrix}$, und $D = \begin{pmatrix} 1 & 2 & 3 \\ 5 & 2 & 3 \end{pmatrix}$.

```
A <- matrix(c(1, 3, 2, 4), nrow = 2)
B <- matrix(c(4, -2, -3, 1), nrow = 2)
C <- matrix(c(2, -1, 2, 0, 9, 8), nrow = 2)
D <- matrix(c(1, 5, 2, 2, 3, 3), nrow = 2)
```

1. AB .

$$\begin{aligned} AB &= \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1(4) + 2(-2) & 1(-3) + 2(1) \\ 3(4) + 4(-2) & 3(-3) + 4(1) \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 4 & -5 \end{pmatrix}. \end{aligned}$$

```
A %*% B
##      [,1] [,2]
## [1,]    0   -1
## [2,]    4   -5
```

2. BA .

$$\begin{aligned} BA &= \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 4(1) + (-3)(3) & 4(2) + (-3)(4) \\ -2(1) + 1(3) & -2(2) + 1(4) \end{pmatrix} = \begin{pmatrix} -5 & -4 \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

```
B %*% A
##      [,1] [,2]
## [1,]   -5   -4
## [2,]    1    0
```

$AB \neq BA$! Matrixmultiplikation ist nicht kommutativ

3. CD . Geht nicht.

4. BC .

$$\begin{aligned} BC &= \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 & 9 \\ -1 & 0 & 8 \end{pmatrix} \\ &= \begin{pmatrix} 4(2) + (-3)(-1) & 4(2) + (-3)(0) & 4(9) + (-3)(8) \\ -2(2) + 1(-1) & -2(2) + 1(0) & -2(9) + 1(8) \end{pmatrix} \\ &= \begin{pmatrix} 11 & 8 & 12 \\ -5 & -4 & -10 \end{pmatrix} \end{aligned}$$

```
B %*% C
##      [,1] [,2] [,3]
## [1,]   11    8   12
## [2,]   -5   -4  -10
```

5. CB . Geht nicht.

Weitere Eigenschaften

Seien A , B , und C Matrizen mit angemessener Grösse und $r \in \mathbb{R}$. Dann

1. $A(BC) = (AB)C$
2. $(rA)B = r(AB) = A(rB)$
3. $A(B + C) = AB + AC$
4. $(A + B)C = AC + BC$
5. Es gibt eine $n \times n$ Matrix I , so dass für eine $n \times n$ Matrix M : $IM = MI = M$.

Die $n \times n$ Matrix I heisst *Einheitsmatrix*.

$$I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Lineare Gleichungssysteme. Aus der Schule kennen wir lineare Gleichungssysteme der Form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

Wir können das jetzt kurz und knackig schreiben als

$$Ax = b$$

mit

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{und} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Die Matrixgleichung $Ax = b$ stellt ein System von linearen Gleichungen dar.

Beispiel

Betrachte das lineare Gleichungssystem

$$\begin{aligned} 2x - y &= 0 \\ x + z &= 4. \\ x + 2y - 2z &= -1 \end{aligned}$$

1. Wir schreiben das als $A\mathbf{x} = \mathbf{b}$. A ist die Matrix mit den Koeffizienten, \mathbf{x} ist der Kolonnenvektor mit den Variablen, und \mathbf{b} ist der Kolonnenvektor mit Konstanten. Wir haben dann

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 2 & -2 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \text{ und } \mathbf{b} = \begin{pmatrix} 0 \\ 4 \\ -1 \end{pmatrix}.$$

Die Gleichung ist

$$\begin{pmatrix} 2 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ -1 \end{pmatrix}.$$

Für die Lösung müssen wir zuerst den Begriff der *Inversen* einführen. Wir wissen, dass es für jede Zahl $x \neq 0$ eine inverse Zahl $x^{-1} = 1/x$ gibt mit $xx^{-1} = 1$. Gibt es etwas Ähnliches für Matrizen? Gibt es für jede $n \times n$ Matrix eine Matrix B , so dass $AB = BA = I$? Die Antwort ist nein, nicht immer.

Definition 6

Eine $n \times n$ Matrix A ist *umkehrbar*, wenn es eine $n \times n$ Matrix B gibt mit $AB = BA = I$. Die Matrix B ist dann *eine Inverse of A*.

Beispiel

Sei $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ und $B = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$.

```
A <- matrix(c(2, 1, 1, 1), nrow = 2)
B <- matrix(c(1, -1, -1, 2), nrow = 2)
```

$$\begin{aligned} AB &= \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2(1) + 1(-1) & 2(-1) + 1(2) \\ 1(1) + 1(-1) & 1(-1) + 1(2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{und} \\ BA &= \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1(2) - 1(1) & 1(1) - 1(1) \\ -1(2) + 2(1) & -1(1) + 2(1) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

```
A %*% B

##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1

B %*% A

##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

Eigenschaften von invertierbaren Matrizen

1. Wenn eine $n \times n$ Matrix invertierbar ist, dann ist diese eindeutig.
2. Wenn A eine $n \times n$ invertierbare Matrix, dann ist die Inverse der Inversen von A gleich A .
3. Wenn A und B invertierbare $n \times n$ Matrizen, dann $(AB)^{-1} = B^{-1}A^{-1}$.

Wir nennen nun *die Inverse* einer quadratischen Matrix A kurz A^{-1} , mit der Eigenschaft $AA^{-1} = A^{-1}A = I$.

Theorem 1

Sei A eine $n \times n$ Matrix. Dann sind folgende Aussagen äquivalent:

1. A ist umkehrbar.
2. Für jedes $\mathbf{b} \in \mathbb{R}^n$ hat die Gleichung $A\mathbf{x} = \mathbf{b}$ exakt eine Lösung.

Theorem 2

Seien A und B $n \times n$ Matrizen und I die $n \times n$ Einheitsmatrix. Wenn $AB = I$, dann sind A und B invertierbar und $A^{-1} = B$.

Zurück zu unserem Gleichungssystem $A\mathbf{x} = \mathbf{b}$. Wenn A invertierbar ist (man sagt auch *regulär*), dann

$$\mathbf{x} = A^{-1}\mathbf{b}$$

```
A <- matrix(c(2, 1, 1, -1, 0, 2, 0, 1, -2), nrow = 3)
b <- c(0, 4, -1)
A

##      [,1] [,2] [,3]
```

```
## [1,]    2   -1    0
## [2,]    1    0    1
## [3,]    1    2   -2

b

## [1]  0  4 -1
```

Die Lösung wird hier die folgende sein. Man bekommt sie mit `solve()`:

```
Ainv <- solve(A)
x <- Ainv %*% b
x

##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

Kontrolle:

```
A %*% x
```

Definition 7

Sei $A = (a_{ij})$ eine $m \times n$ Matrix. Die *Transponierte* von A , notiert mit A^T , ist die Matrix, deren i te Kolonne die i te Zeile von A ist, oder deren j te Zeile die j te Kolonne von A ist. A^T ist eine $n \times m$ Matrix. Wir schreiben $A^T = (a_{ji}^T)$ mit $a_{ji}^T = a_{ij}$.

Der j ite-Eintrag von A^T ist der ij te Eintrag von A .

Beispiel

$$1. \quad A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}.$$

$$A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

Die R-Funktion zum Transponieren ist `t()`

```
A <- matrix(c(1, 2, 3, 4, 5, 6), byrow = TRUE, nrow = 2)
t(A)
```

$$2. \quad B = \begin{pmatrix} -1 & 0 & 6 \\ -4 & 1 & 9 \\ 2 & 3 & 0 \end{pmatrix}.$$

$$B^T = \begin{pmatrix} -1 & -4 & 2 \\ 0 & 1 & 3 \\ 6 & 9 & 0 \end{pmatrix}.$$

```
B <- matrix(c(-1, -4, 2, 0, 1, 3, 6, 9, 0), nrow = 3)
t(B)
```

Eigenschaften der Transponierten

Seien A und B Matrizen mit angemessener Grösse und $r \in \mathbb{R}$. Dann

1. $(A^T)^T = A$.
2. $(A + B)^T = A^T + B^T$.
3. $(rA)^T = rA^T$.
4. $(AB)^T = B^T A^T$.

```
t(A %*% B)
```

```
##      [,1] [,2]
## [1,]    -3   -12
## [2,]    11    23
## [3,]    24    69
```

```
t(B) %*% t(A)
```

```
##      [,1] [,2]
## [1,]    -3   -12
## [2,]    11    23
## [3,]    24    69
```

Kapitel 13

Das Allgemeine Lineare Modell (LM)

Das Allgemeine Lineare Modell (LM) ist ein breit angewandtes statistisches Modell im den Gesundheitswissenschaften. Es ist eines der einfachsten Beispiele, um wichtige Aspekte des *statistischen Modellierens* zu demonstrieren. Das Allgemeine Lineare Modell stellt einen allgemeinen Rahmen dar für eine grosse Menge von Modellen, deren gemeinsames Ziel ist:

- Erklären oder vorhersagen
- einer *quantitativen abhängigen* Variablen
- durch eine Menge von unabhängigen Variablen, die *kategorial oder kontinuierlich* sein können.

LM umfassen *t*-Tests, die einfache Regression, klassische Varianzanalysen, Kovarianzanalysen und multiple Regressionen. Wir werden auch die klassischen Varianzanalysen als Regressionen mit kategorialen Eingangsgrössen (d.h. mit Indikatorvariablen als Eingangsgrössen) behandeln.

Gegeben ist also eine quantitative abhängige Variable, die Zielgröße. Bis auf Zufallsfehler ist diese Variable eine *lineare Funktion* von mehreren Eingangsgrössen (oder Prädiktoren, Kovariablen). Das Ziel ist dann, die Parameter zu schätzen, ihre Relevanz zu studieren und die Fehlervarianz zu schätzen.

Mit linearen Modellen, die nicht nur einen, sondern mehrere Eingangsgrössen (Prädiktoren oder Kovariablen) in das Modell einbeziehen, versucht man den Effekt jedes einzelnen Prädiktors auf die Zielgröße zu quantifizieren. Wir möchten dann den Effekt jeder einzelnen Eingangsgröße für den Effekt von anderen Kovariablen *korrigieren*.

In den nächsten Abschnitten betrachten wir zuerst das Modell und befassen uns dann mit der Schätzung der Parameter. In einem nächsten Abschnitt betrachten wir Erwartungswerte und Varianz der Schätzer. Dann betrachten wir die Verteilung der Schätzer. Damit sind wir dann in der Lage, statistische Hypothesen zu testen und Intervallschätzungen bezüglich den Parametern zu vollziehen. Wir führen dann die Varianzanalyse ein, um gemeinsame Hypothesen bezüglich Parametern zu testen. Die Theorie wird mit R anhand eines Beispiels nachvollzogen. Am Schluss des Kapitels betrachten wir dann die Residuenanalyse als Mittel der Überprüfung der Annahmen.

13.1 Das Modell

Modellformel.

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n. \quad (13.1.1)$$

Wir nehmen meistens an, dass die Fehler unabhängig und gleichverteilt sind, ϵ_i i.i.d., dass die Fehler Erwartungswert 0 haben, $E(\epsilon_i) = 0$, und dass die Varianz der Fehler konstant ist, $\text{Var}(\epsilon_i) = \sigma^2$.

Notation. Im Allgemeinen werden beobachtbare Zufallsvariable mit lateinischen Grossbuchstaben notiert, ebenso Matrizen. Realisierte Beobachtungen werden mit kleinen lateinischen Buchstaben notiert. Vektoren werden fett notiert, Skalare und Matrizen normal. Unbekannte Parameter und nicht beobachtbare Grössen werden griechisch notiert.

- $\mathbf{Y} = \{Y_i; i = 1, \dots, n\}$: n -Vektor der Zielgrössen
- $\mathbf{x}_j = \{x_{ij}; i = 1, \dots, n\}$: n -Vektor des j ten Prädiktors
- $\mathbf{x}_i = \{x_{ij}; j = 1, \dots, p\}$: p -Vektor der Prädiktoren für die i te Beobachtung
- $\boldsymbol{\beta} = \{\beta_j; j = 1, \dots, p\}$: p -Vektor der unbekannten Parameter
- n ist die Stichprobengrösse, p ist die Anzahl der Prädiktoren

Die β_j (und σ^2) werden *Koeffizienten* oder *Parameter* des Modells genannt. Die Parameter sind unbekannt und die Fehler ϵ_i sind nicht beobachtbar. Die Y_i und die x_{ij} sind beobachtbar.

Modell mit Vektornotation. Durch Notation mit Vektoren können wir das Modell kürzer schreiben. Mit dem (transponierten) Vektor der Eingangsgrössen $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ und dem Parametervektor $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ schreiben wir dann kurz mit dem Skalarprodukt $\mathbf{x}_i^T \boldsymbol{\beta}$

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n. \quad (13.1.2)$$

Modell mit Matrixnotation. In Matrixform können wir diese n Gleichungen mit p Parametern kompakt zusammenfassen in der Form

$$\mathbf{Y} = \mathbf{X} \times \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (13.1.3)$$

Man nennt die Matrix X die **Design-Matrix**.

Ausgeschrieben ist dies

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \quad (13.1.4)$$

Die erste Eingangsgrösse ist meistens eine Konstante, $x_{i1} \equiv 1$, wir haben dann wie bei der einfachen Regression ein *Intercept* im Modell (das jetzt β_1 statt α heisst).

$$Y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_j x_{ij} + \dots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n. \quad (13.1.5)$$

Wir nehmen an, dass die Stichprobengrösse n grösser ist als die Anzahl Parameter p , $n > p$.

Zudem ist eine Bedingung, dass die p Spalten von X *linear unabhängig* sind. Das bedeutet, dass ein Spaltenvektor von X (Prädiktor) nicht als Linearkombination von anderen Spaltenvektoren geschrieben werden kann. Die Folge wäre, dass das unten eingeführte Optimierungsproblem dann nicht gelöst werden kann (weil dann die $p \times p$ Matrix $X^T X$ nicht invertierbar sein wird).

Stochastische Modelle. Die Fehler ϵ_i sind Zufallsvariable und damit sind auch die Y_i Zufallsvariable. Die Fehler $\epsilon_1, \dots, \epsilon_n$ bilden dann wieder eine i.i.d. Zufallstichprobe. Die stochastische Natur der Fehler haben ihren Ursprung in verschiedenen Quellen. Alle Arten von Messfehlern oder Unmöglichkeit der Erfassung von nicht-systematischen Effekten werden in dieser Zufallsvariable mit Erwartungswert 0 subsumiert. Die Prädiktorvariablen x_{ij} werden *nicht* als zufällig betrachtet.

Die stochastische Natur erlaubt es uns, Unsicherheit zu quantifizieren, "Signifikanz" von Prädiktoren zu bestimmen und einen guten Kompromiss zwischen Modellkomplexität (Anzahl Parameter) und Modellanpassung zu finden.

Die beobachteten Zielgrössen in den Daten werden als Realisierungen von Zufallsvariablen Y_1, \dots, Y_n betrachtet; die x_{ij} sind nicht zufällig und gleich den beobachteten Prädiktoren in den Daten.

Linearität. Das Modell heisst *linear*, weil es linear *in den Koeffizienten* β_j ist; der Effekt eines Prädiktors x_j ist also über den ganzen Bereich von x_j konstant. Ein Modell mit

quadratischen und kubischen Eingangsgrößen wie

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \epsilon_i \quad (13.1.6)$$

ist in diesem Sinne auch linear, obwohl wir da eine quadratische und kubische Funktion in den x haben! Der Effekt von x , von x^2 und von x^3 ist über den Wertebereich dieser Größen konstant, nämlich β_1, β_2 und β_3 . Ein Beispiel für ein *nicht-lineares* Modell wäre $Y_i = \beta_1 x_i^{\beta_2} + \epsilon_i$. Hier ist der Effekt von x nicht konstant über den Wertebereich von x . Nichtlineare Modelle in der Wissenschaft sind wesentlich komplizierter und werden hier nicht behandelt.

13.2 Spezialfälle

Einfache Regression. Die einfache Regression haben wir im vorigen Kapitel behandelt.

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \quad (13.2.1)$$

Einstichproben t-Test. Dieses Modell hat einen Parameter β .

$$Y_i = \beta + \epsilon_i$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \quad (13.2.2)$$

Ein One-sample- t -Test ist äquivalent mit einem linearen Modell, einer Regression von Y auf ein Intercept.

Zwei-Stichproben t -Test. Dieses Modell hat zwei Parameter, β_1 und β_2 .

$$Y_i = \beta_1 + \beta_2 I(x_i) + \epsilon_i, \quad I(x_i) = \begin{cases} 1, & \text{wenn } x_i = 1 \\ 0, & \text{wenn } x_i = 0 \end{cases}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \quad (13.2.3)$$

Bei kategorialen Eingangsgrößen braucht es gleich viele *Dummy*-Variablen in der Design-Matrix X wie Anzahl Kategorien der entsprechenden Eingangsgrößen. In R ist defaultmäßig die sogenannte *Effekt*-Parameterisierung eingestellt: Der erste Parameter steht für den Erwartungswert in der Referenzgruppe (Intercept). Das heisst, die erste Kolonne besteht aus Einsen. Der zweite Parameter steht für den *erwarteten Unterschied* in der Zielgröße zwischen der zweiten und der ersten Gruppe. Daher stehen in der zweiten Kolonne von X die Einträge 0 oder 1, je nachdem ob die entsprechenden Beobachtung in der Referenzgruppe ist (0) oder in der zweiten Gruppe (1). Bei der Effekt-Parametrisierung ist dann $\beta_1 = \mu_1$ und $\beta_2 = \mu_2 - \mu_1$. Die Quantität von Interesse ist dann β_2 .

Ein Zwischengruppen- t -Test mit homogenen Varianzen ist dann äquivalent mit einem linearen Modell, einer Regression von Y auf eine zweiseitige Eingangsgröße.

Beispiel t -Test. Wir lesen die schon bekannten Daten ein der Datei

<https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/omega.csv>

Uns interessieren im Folgenden wieder nur die Variablen `weight` und `gender`. Wir wollen den Zusammenhang verstehen zwischen einem t -Test (Gewicht zwischen Männern und Frauen) und einem linearen Modell (Regression vom Gewicht auf das Geschlecht).

```
url <- "https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/omega.csv"
d.omega <- read.csv(url, stringsAsFactors = TRUE)
## str(d.omega)

by(d.omega$weight, d.omega$gender, psych::describe)

## d.omega$gender: female
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 21 80.1 14.1    81.6    80.6 17.4  54 105 50.8 -0.18   -1.09 3.08
## -----
## d.omega$gender: male
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 24 91.8 16.1    85.7    89.6 9.41 74.4 140 65.3  1.3    1.14 3.29
```

Klassisch als Two-sample *t*-Test:

```
t.test(x = d.omega$weight[d.omega$gender == "female"], y = d.omega$weight[d.omega$gender == "male"] ,  
       var.equal = TRUE)  
  
##  
## Two Sample t-test  
##  
## data: d.omega$weight[d.omega$gender == "female"] and d.omega$weight[d.omega$gender == "male"]  
## t = -3, df = 43, p-value = 0.01  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -20.8 -2.5  
## sample estimates:  
## mean of x mean of y  
## 80.1 91.8
```

Alternative: Mit **formula** und **data**-Argument:

```
t.test(weight ~ gender, data = d.omega, var.equal = TRUE)  
  
##  
## Two Sample t-test  
##  
## data: weight by gender  
## t = -3, df = 43, p-value = 0.01  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## -20.8 -2.5  
## sample estimates:  
## mean in group female mean in group male  
## 80.1 91.8
```

Jetzt als lineares Modell (**female** ist Referenz, da alphabetisch erste Kategorie !)

```
mod <- lm(weight ~ gender, d.omega)  
mod  
  
##  
## Call:  
## lm(formula = weight ~ gender, data = d.omega)  
##  
## Coefficients:  
## (Intercept)  gendermale  
##             80.1          11.7
```

Hier ist $(\text{Intercept}) = \beta_1 = \mu_1$ (Durchschnittsgewicht Frauen) und $\text{gendermale} = \beta_2 = \mu_2 - \mu_1$. (Erwarteter Unterschied im Gewicht von Männern relativ zu Frauen).

Allgemeine Ziele von Regressionsanalysen. Die Ziele von Regressionsanalysen sind:

- Man will eine gute *Anpassung* des Modells an die Daten, d.h. die “Reste” des Modells sollen klein sein. Das machen wir über die *Methode der kleinsten Quadrate*.

- Zudem will man “gute” Schätzungen der *Parameter* des Modells. Damit kann man die Veränderung der Zielgröße quantifizieren, wenn man Eingangsgrößen variiert.
- Ein weiteres Ziel ist die *Vorhersage* der abhängigen Variablen bei neuen Daten als Eingangsgrößen.
- Unsicherheit und Signifikanz der drei obigen Ziele. Das führt zu statistischen Tests und Konfidenzintervallen.
- Entwicklung eines guten Modells. In einem interaktiven Prozess werden Teile des Modells verändert um zu einem besseren Modell zu gelangen.

13.3 Die Methode der kleinsten Quadrate

Die Parameter des Modells (der p -dimensionale Parametervektor β) sind unbekannt und müssen aus den Daten geschätzt werden. Wir verallgemeinern hier die schon bekannte Methode der kleinsten Quadrate auf den mehrdimensionalen Fall. Das Prinzip ist analog zum grundlegenden Schätzproblem (Schätzen eines Erwartungswert im Einstichprobenfall, also in einem Modell ohne Eingangsgrößen) und auch analog zur einfachen Regression im letzten Kapitel.

Wir gehen von einem linearen Modell $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ aus. Der Kleinste-Quadrat-Schätzer für $\hat{\beta}$ ist

$$\hat{\beta} = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2. \quad (13.3.1)$$

$\|\cdot\|$ notiert die Euklidische Norm in \mathcal{R}^n . Das ist die **Länge** des Vektors $\mathbf{Y} - \mathbf{X}\beta$. $\|\mathbf{Y} - \mathbf{X}\beta\|^2$ ist dann eine **quadrierte Länge**.

$\|\mathbf{Y} - \mathbf{X}\beta\|^2$ ist nichts Neues, es ist eine andere Schreibweise für die Quadratsumme $\sum_{i=1}^n (Y_i - x_i^T \beta)^2$, und hat nun eine geometrische Deutung einer quadrierten Länge.

Die Größe, die (13.3.1) minimiert, kann wie bei der einfachen Regression über Ableiten von $\|\mathbf{Y} - \mathbf{X}\beta\|^2$ und Nullsetzen bestimmt werden:

$$(-2)\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0} \quad (13.3.2)$$

Das führt uns zu den *Normalgleichungen*¹

$$\mathbf{X}^T \mathbf{X} \begin{matrix} p \times p \\ \hat{\beta} \end{matrix} = \mathbf{X}^T \mathbf{Y} \quad (13.3.3)$$

Das sind p lineare Gleichungen für p Unbekannte (Komponenten von $\hat{\beta}$). Wenn die Spalten der Design-Matrix X linear unabhängig sind, dann ist die $p \times p$ -Matrix $\mathbf{X}^T \mathbf{X}$

¹Weiter unten wird klar werden, wieso diese Gleichungen “Normal”-Gleichungen heißen.

umkehrbar. Der Kleinste-Quadrat-Schätzer ist dann

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y} \quad (13.3.4)$$

Kurzer Rückblick: Beim grundlegenden Einstichproben-Schätzproblem ohne Prädiktoren ist die Design-Matrix eine Kolonne von Einsen und (13.3.4) reduziert sich dann auf das bekannte $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$.

Aus den Residuen $r_i = Y_i - \mathbf{x}_i^T \hat{\beta}$ bekommt man schliesslich eine Schätzung für σ^2 ,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n r_i^2 \quad (13.3.5)$$

Beim grundlegenden Schätzproblem ohne Eingangsgrössen war die geschätzte Varianz $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n r_i^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Der Freiheitsgrad war dort $n-1$. Im allgemeinen Fall ist der Freiheitsgrad **Anzahl Beobachtungen minus Anzahl Parameter**, also $n-p$.

13.4 Annahmen

1. Das Modell ist korrekt: $E(\epsilon_i) = 0$ für alle i .
2. Homoskedastizität: $\text{Var}(\epsilon_i) = \sigma^2$ ist konstant für alle i .
3. Die Fehler sind unabhängig: $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$
4. Die Fehler haben eine gemeinsame Normalverteilung

Falls Annahme 2 verletzt ist, kann man die abhängige Variable transformieren oder *gewichtete Kleinste-Quadrat* brauchen (am Schluss des Kapitels für Interessierte). Wenn Annahme 3 verletzt ist, gibt es als Alternativen *verallgemeinerte kleinste Quadrate* und *Lineare Gemischte Modelle*. Diese werden wir später einführen. Falls Annahme 4 verletzt ist, braucht man robuste Methoden (oder n gross). Wenn Annahme 1 verletzt ist, dann braucht man ein anderes Modell. Wir kommen später im Rahmen der Residuenanalyse auf die Überprüfung der Annahmen zurück.

13.5 Die Geometrie der Regression

Wir wollen das Prinzip der multiplen Regression geometrisch darstellen. Wir werden darin Aspekte des Geometrieunterrichts aus der Schulzeit wiedererkennen. Eine Regression kann man als eine spezielle *Abbildung* in einem n -dimensionalen Raum \mathcal{R}^n betrachten. Man kann zeigen, dass die Abbildung von $\mathbf{Y} \in \mathcal{R}^n$ auf die angepassten

Werte $\hat{\mathbf{Y}}$ im p -dimensionalen Unterraum von \mathcal{R}^n eine *orthogonale Projektion* ist². Der Kleinstes-Quadrat-Schätzer $\hat{\beta}$ erfüllt dann, dass $\hat{\mathbf{Y}} = X\hat{\beta}$ der Punkt im von den Prädiktoren aufgespannten Raum ist, der *am nächsten* ist zu \mathbf{Y} .

Dies ist in der Abbildung 13.1 rechts illustriert mit einer Regression auf 2 Eingangsgrößen. Links ist eine klassische Darstellung im sogenannten Variablen-Raum (n Datenpunkte mit Koordinaten x_1, x_2 und y). Rechts sehen wir eine Darstellung im Beobachtungsraum (Vektoren $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}$ im n -dimensionalen Raum³). Wir sehen, dass $\hat{\mathbf{Y}}$ der “nächste” Ort in der durch die Prädiktoren \mathbf{x}_1 und \mathbf{x}_2 aufgespannten Ebene relativ zu \mathbf{y} ist.

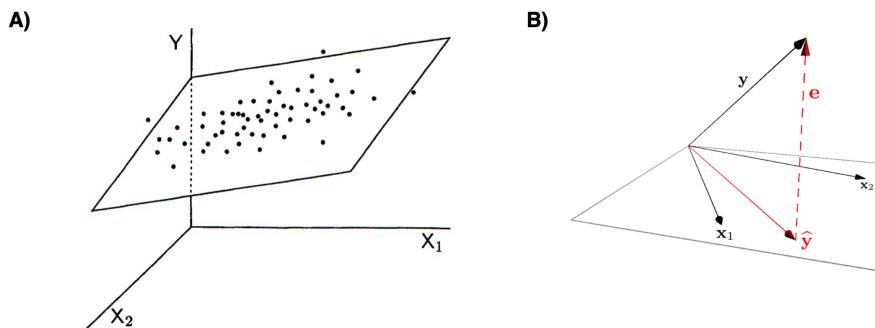


Abbildung 13.1: Regression auf zwei Eingangsgrößen x_1 und x_2 . Links: klassisch dargestellt im “Variable space”: n Punkte im dreidimensionalem Raum. Rechts: Darstellung im “Observation space”, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}$ (Vektoren) im n -dimensionalen Raum.

² $\hat{\mathbf{Y}} = X\hat{\beta} = X(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{P}\mathbf{Y}$. Die *Projektionsmatrix* oder *Hut-Matrix* ist die $n \times n$ Matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

³Da wir hier nur drei Vektoren zeichnen müssen, wurde der n -dimensionale Raum so gedreht, dass man nur noch drei Dimensionen sieht. Das gleiche passiert, wenn wir einen durchsichtigen Würfel mit zwei Vektoren so drehen, dass wir nur noch die Ebene mit den zwei Vektoren sehen.

Einfache versus multiple Regression. Im Allgemeinen ist es nicht angebracht, eine multiple Regression durch mehrere einfache Regressionen zu ersetzen (mit je einer Eingangsgrösse).

Die Lösungen der einfachen Regressionen sind nur dann gleich der Lösung der multiplen Regression, wenn die Prädiktoren **orthogonal**, also “senkrecht” zueinander sind⁴.

Multiple Regressionen sagen also *viel mehr* aus als einfache Regressionen. Im Allgemeinen ist die multiple Regression die Methode der Wahl, um Effekte von mehr als einem Prädiktor gleichzeitig einzubeziehen. Folgendes Beispiel einer Regression auf zwei Eingangsgrössen soll das illustrieren. Wir brauchen dazu die Daten **States** aus der package **carData**. Die Variablen sind:

- **SATM**: Average score of graduating high-school students in the state on the math component of the Scholastic Aptitude Test (a standard university admission exam).
- **percent**: Percentage of graduating high-school students in the state who took the SAT exam.
- **pay**: Average teacher’s salary in the state, in 1000s.

Wir stellen jetzt diese Regression geometrisch dar (Abbildung 13.2). Links sehen wir die Regression von **SATM** auf **percent** und **pay**. Rechts sehen wir zusätzlich die beiden einfachen Regression von **SATM** auf **percent** und von **SATM** auf **pay**.

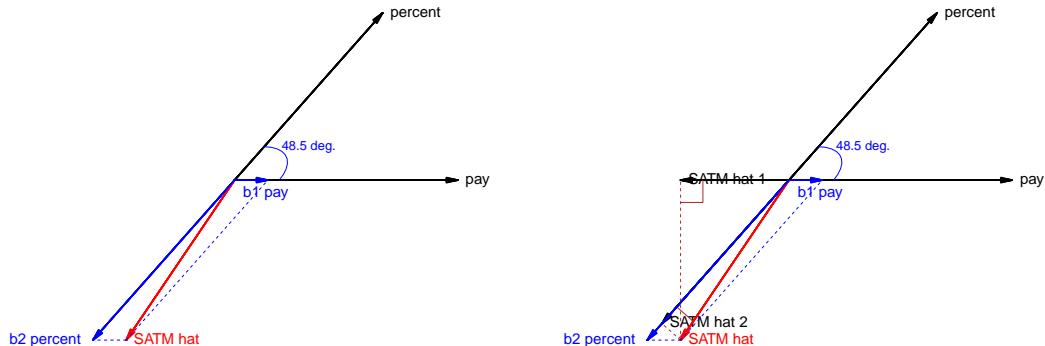


Abbildung 13.2: Links: Multiple Regression von **SATM** auf **percent** und **pay**, Rechts: Mit beiden einfachen Regressionen.

⁴Für die Interessierten: Wenn die Spalten von X orthogonal sind, dann ist $X^T X$ eine Diagonalmatrix mit Einträgen $\sum x_{i1}^2, \sum x_{i2}^2, \dots, \sum x_{ip}^2$ und $\hat{\beta}_j$ hängt dann nur von y_i und vom j ten Prädiktor ab.

Wir haben hier ein Beispiel für das sogenannte *Simpson-Paradoxon*: Der Effekt von `pay` dreht (ändert das Vorzeichen), wenn er nicht für `percent` kontrolliert wird. Gegeben `percent` ist der Effekt von `pay` auf SATM positiv. Unbedingt ist der Effekt von `pay` auf SATM negativ. Das heisst, Resultate beim Test sind umso besser, je weniger die Lehrer verdienen! Bei gleichbleibendem Anteil von Teilnehmern aber ist der Zusammenhang zwischen Verdienst und Resultat positiv.

Machen wir die multiple und die beiden einfachen Regressionen mit `lm()`:

```
lm(SATM ~ pay + percent, data = States)$coef  ## multiple regression on pay and percent

## (Intercept)      pay      percent
##      513.699     0.972    -1.374

lm(SATM ~ pay, data = States)$coef  ## single on pay

## (Intercept)      pay
##      595.19       -3.16

lm(SATM ~ percent, data = States)$coef  ## single regression on percent

## (Intercept)      percent
##      538.97      -1.23
```

Wir schauen uns diese Regression (auf zwei Eingangsgrößen) noch im Variablenraum an. Im zweidimensionalen Raum (Abbildung 13.3) sehen wir, dass der Effekt von `pay` auf SATM bei der einfachen Regression negativ ist. Im dreidimensionalen Raum (Abbildung 13.4) sehen wir schön, dass der Effekt (Steigung) von `pay` bei fixiertem `percent` positiv ist.

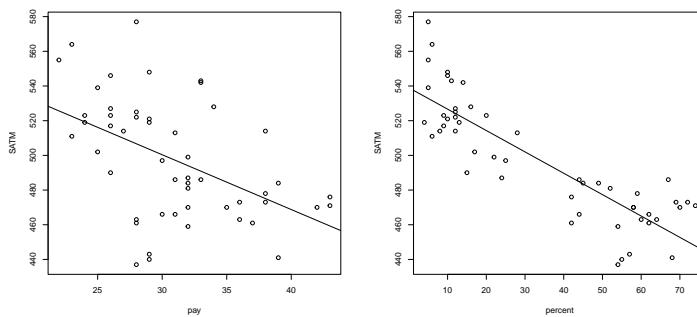


Abbildung 13.3: Einfache Regressionen von SATM auf percent und pay

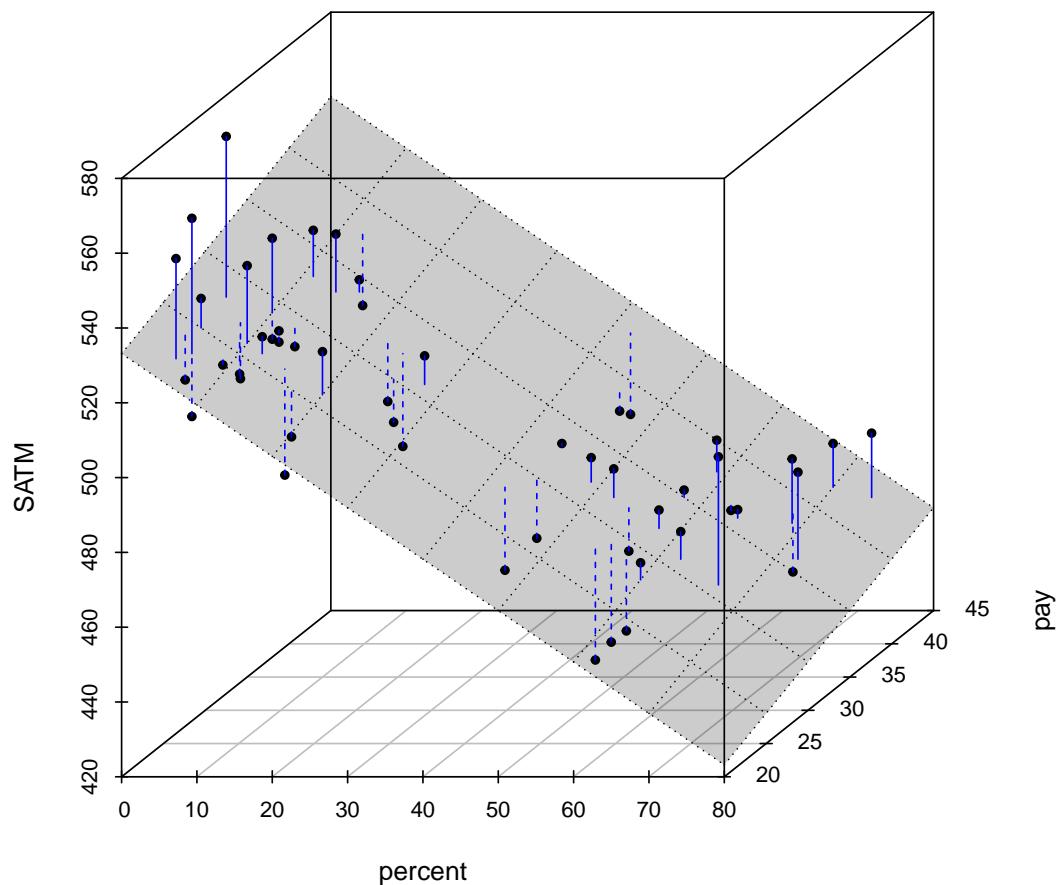


Abbildung 13.4: Regression von SATM auf percent und pay

13.6 Erwartungswerte, Varianz und Verteilung der Schätzer

Dieser kurze Abschnitt behandelt Erwartungswert und Varianz sowie die Verteilung des Schätzers. Damit sind wir dann in der Lage, Hypothesen zu testen und Schätzungen bezüglich den Parametern zu machen.

13.6.1 Erwartungswerte und Varianz der Schätzer

Ohne Annahme bezüglich der Verteilung gilt für Kleinste-Quadrat-Schätzer:

- Der Schätzer ist *unverzerrt*: $E(\hat{\beta}) = \beta$
- Für die $p \times p$ *Kovarianzmatrix* von $\hat{\beta}$

$$\text{Cov}(\hat{\beta}) = \begin{pmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{Var}(\hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_2, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_p, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_p, \hat{\beta}_2) & \cdots & \text{Var}(\hat{\beta}_p) \end{pmatrix} \quad (13.6.1)$$

gilt

$$\boxed{\text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}}. \quad (13.6.2)$$

Daraus werden die *Standardfehler* berechnet⁵.

- $E(\hat{Y}) = E(Y) = X\beta$
- $E(r) = \mathbf{0}$

Einstichprobenfall. Im Einstichprobenfall ohne Prädiktoren haben wir die bereits bekannten Momente,

- $E(\hat{\mu}) = E(\bar{Y}) = \mu$
- Wegen $(X^T X)^{-1} = 1/n$, haben wir $\text{Var}(\hat{\mu}) = \text{Var}(\bar{Y}) = \sigma^2/n$.

⁵

Beweis. Für die Interessierten*: Für Zufallsvariable Y gilt $\text{Var}(aY) = a^2 \text{Var}(Y)$. Ein Analogon gibt es jetzt für Zufallsvektoren. Die $n \times n$ Kovarianzmatrix Σ für die Fehler ist gemäss 13.4

$$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \quad (13.6.3)$$

Man kann nun zeigen, dass für Matrix A und Zufallsvektor \mathbf{Y} mit Kovarianzmatrix Σ gilt: $\text{Var}(AY) = A\Sigma A^T$. Also ist $\text{Var}(\hat{\beta}) = \text{Var}((X^T X)^{-1} X^T \mathbf{Y}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} = (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$.

13.6.2 Verteilung der Schätzer bei Normalverteilung

Wir nehmen nun zusätzlich an, dass ϵ_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Dann gilt

1. Parameterschätzungen: *multivariat normalverteilt*, $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^T X)^{-1})$
2. Geschätzte Residualvarianz: $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi_{n-p}^2$

Die Normalverteilung der Fehler ist oft nicht erfüllt. Für grosse n sind die Aussagen dann trotzdem annähernd wahr (Zentraler Grenzwertsatz). Das ist dann die herkömmliche Begründung in der Praxis für die Konstruktion von Konfidenzintervallen und Tests für die Parameter des Modells.

Wir kennen nun Erwartungswerte, Varianz und Verteilung der Schätzer. Wir können jetzt, wie bereits gelernt, Hypothesen testen und Parameter schätzen.

13.7 Tests und Konfidenzbereiche

Wir nehmen an, dass ϵ_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ oder n “gross”. Wir haben gesehen, dass der Schätzer $\hat{\beta}$ normalverteilt ist.

13.7.1 *t*-Test

Wir können jetzt die *Nullhypothese*

$$H_{0,j} : \beta_j = 0 \quad (13.7.1)$$

versus die Alternative $H_{A,j} : \beta_j \neq 0$ testen. σ^2 ist unbekannt, wir ersetzen es mit $\hat{\sigma}^2$ und erhalten die bereits bekannte *t*-Statistik

$$T_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p}$$

(13.7.2)

mit $\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2(X^T X)_{jj}^{-1}}$. Der zugehörige Test heisst ***t*-Test**.

Beim Betrachten der individuellen Tests kann es vorkommen, dass alle individuellen Tests die Nullhypothese nicht verwerfen, obwohl es wahr ist, dass einige Prädiktoren einen signifikanten Effekt haben. Dieses “Paradox” haben wir oben geometrisch dargestellt, es entsteht wegen der Korrelation der Prädiktorvariablen. Ein individueller *t*-Test quantifiziert den Effekt des j ten Prädiktors nach Subtrahieren des linearen Effekts von allen anderen Prädiktoren auf Y . Das haben wir oben in der geometrischen Darstellung illustriert.

Konfidenzintervalle für die Parameter. Für Irrtumswahrscheinlichkeit α ist dann ein $(1 - \alpha) \times 100\%$ Konfidenzintervall für die unbekannten Parameter β_j

$$\boxed{\hat{\beta}_j \pm t_{n-p, 1-\alpha/2} \text{se}(\hat{\beta}_j)} \quad (13.7.3)$$

Ein solches Intervall *überdeckt* das wahre β_j mit Wahrscheinlichkeit $1 - \alpha$.

13.7.2 Modellvergleich und F-Test

Um zu testen, ob es überhaupt einen Effekt von (einer Menge von) Prädiktoren gibt, können wir eine *simultane Nullhypothese* bilden wie

$$H_0 : \beta_2 = \dots = \beta_p = 0, \quad (13.7.4)$$

versus die Alternative $H_A : \beta_j \neq 0$ für mindestens ein $j \in 2, \dots, p$. Wir testen dann ein eingeschränktes (restricted) gegen ein volles Modell:

- *Eingeschränktes* Modell: Eine Untermenge von Parametern wird Null gesetzt (d.h. sie spielen keine Rolle bezüglich der Zielgröße)
- *Volles* Modell: Alle Parameter werden geschätzt.

Ein solcher Test wird über eine *Varianzanalyse* (ANOVA, Analysis of variance) gemacht. Wir haben in 13.5 gesehen, dass die Abbildung $\mathbf{Y} \rightarrow \hat{\mathbf{Y}}$ eine orthogonale Projektion von \mathbf{Y} auf einen p -dimensionalen Unterraum in \mathcal{R}^n darstellt. Der n -dimensionale Vektor der Residuen $\mathbf{Y} - \hat{\mathbf{Y}}$ steht also **senkrecht** auf dem Vektor $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$. Man kann demnach folgende Zerlegung von $\mathbf{Y} - \bar{\mathbf{Y}}$ über den Satz von PYTHAGORAS beweisen:

$$TSS = ESS + RSS$$

Diese Zerlegung haben wir schon in der einfachen Regression gebraucht, ohne dort den Pythagoras zu “sehen”.

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \quad (13.7.5)$$

Die Anzahl der Freiheitsgrade für die durch die Regression erklärte Quadratsumme ESS ist der Unterschied der Freiheitsgrade aus dem reduzierten Modell (in diesem Fall $n - 1$, da nur das Intercept bleibt) und dem vollen Modell, $n - p$, in dieser Situation also $(n - 1) - (n - p) = p - 1$. Die Anzahl Freiheitsgrade für die Fehlerquadratsumme ist dann $n - p$ und die der totalen Quadratsumme ist die Summe der beiden,

$$df_{Total} = df_{Explained} + df_{Residual}. \quad (13.7.6)$$

Eine solche Zerlegung wird mit einer ANOVA-Tabelle dargestellt:

	Sum of Squares (SS)	Freiheitsgrade (df)	Mean Squares (MS)	$E(MS)$
Regression	$\ \hat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2$	$p - 1$	$\ \hat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2 / (p - 1)$	$\sigma^2 + \frac{\ E(\mathbf{Y}) - E(\bar{\mathbf{Y}})\ ^2}{p-1}$
Fehler	$\ \mathbf{Y} - \hat{\mathbf{Y}}\ ^2$	$n - p$	$\ \mathbf{Y} - \hat{\mathbf{Y}}\ ^2 / (n - p)$	σ^2
Total	$\ \mathbf{Y} - \bar{\mathbf{Y}}\ ^2$	$n - 1$	—	—

Tabelle 13.1: ANOVA-Tabelle

Die Grösse für den Erwartungswert der mittleren Quadratsumme der Regression beweisen wir hier nicht. Wir schreiben die Tabelle abgekürzt:

	SS	df	MS
Regression	ESS	$p - 1$	EMS
Fehler	RSS	$n - p$	RMS
Total	TSS	$n - 1$	

Tabelle 13.2: ANOVA-Tabelle kurz

F-Statistik und Verteilung. Wenn H_0 wahr ist, dann gibt es keinen Effekt der Prädiktoren. Der zweite Summand des Erwartungswertes von EMS wird dann null und der Erwartungswert von EMS ist – wie für RMS – ebenfalls σ^2 .

Das hat zur Folge, dass der Quotient $\frac{EMS}{RMS}$ unter der Nullhypothese F -verteilt ist mit Zählerfreiheitsgrad $p - 1$ und Nennerfreiheitsgrad $n - p$ ⁶:

$$F = \frac{EMS}{RMS} \sim F_{p-1, n-p} \quad (13.7.7)$$

Dieser Test heisst **F-Test**. Die F -Verteilung ist in R implementiert mit `rf()`, `df()`, `qf()`, `pf()`.

Goodness-of-fit. Ein Mass für die Anpassungsgüte (“goodness of fit”) ist das bereits oben eingeführte R^2

$$R^2 = \frac{ESS}{TSS} \quad (13.7.8)$$

Diese Grösse stellt die Proportion der totalen Quadratsumme dar, die durch die Regression erklärt wurde.

⁶Weil $EMS/\sigma^2 \sim \chi^2_{p-1}/(p-1)$ und $RMS/\sigma^2 \sim \chi^2_{n-p}/(n-p)$ und der Definition der F -Verteilung.

13.8 Beispiel: Lineares Modell für Fertilität

Generische Funktionen für Modellobjekte der Klasse `lm`

- `print()`: Einfacher Output
- `summary()`: Standard Regression Output
- `coef()`: (oder `coefficients()`) extrahiert Koeffizienten
- `residuals()`: (oder `resid()`) extrahiert Residuen
- `fitted()`: (oder `fitted.values()`) extrahiert angepasste Werte
- `anova()`: Vergleich von verschachtelten Modellen
- `plot()`: Diagnostische Plots
- `confint()`: CI's für Koeffizienten
- `vcov()`: Geschätzte Kovarianzmatrix
- `predict()`: Vorhersagen für neue Daten
- `deviance()`: RSS (später)
- `logLik()`: log-likelihood (später)
- `AIC()`: Informationskriterium (später)

Viele andere Modelle in R haben analoge Funktionen. Später werden wir *generalisierte lineare Modelle* kennenlernen (`glm()`). Bei `summary()` wird dann im Hintergrund statt `summary.lm()` die Funktion `summary.glm()` ausgeführt (für Objekte der Klasse `glm`).

Modellobjekt ausdrucken. Bis jetzt hatten wir uns nur die Punktschätzungen:

```
mod <- lm(Fertility ~ ., data = swiss)
mod

##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Coefficients:
## (Intercept)      Agriculture    Examination     Education      Catholic
##             66.915          -0.172         -0.258        -0.871          0.104
## Infant.Mortality
##                 1.077

## print(mod) ## dasselbe
```

Wollen wir die Standardfehler, die t -Statistiken und die p -Wert, sowie einen F -Test, bekommen wir das mit `summary()`. Als Argument übergeben wir das angepasste Modellobjekt (“the fitted object”):

```
summary(mod)

##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -15.274 -5.262  0.503  4.120 15.321 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 66.9152   10.7060   6.25  1.9e-07  
## Agriculture -0.1721    0.0703  -2.45  0.0187  
## Examination -0.2580    0.2539  -1.02  0.3155  
## Education    -0.8709    0.1830  -4.76  2.4e-05  
## Catholic      0.1041    0.0353   2.95  0.0052  
## Infant.Mortality 1.0770    0.3817   2.82  0.0073  
## 
## Residual standard error: 7.17 on 41 degrees of freedom
## Multiple R-squared:  0.707, Adjusted R-squared:  0.671 
## F-statistic: 19.8 on 5 and 41 DF,  p-value: 5.59e-10
```

Reproduzieren der Größen im summary-Output. Wir wollen jetzt die Größen in diesem Output mit der Theorie, die wir gelernt haben, reproduzieren. Natürlich wird das später alles R für uns machen, aber um die Theorie zu wiederholen, reproduzieren wir die Resultate einmal “von Hand”.

Wir schauen uns zuerst die Design-Matrix X an:

```
X <- model.matrix(mod)
head(X)

##           (Intercept) Agriculture Examination Education Catholic Infant.Mortality
## Courtelary          1       17.0        15       12      9.96      22.2
## Delemont            1       45.1         6       9      84.84      22.2
## Franches-Mnt         1       39.7         5       5      93.40      20.2
## Moutier              1       36.5        12       7      33.77      20.3
## Neuveville           1       43.5        17       15      5.16      20.6
## Porrentruy           1       35.3         9       7      90.57      26.6
```

Da wir ein Modell mit Intercept haben, besteht die erste Kolonne aus Einsen. p ist die Anzahl Kolonnen und n ist die Anzahl Zeilen dieser Matrix:

```
dim(X) ##dimension von X

## [1] 47  6

n <- dim(X)[1]
p <- dim(X)[2]
```

Die Freiheitsgrade sind also $n - p$, Anzahl Beobachtungen minus Anzahl Parameter:

```
(df <- n - p)
```

```
## [1] 41
```

Die Komponenten von $\hat{\beta} = (X^T X)^{-1} X^T Y$ sind

```
Y <- swiss$Fertility
bhat <- solve(t(X) %*% X) %*% t(X) %*% Y
```

```
## [,1]
## (Intercept) 66.915
## Agriculture -0.172
## Examination -0.258
## Education -0.871
## Catholic 0.104
## Infant.Mortality 1.077
```

```
## coef(mod) ##dasselbe
```

Jetzt berechnen wir die Residuen-Quadratsumme RSS :

```
(RSS <- sum(resid(mod)^2))
```

```
## [1] 2105
```

```
## deviance(mod) ##dasselbe
```

Die geschätzte Fehlervarianz $\hat{\sigma}^2$ ist

```
(sigma2hat <- RSS/df)
```

```
## [1] 51.3
```

und $\hat{\sigma}$ ist damit

```
sqrt(sigma2hat)
```

```
## [1] 7.17
```

Die (geschätzte) Kovarianzmatrix $\hat{\sigma}^2(X^T X)^{-1}$ des Schätzers ist

```
kovmat <- sigma2hat * solve(t(X) %*% X)
round(kovmat, 4)
```

	(Intercept)	Agriculture	Examination	Education	Catholic	Infant.Mortality
## (Intercept)	114.6192	-0.4849	-1.2026	-0.2813	-0.0222	-3.2658
## Agriculture	-0.4849	0.0049	0.0044	0.0048	-0.0005	0.0066
## Examination	-1.2026	0.0044	0.0645	-0.0273	0.0051	0.0003
## Education	-0.2813	0.0048	-0.0273	0.0335	-0.0030	0.0123
## Catholic	-0.0222	-0.0005	0.0051	-0.0030	0.0012	-0.0027
## Infant.Mortality	-3.2658	0.0066	0.0003	0.0123	-0.0027	0.1457

```
## vcov(mod) ##dasselbe
```

Die Standardfehler sind die Wurzel der Diagonalelemente dieser Kovarianzmatrix, $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2(X^T X)^{-1}_{jj}}$:

```
(se <- sqrt(diag(vcov(mod))))
```

	(Intercept)	Agriculture	Examination	Education	Catholic
##	10.7060	0.0703	0.2539	0.1830	0.0353
## Infant.Mortality					
##	0.3817				

Die Werte der t -Statistiken sind die standardisierten Schätzungen:

```
(t <- bhat/se)
```

	[,1]
## (Intercept)	6.25
## Agriculture	-2.45
## Examination	-1.02
## Education	-4.76
## Catholic	2.95
## Infant.Mortality	2.82

Die p -Werte sind die Wahrscheinlichkeiten, solche t -Werte oder noch extremere zu beobachten (unter H_0):

```
(pval <- (1 - pt(abs(t), df)) * 2)
```

	[,1]
## (Intercept)	1.91e-07
## Agriculture	1.87e-02
## Examination	3.15e-01
## Education	2.43e-05
## Catholic	5.19e-03
## Infant.Mortality	7.34e-03

Führen wir das zusammen und vergleichen es mit dem `summary()`-Output von oben.

```
data.frame(estimate = bhat, SE = se, t = t, pval = pval)

##          estimate      SE       t     pval
## (Intercept) 66.915 10.7060  6.25 1.91e-07
## Agriculture -0.172  0.0703 -2.45 1.87e-02
## Examination -0.258  0.2539 -1.02 3.15e-01
## Education    -0.871  0.1830 -4.76 2.43e-05
## Catholic      0.104  0.0353  2.95 5.19e-03
## Infant.Mortality 1.077  0.3817  2.82 7.34e-03

## summary(mod)$coef ## dasselbe
```

Im `summary()` haben wir noch Quantile der Residuen

```
quantile(residuals(mod))

##      0%     25%     50%     75%    100%
## -15.274  -5.262   0.503   4.120  15.321
```

Jetzt reproduzieren wir noch den globalen F -Test für die Hypothese, dass keine der $p - 1 = 5$ Eingangsgrößen einen Effekt auf Fertilität hat.

$$H_0 : \beta_2 = \cdots = \beta_p = 0 \quad (13.8.1)$$

Dazu brauchen wir das eingeschränkte (restricted) Modell. In diesem Modell, auch *Nullmodell* genannt, gilt diese Nullhypothese, d.h. alle Parameter ausser dem Intercept werden Null gesetzt:

```
mod0 <- lm(Fertility ~ 1, data = swiss)
mod0

##
## Call:
## lm(formula = Fertility ~ 1, data = swiss)
##
## Coefficients:
## (Intercept)
##       70.1
```

Jetzt machen wir die Varianzanalyse. Wir vergleichen das volle Modell mit dem Nullmodell. Das geht in R mit `anova()`:

```
anova(mod0, mod)

## Analysis of Variance Table
##
## Model 1: Fertility ~ 1
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##             Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1       46 7178
## 2       41 2105  5     5073 19.8 5.6e-10
```

Die Residuen-Quadratsumme für das eingeschränkte Modell ist:

```
(RSS0 <- sum(residuals(mod0)^2))

## [1] 7178

(df0 <- n - 1)

## [1] 46
```

Diese ist natürlich grösser, da das Modell keine Eingangsgrössen hat. Die Freiheitsgrade sind $df_0 = n - 1$. Die durch das Modell erklärte Quadratsumme ist $ESS = RSS_0 - RSS$ mit Freiheitsgrad $df_0 - df = (n - 1) - (n - p) = p - 1$.

Wir berechnen nun die mittleren Quadratsummen, und die F -Statistik (die Menge an verfügbarem Fit (pro Freiheitsgrad), der erreicht wurde, wenn man vom Nullmodell zum vollen Modell geht),

$$F = \frac{EMS}{RMS} = \frac{(RSS_0 - RSS)/(df_0 - df)}{RSS/df} = \frac{(7177.955 - 2105.043)/5}{2105.043/41},$$

```
EMS <- (RSS0 - RSS)/(p - 1)
RMS <- RSS/(n - p)
(F <- EMS/RMS)

## [1] 19.8
```

und den zugehörigen p -Wert:

```
(pval <- 1 - pf(F, df1 = p - 1, df2 = n - p))

## [1] 5.59e-10
```

Damit haben wir den globalen F -Test aus dem `summary()`-Output reproduziert. Schliesslich reproduzieren wir noch das R^2 :

```
(R2 <- (RSS0 - RSS)/RSS0)

## [1] 0.707
```

Für Konfidenzintervalle für die Parameter β_j gibt es die Funktion `confint()`:

```
round(confint(mod), 3)

##              2.5 % 97.5 %
## (Intercept) 45.294 88.536
## Agriculture -0.314 -0.030
## Examination -0.771  0.255
## Education    -1.241 -0.501
## Catholic     0.033  0.175
## Infant.Mortality 0.306  1.848
```

Das können wir reproduzieren mit

```
bhat - qt(0.975, df) * se
bhat + qt(0.975, df) * se
```

Ein beliebter Output ist

```
round(cbind(summary(mod)$coef, confint(mod)), 3)

##           Estimate Std. Error t value Pr(>|t|) 2.5 % 97.5 %
## (Intercept) 66.915   10.706   6.25   0.000 45.294 88.536
## Agriculture -0.172    0.070  -2.45   0.019 -0.314 -0.030
## Examination -0.258    0.254  -1.02   0.315 -0.771  0.255
## Education    -0.871    0.183  -4.76   0.000 -1.241 -0.501
## Catholic      0.104    0.035   2.95   0.005  0.033  0.175
## Infant.Mortality 1.077    0.382   2.82   0.007  0.306  1.848
```

13.9 Testen von Hypothesen als Modellvergleiche

Jeder Hypothesentest bezüglich einzelnen Parametern kann als Vergleich von statistischen Modellen formuliert werden. Das Testen von Parametern ist also identisch mit dem Vergleich von zwei Modellen. Oben haben wir das Modell mit allen Eingangsgrößen verglichen mit dem Modell ohne Eingangsgröße, um zu testen, ob die Nullhypothese $\beta_2 = \dots = \beta_p = 0$ verworfen werden kann, ob das Nullmodell zugunsten des vollen Modells verworfen werden kann. Das war der Fall.

```
anova(mod0, mod)

## Analysis of Variance Table
##
## Model 1: Fertility ~ 1
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##             Infant.Mortality
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1     46 7178
## 2     41 2105  5     5073 19.8 5.6e-10
```

Man kann jetzt aber beliebige *verschachtelte* Modelle miteinander über einen *F*-Test vergleichen. Ein einfacheres (eingeschränktes) Modell erhält man dadurch, dass man gewisse Parameter des uneingeschränkten Modells fixiert (oft auf den Wert Null). Dann ist das eingeschränkte Modell verschachtelt im uneingeschränkten. Im Gegensatz dazu sind z.B. die Modelle

```
## Fertility ~ Examination + Catholic
```

und

```
## Fertility ~ Examination + Education
```

nicht verschachtelt.

Beispiel 1. Angenommen, wir wollen testen, ob der Effekt von `Catholic`, kontrolliert für alle anderen Eingangsgrößen, signifikant ist. Wir vergleichen dann das Modell mit

allen Eingangsgrößen ausser **Catholic** mit dem Modell mit allen Eingangsgrößen. Mit `update()` kann man Eingangsgrößen aus einem bestehenden Modell verändern. Man kann das Modell natürlich auch neu aufschreiben.

```
mod1 <- lm(Fertility ~ ., swiss)
mod0 <- update(mod1, . ~ . - Catholic)
## mod0<-lm(Fertility~.-Catholic,swiss) ##äquivalent
anova(mod0, mod1)

## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture + Examination + Education + Infant.Mortality
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     42 2553
## 2     41 2105  1      448 8.72 0.0052
```

Der `anova()`-Output sagt uns, dass das uneingeschränkte Modell (`mod1`) zusätzlich zu den Eingangsgrößen vom eingeschränkten Modell (`mod0`) noch die Eingangsgröße `catholic` hat. Der p -Wert von diesem F -Test ist **der gleiche** wie der p -Wert für den Koeffizienten im vollen Modell. Die einzelnen t -Tests für $H_{0,j}$ quantifizieren ja den Effekt des j ten Prädiktors, nachdem der Effekt aller anderen Prädiktoren auf die Zielgröße subtrahiert wurde!

Das ist der Test für die Hypothese, dass *der* Effekt von **Catholic**, der über den Effekt von allen anderen Eingangsgrößen hinausgeht, gleich Null ist. Diese Hypothese kann verworfen werden.

Beispiel 2. Wir wollen den Effekt von **Catholic** testen, kontrolliert für den Effekt von **Education** und **Examination**.

```
mod1 <- lm(Fertility ~ Education + Examination + Catholic, swiss)
mod0 <- update(mod1, . ~ . - Catholic)
anova(mod0, mod1)

## Analysis of Variance Table
##
## Model 1: Fertility ~ Education + Examination
## Model 2: Fertility ~ Education + Examination + Catholic
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     44 3550
## 2     43 3052  1      498 7.02 0.011
```

Das ist der Test für die Hypothese, dass *der* Effekt von **Catholic**, der über den Effekt von **Education** und **Examination** hinausgeht, gleich Null ist. Diese Hypothese kann verworfen werden.

Der p -Wert von diesem F -Test ist jetzt nicht mehr einer der t -Tests, da letztere immer einen individuellen Effekt testen, kontrolliert für *alle anderen* Größen im Modell.

Beispiel 3. Wir wollen testen, ob `Catholic` und/oder `Education` einen Effekt hat, kontrolliert für alle anderen Eingangsgrößen. Wir haben also die gemeinsame Nullhypothese $H_0 : \beta_{cath} = \beta_{educ} = 0$:

```
mod1 <- lm(Fertility ~ ., swiss)
mod0 <- update(mod1, . ~ . - Catholic - Education)
anova(mod0, mod1)

## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture + Examination + Infant.Mortality
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     43 3304
## 2     41 2105  2      1198 11.7 0.000097
```

Die Hypothese kann verworfen werden. Auch der p -Wert von diesem F -Test nicht mehr einer der t -Tests, da es um den Test einer anderen Hypothese geht.

Sequentielle Tests. Was passiert, wenn wir in die `anova()`-Funktion **ein** Modellobjekt übergeben (statt wie oben **zwei**)?

```
anova(mod)

## Analysis of Variance Table
##
## Response: Fertility
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## Agriculture       1    895    895  17.43 0.00015
## Examination       1   2210   2210  43.05 6.9e-08
## Education         1    892    892  17.37 0.00015
## Catholic          1    667    667  12.99 0.00084
## Infant.Mortality 1    409    409   7.96 0.00734
## Residuals        41   2105    51
```

`anova()` macht *sequentielle Tests*, und zwar in der Reihenfolge, wie die Prädiktoren in das Modell eingegeben wurden. Die erste Zeile testet `Agriculture`, kontrolliert für nichts, die zweite Zeile testet `Examination`, kontrolliert für `Agriculture`, die dritte Zeile testet `Education`, kontrolliert für `Agriculture`, `Examination`, usw. Die letzte Zeile testen `Infant.Mortality`, kontrolliert für alle anderen. Mit `update()` kann man obigen Output reproduzieren.

```
fit0 <- lm(Fertility ~ 1, swiss)
fit1 <- update(fit0, . ~ . + Agriculture)
fit2 <- update(fit1, . ~ . + Examination)
fit3 <- update(fit2, . ~ . + Education)
fit4 <- update(fit3, . ~ . + Catholic)
fit5 <- update(fit4, . ~ . + Infant.Mortality)
anova(fit0, fit1, fit2, fit3, fit4, fit5, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ 1
## Model 2: Fertility ~ Agriculture
## Model 3: Fertility ~ Agriculture + Examination
## Model 4: Fertility ~ Agriculture + Examination + Education
## Model 5: Fertility ~ Agriculture + Examination + Education + Catholic
## Model 6: Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     46  7178
## 2     45 6283  1      895 17.43 0.00015
## 3     44 4073  1     2210 43.05 6.9e-08
## 4     43 3181  1      892 17.37 0.00015
## 5     42 2514  1      667 12.99 0.00084
## 6     41 2105  1      409  7.96 0.00734
```

Wenn man die (partiellen) F -Tests will, die für die jeweils anderen Prädiktoren im grossen Modell kontrollieren, dann braucht man `drop1()`. Die Reihenfolge spielt dann keine Rolle:

```
drop1(fit5, test = "F")

## Single term deletions
##
## Model:
## Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##                   Df Sum of Sq  RSS AIC F value    Pr(>F)
## <none>              2105 191
## Agriculture       1      308 2413 195    5.99  0.0187
## Examination       1      53 2158 190    1.03  0.3155
## Education          1     1163 3268 209   22.64 0.000024
## Catholic           1      448 2553 198    8.72  0.0052
## Infant.Mortality  1      409 2514 197    7.96  0.0073
```

Die p -Werte sind dann identisch mit den p -Werten der individuellen t -Tests. Man hat also dasselbe Resultat wie im `summary()`-Output. Dort spielt die Reihenfolge auch keine Rolle.

Wir sehen im Output von `drop1()` eine neue Grösse, das AIC. Das führen wir im nächsten Kapitel ein.

13.10 Residuenanalyse und Modellannahmen

Die Residuen $r_i = Y_i - \hat{Y}_i$ stellen Approximationen für die nicht beobachteten Fehler ϵ_i dar. Sie können gebraucht werden, um Modellannahmen zu testen.

Der Tukey-Anscombe Plot. In einem *Tukey-Anscombe Plot* werden die Residuen r_i gegen die angepassten Werte \hat{Y}_i aufgetragen. Der Grund dafür ist, dass die Korrelation

zwischen den angepassten Werte und den Residuen Null ist. Das haben wir oben geometrisch gezeigt. Der Vektor der Residuen steht senkrecht auf dem Vektor des linearen Prädiktors.

Im Idealfall fluktuieren die Punkte zufällig um die horizontale Achse um Null, wie das die Abbildung 13.5 zeigt.

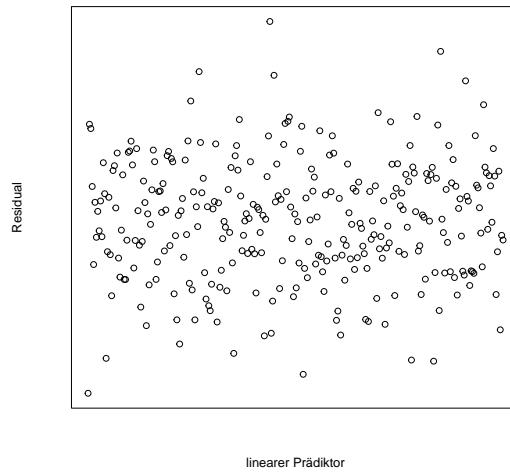


Abbildung 13.5: Tukey-Anscombe Plot bei erfüllten Modellannahmen.

Oft sieht man eine nicht konstante Variabilität der Residuen, was darauf hinweist, dass die Varianz der ϵ_i nicht konstant ist (Das war z.B. der Fall beim t -Test mit nicht-homogenen Varianzen). Das sieht man in der Abbildung 13.6 (A-C). Wenn der Plot einen Trend zeigt wie in D (Erwartungswert der Fehler ist nicht Null), dann ist das ein Hinweis auf einen systematischen Fehler (Hier müsste man einen quadratischen Term als Eingangsgröße hinzufügen).

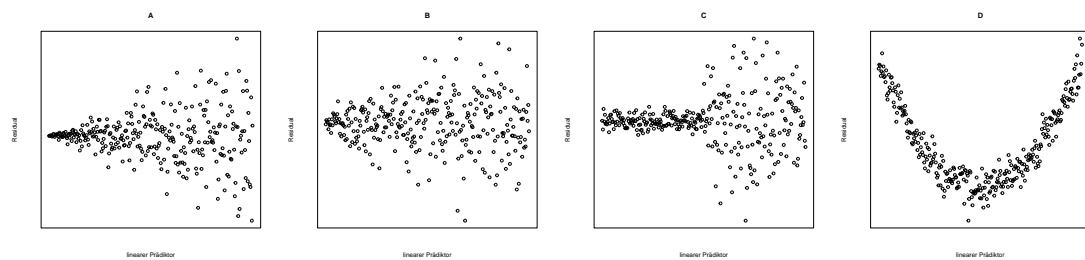
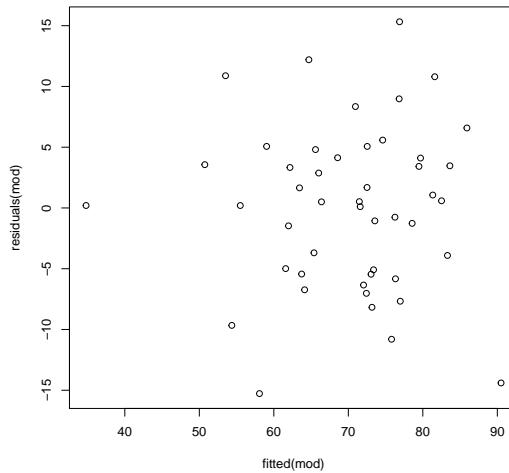


Abbildung 13.6: A: linear wachsende Standardabweichung, B: nicht-linear wachsende Standardabweichung, C: zwei Gruppen mit unterschiedlicher Varianz, D: fehlender quadratischer Term im Modell.

Falls es einen systematischen Zusammenhang gibt zwischen der Variabilität und den angepassten Werten, kann man die abhängige Variable transformieren. Falls die Standardabweichung linear wächst, stabilisiert oft die Logarithmusfunktion die Varianz.

Schauen wir uns den Tukey-Anscombe-Plot noch an für das Modell von Fertilität. Das wäre einigermassen akzeptabel.

```
plot(fitted(mod), residuals(mod))
```



```
## plot(mod,which=1) ## in R implementiert, 4 plots, davon den ersten nehmen
```

Quantil-Quantil-Plot. Die Annahmen bezüglich der Normalverteilung der Fehler können über den Quantil-Quantil-Plot (QQ-Plot) überprüft werden. Empirische Quantile der Residuen (auf der y -Achse) werden gegen die theoretischen Quantile der Standardnormalverteilung (x -Achse) aufgetragen. Bei normalverteilten Daten geht dieser Plot in Richtung einer Geraden mit Achsenabschnitt μ und Steigung σ . Abbildungen 13.7 und 13.8 zeigen QQ-Plots für normalverteilte- und nicht-normalverteilte Daten:

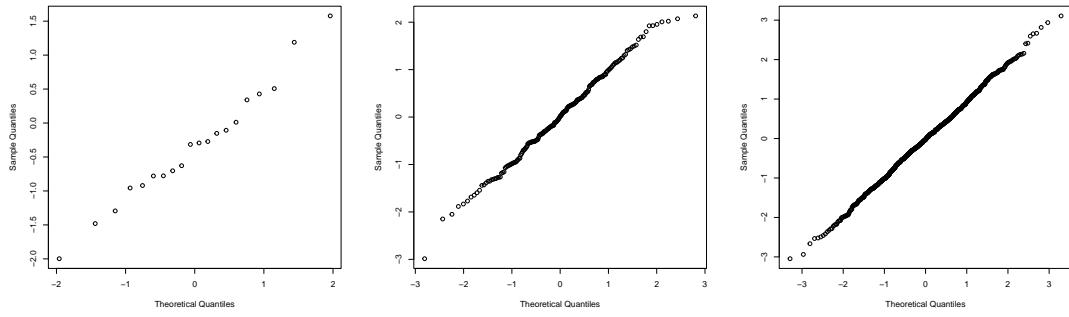


Abbildung 13.7: QQ-plots für i.i.d. normalverteilte Daten

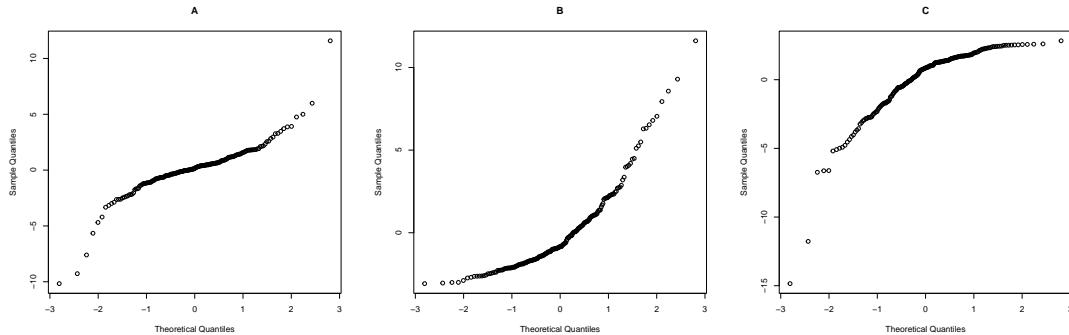
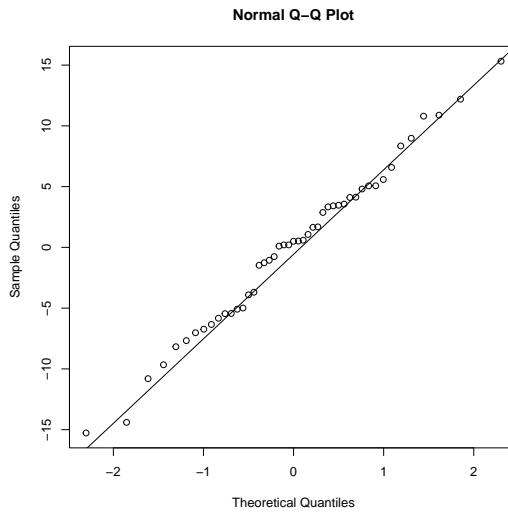


Abbildung 13.8: QQ-plots für A) langschwänzige Daten, B) rechtsschiefe Daten, C) linksschiefe Daten

Für unser Modell von Fertilität hätten wir

```
qqnorm(residuals(mod))
qqline(residuals(mod))
```



Als Steigung erkennen wir die Wurzel der Fehlervarianz $\hat{\sigma} = 7.165$.

Es muss hier nochmals betont werden, dass man die Normalverteilung der Residuen überprüft. **Die Normalverteilung der Y_i zu überprüfen ist sinnlos, da ja der lineare Prädiktor für jedes i verschieden ist.**

Verallgemeinerte Methode der kleinsten Quadrate*. Im linearen Modell sind die Fehler unabhängig und haben gleiche Varianz, die Kovarianzmatrix der Fehler ist dann

$$\begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Im allgemeinen Fall sind die Fehler korreliert mit bekannter Kovarianzmatrix Σ ,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \Sigma)$$

Das führt zur *verallgemeinerten Methode der kleinsten Quadrate*. Ein Spezialfall davon sind *gewichtete kleinste Quadrate*, wenn alle Elemente in den Nebendiagonalen von Σ null sind. Diese Methode kommt zum Zuge, wenn die Varianzen der beobachteten Werte nicht konstant sind und keine Korrelation zwischen den beobachteten Störgrößen vorliegt. Der Schätzer ist dann

$$\hat{\boldsymbol{\beta}}_{\text{weighted}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n w_i (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2,$$

mit Gewichten $w_i = 1/z_i$, und z_i als der Varianz von Y_i . Für die gewichtete Regression gibt es das Argument `weights` in `lm()`. Gewichtete Regression wird z.B. dann angewandt, wenn die Beobachtungen verschieden gewichtet werden sollen. Stellen z.B.

einzelne Beobachtungen Durchschnitte von mehreren Messungen dar, dann ist $w_i = n_i$, da ja $\text{Var}(\epsilon_i) = \sigma^2/n_i$.

Korrelierte Daten. Den Fall von korrelierten Fehlern, d.h. wenn die Unabhängigkeitsannahme verletzt ist, werden wir später mit den linearen gemischten Modellen (*Linear Mixed Models, LMM*) untersuchen, wenn wir wiederholte Messungen analysieren (Erweiterungen vom gepaarten *t*-Test).

Kapitel 14

Overfitting und Modellwahl

Wir haben gesehen, dass wir statistische Tests *von Parametern* über einen Vergleich von verschachtelten Modellen mit einem *F*-Test formulieren können. Ein einfacheres (eingeschränktes) Modell erhält man dadurch, dass man gewisse Parameter des uneingeschränkten Modells fixiert (oft Null).

Ein anderes Ziel ist häufig die *Modellwahl*. Wenn Modelle verschachtelt sind (eingeschränktes Modell als Spezialfall von einem grösseren Modell), dann kann man obiges Verfahren über den *F*-Test auch für die Modellwahl anwenden. Haben wir hingegen Modelle, die nicht verschachtelt sind, dann kann man nicht einen Modellvergleich über einen *F*-Test anwenden. Es braucht dann andere Kennwerte für die Modellwahl. Zuerst wollen wir das Problem der Überanpassung darstellen.

14.1 Überanpassung

Wie im Leben sollte man auch in der Statistik *nicht ohne Not Vielfältigkeit hinzufügen*, das hat WILLIAM VON OCCAM bereits 1320 formuliert. Dieses Sparsamkeitsprinzip, bekannt unter dem Namen *Ockhams Rasiermesser* ist ein heuristisches Forschungsprinzip, welches höchstmögliche Sparsamkeit bei der Bildung von Hypothesen und Theorien fordert. JOHN PONCE OF CORK schrieb 1639: *Variation must be taken as random until there is positive evidence to the contrary; and new parameters in laws, when they are suggested, must be tested one a time unless there is specific reason to the contrary.*

Könnte man z.B. R^2 brauchen für die Modellwahl? R^2 war definiert als

$$\frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Diese Grösse wird aber *immer* wachsen, wenn man zusätzliche Prädiktoren als Eingangsgrössen ins Modell nimmt. Je komplexer ein Modell (je mehr Parameter), umso grösser würde dieses Kriterium, und man würde immer das komplexere Modell wählen, wenn man nur die Anpassung als Kriterium betrachten würde. Wir haben gesehen, dass es immer möglich ist, zu einem (saturierten) Modell zu kommen, das perfekt zu den Daten passt (d.h. mit $R^2 = 1$).

Komplizierte Modelle sind aber schlecht für die Vorhersage für *neue* Daten. Jeder Koeffizient muss geschätzt werden und ist mit Variabilität behaftet. Diese Unsicherheiten addieren sich bezüglich der Variabilität der geschätzten (Hyper)-ebene, die wir in der geometrischen Betrachtung gesehen haben. Das nennt man *Überanpassung* oder *overfitting*. Das war in Abbildung 6.9 schön ersichtlich. Das rote Modell ist dort für die Vorhersage von neuen Daten offensichtlich schlecht.

Mit zunehmender Komplexität wird der systematische Fehler kleiner, diesbezüglich hat man also nichts zu verlieren mit vielen Eingangsgrößen. Die Varianz der Vorhersagen aber wird immer grösser mit zunehmender Komplexität. Es gibt also immer ein Spannungsfeld zwischen *Bias* und *Varianz*¹.

Simulation. In den Abbildungen 14.1 und 14.2 ist das Gesagte mit einer Simulation dargestellt. Es wurden $n = 100$ Daten aus folgendem Modell gezogen.

$$Y_i = 5 + 2x_i^2 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 0.1).$$

Es wurden dann vier Modelle angepasst:

- $Y_i = \beta_1 + \epsilon_i$ (“Null”)
- $Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ (“linear”)
- $Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i$ (“quadratisch”)
- $Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \dots + \beta_{10} x_i^9 + \epsilon_i$ (“Polynom neunten Grades”)

Wir sehen, dass das Nullmodell sehr schlecht passt. Das lineare Modell sieht schon sehr gut aus, wir sehen aber, dass das quadratische Modell noch besser passt. Das Polynom neunten Grades ist schon sehr unschön.

Anschliessend wurden vier mal neue Daten aus demselben Modell gezogen und die Modelle angepasst. Nur das Nullmodell und das Polynom neunten Grades wurden dargestellt. Wir können nun den Unterschied zwischen Bias und Varianz schön sehen. Das Nullmodell scheint klar falsch, verzerrt, aber ist es über alle vier Simulationen sehr ähnlich, hat also wenig Varianz. Im Gegensatz dazu scheint das komplexe Modell korrekt für alle 4 Simulationen (wenig verzerrt). Das Modell ist aber sehr variabel. Jeder Datensatz gibt ein ganz anderes geschätztes Modell.

Das bedeutet also, dass bei der Modellwahl nicht nur die Anpassung, sondern eben auch die Komplexität (und damit die Varianz) eine Rolle haben muss. Bei der Modellwahl braucht es also Kriterien, die komplexe Modelle *bestrafen*. Wir schauen uns jetzt zwei Statistiken, die beide für Modellkomplexität kontrollieren. Dazu gehören das wichtige *AIC* und der (etwas weniger wichtige) *adjusted R²*.

¹Dieses Spannungsfeld erleben wir auch in anderen Lebensbereichen.

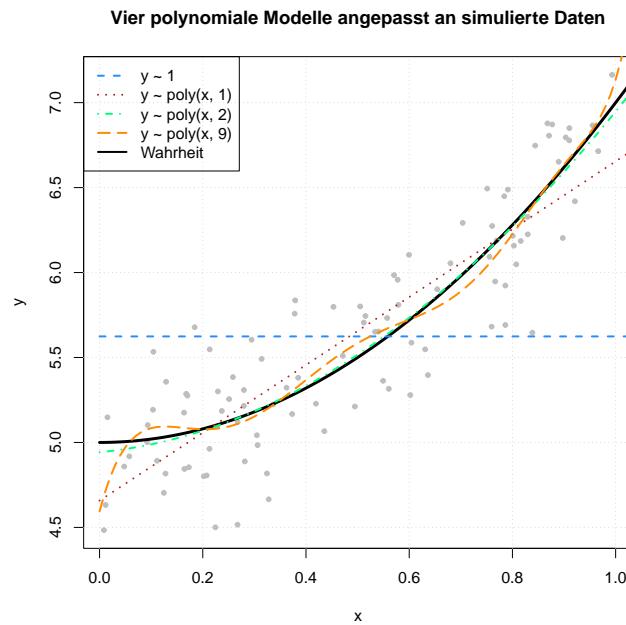


Abbildung 14.1: Simulation und Analyse mit vier Modellen.

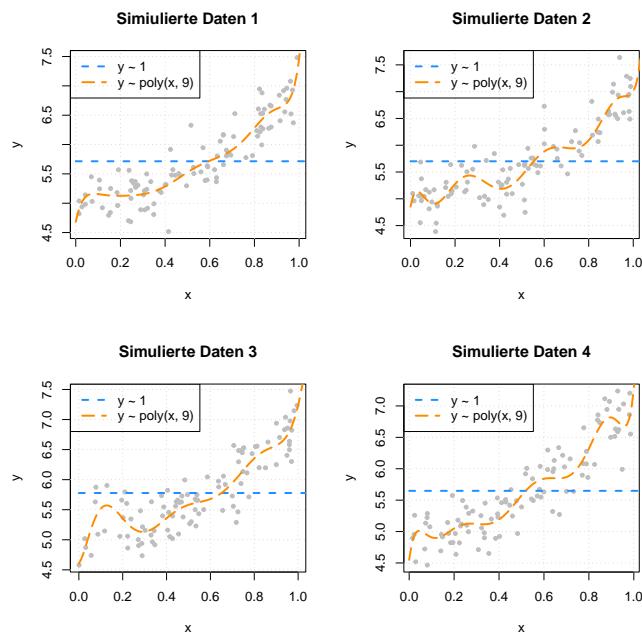


Abbildung 14.2: Vier neue Stichproben mit nachfolgender Analyse.

Bias-Varianz-Zerlegung*. Für die mathematisch Interessierten: Angenommen, wir haben eine datengenerierende Funktion

$$Y = f(x) + \epsilon \quad (14.1.1)$$

mit $E(\epsilon) = 0$ und $\text{Var}(\epsilon) = \sigma^2$. Wir möchten nun eine Funktion $\hat{f}(x; D)$ finden, die die wahre Funktion $f(x)$ approximiert. Dazu haben wir eine Stichprobe (Trainingsdaten) $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$.

Dafür wollen wir $(y - \hat{f}(x; D))^2$ minimal für x_1, \dots, x_n und für Punkte ausserhalb *unseres* Samples. Es gibt keine perfekte Lösung für eine solche Funktion, da ja die y_i Zufallsfehler ϵ_i beinhalten. Es wird also immer ein *irreduzibler Fehler* zu akzeptieren sein.

Mann kann dann zeigen, dass für jede gewählte Funktion \hat{f} der *erwartete Fehler* für eine nicht beobachtete Stichprobe (x^*, y^*) zerlegt werden kann²:

$$\underbrace{\mathbb{E}_{D, \epsilon} [(y^* - \hat{f}(x^*, D))^2]}_{\text{Erwarteter Fehler}} = \underbrace{\left(\text{Bias}_D [\hat{f}(x^*, D)] \right)^2}_{\text{Quadrierter Bias}} + \underbrace{\text{Var}_D [\hat{f}(x^*, D)]}_{\text{Varianz}} + \underbrace{\sigma^2}_{\text{Irreduzibel}} \quad (14.1.2)$$

Der irreduzible Fehler stellt eine untere Schranke für die erwartete Abweichung dar. Je komplexer ein Modell ist, umso mehr Datenpunkte wird das Modell erfassen und umso kleiner wird der Bias. Auf der anderen Seite wird die höhere Komplexität das Modell mehr schwanken lassen, was die Varianz grösser macht.

14.2 Modellwahl

AIC. Sei $\hat{\beta}$ der Maximum-Likelihood-Schätzer (der bei Normalverteilung identisch ist mit dem Kleinsten-Quadrat-Schätzter). Das *AIC* (Akaike information criterion) ist dann

$$AIC = -2l(\hat{\beta}) + 2p, \quad (14.2.1)$$

mit p als der Anzahl Parameter und $l(\hat{\beta})$ als der logarithmierten Likelihood am Maximum der Likelihood-Funktion. Bei hoher Anpassung des Modells an die Daten ist die Likelihood gross, aber das Modell ist zu *kompliziert* (viele Parameter) und wir haben Überanpassung. Modellkomplexität wird mit dem Strafterm $2p$ bestraft. Das präferierte Modell ist dasjenige mit dem kleineren AIC. Die Implementation in R ist `AIC()`.

²

- $\text{Bias}_D [\hat{f}(x^*, D)] = \mathbb{E}_D [\hat{f}(x^*, D)] - f(x^*)$: Verzerrung, verursacht durch ein zu einfaches Modell.
- $\text{Var}_D [\hat{f}(x^*, D)] = \mathbb{E}_D [(\hat{f}(x^*, D) - \mathbb{E}_D[\hat{f}(x^*, D)])^2]$: Varianz des Modells.
- σ^2 : Irreduzibler Fehler, untere Schranke für die erwartete Abweichung.

Adjusted R^2 . Eine (seltener gebrauchte) Alternative ist der *adjusted R^2* , den die `summary()`-Funktion auch ausdrückt. Diese Grösse korrigiert ebenfalls für die Modellkomplexität.

$$\text{adjusted } R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p} = 1 - \frac{RMS}{TMS} \quad (14.2.2)$$

mit p als der Anzahl von Prädiktoren. Im Gegensatz zum R^2 kann der adjusted R^2 kleiner werden wenn man Prädiktoren hinzufügt.

Beispiel Polynomiale Regression. Wir hatten in obiger Simulation vier Modelle zur Auswahl. Welches Modell ist optimal für die Vorhersage? Es sind verschachtelte Modelle, wir könnten hier auch die bekannten F -Tests brauchen. Wir wollen aber die Modellwahl über AIC und adjusted R^2 machen.

```
## nParameter logLik     R2     AIC Adjust.R2
##          1   -98.5 0.000 201.0      0.000
##          2   -24.8 0.771  55.6      0.769
##          3   -17.6 0.802  43.2      0.797
##         10   -15.0 0.812  52.0      0.793
```

Wir sehen, dass AIC (minimal), und adjusted R^2 (maximal) für das quadratische Modell mit drei Parametern optimal werden. R^2 und die Likelihood wächst mit der Anzahl der Parameter. Wir wählen also mit diesem Verfahren das wahre Modell (das wir hier natürlich nur kennen, weil wir aus diesem simuliert haben).

14.3 Schrittweise Prozeduren*

Wir kennen bereits dieses Modell:

```
## Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
```

Es gibt nun $2^5 = 32$ Modelle, die gleich oder verschachtelt in diesem Modell sind. Jede Eingangsgrösse kann im Modell sein oder nicht, das gibt 32 mögliche Kombinationen³.

Obwohl umstritten (und für uns nicht empfohlen), gibt es automatisierte Prozeduren, um zum optimalen Modell zu gelangen, z.B. `step()`. Defaultmässig macht diese Funktion einen schrittweise Rückwärts-Prozedur und sucht nach dem Modell mit optimalem (minimalem) AIC . Das ursprüngliche Modell ist:

```
summary(mod)$coef
```

³Die Anzahl möglicher Modelle wächst exponentiell, für $p = 25$ wären das schon $2^{25} = 33.55$ Millionen mögliche Modelle, bei $p = 30$ hat man schon $2^{30} = 1.07$ Milliarden Modelle.

```
##              Estimate Std. Error t value Pr(>|t|) 
## (Intercept)  66.915   10.7060   6.25 1.91e-07
## Agriculture -0.172    0.0703  -2.45 1.87e-02
## Examination -0.258    0.2539  -1.02 3.15e-01
## Education    -0.871    0.1830  -4.76 2.43e-05
## Catholic      0.104    0.0353   2.95 5.19e-03
## Infant.Mortality  1.077    0.3817   2.82 7.34e-03
```

Die Default-step-Prozedur hat als Resultat ein Modell ohne die Eingangsgrösse **Examination**:

```
stepmod <- step(mod)

## Start: AIC=191
## Fertility ~ Agriculture + Examination + Education + Catholic +
##           Infant.Mortality
##
##              Df Sum of Sq  RSS AIC
## - Examination     1      53 2158 190
## <none>                  2105 191
## - Agriculture     1      308 2413 195
## - Infant.Mortality 1      409 2514 197
## - Catholic         1      448 2553 198
## - Education        1     1163 3268 209
##
## Step: AIC=190
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##              Df Sum of Sq  RSS AIC
## <none>                  2158 190
## - Agriculture     1      264 2422 193
## - Infant.Mortality 1      410 2568 196
## - Catholic         1      957 3115 205
## - Education        1     2250 4408 221
```

Kapitel 15

Kategoriale Eingangsgrößen

Dieses Kapitel behandelt die linearen Modelle, die man klassischerweise als *Varianzanalysen* (ANOVA) bezeichnet. Man meint damit lineare Modelle mit kategorialen Eingangsgrößen. Wir haben den Begriff “Varianzanalyse” aber bereits allgemeiner gefasst als eine Methode innerhalb eines linearen Modells, mit der wir verschiedene Hypothesen testen können.

15.1 Eine kategoriale Eingangsgröße

Den Einstichproben-*t*-Test haben wir als lineares Modell ohne Eingangsgröße interpretiert. Den Zweistichproben-*t*-Test haben wir als lineares Modell mit einer zweiseitigen Eingangsgröße interpretiert. Die kategoriale Eingangsgröße – ein Faktor in R – kann jetzt mehr als zwei Stufen oder Kategorien haben, z.B. k Stufen.

Faktorstufe				
Stufe 1	y_{11}	y_{12}	\dots	y_{1n_1}
Stufe 2	y_{21}	y_{22}	\dots	y_{2n_2}
Stufe i	\vdots			\vdots
Stufe k	y_{k1}	y_{k2}	\dots	y_{kn_k}

Klassisch nennt man ein solches Design eine *einfaktorielle* Varianzanalyse. Für uns ist es ein Spezialfall vom Allgemeinen Linearen Modell.

15.1.1 Modell

Wenn man mit R Modelle anpasst, ist das Verständnis der Parameterisierung wichtig, insbesondere bei kategorialen Eingangsgrößen. Wir unterscheiden zwischen einer *Means-* und einer *Effekt*-Parameterisierung.

Means-Parameterisierung. Kategoriale Eingangsgrößen oder Prädiktoren mit k Kategorien werden mit einem Block von k Indikatorvariablen $x_{i1}, x_{i2}, \dots, x_{ik}$ kodiert.

Das führt uns zuerst zur *Means*-Parameterisierung:

$$Y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + \cdots + \mu_k x_{ik} + \epsilon_i, \quad i = 1, \dots, n \quad (15.1.1)$$

mit

$$x_{ij} = \begin{cases} 1 & \text{Beobachtung } i \text{ gehört zu Kategorie } j \\ 0 & \text{sonst.} \end{cases}$$

ϵ_i beschreibt den kombinierten Effekt aller unbekannten Einflüsse auf die abhängige Variable von Beobachtung i und wird dann meistens wieder als Zufallsvariable mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ angenommen.

Für eine Beobachtung i in Kategorie 1 gilt: $E(Y_i) = \mu_1$. Für eine Beobachtung i in Kategorie analog: $E(Y_i) = \mu_2$, usw. Der erwartete Abstand zwischen Gruppe 3 und 1 ist $\mu_3 - \mu_1$, usw.

Diese Parameterisierung wird häufig abgekürzt geschrieben. Für die Beobachtung j in der i -ten Gruppe (i indexiert jetzt die Gruppen!) ist das Modell

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \dots, k, j = 1, \dots, n_i. \quad (15.1.2)$$

Effekt-Parameterisierung. Im Gegensatz zu den Erwartungswerten in den Gruppen kann man für die Parameter auch *Effekte* brauchen. Das führt uns zur *Effekt*-Parameterisierung:

$$Y_i = \mu + \alpha_1 x_{i1} + \cdots + \alpha_k x_{ik} + \epsilon_i \quad i = 1, \dots, n, \quad (15.1.3)$$

mit μ als dem Gesamtmittel, $\alpha_1 = \mu_1 - \mu, \alpha_2 = \mu_2 - \mu, \dots, \alpha_k = \mu_k - \mu$. Die $\alpha_1, \dots, \alpha_k$ sind dann die erwarteten Abstände **relativ zum Gesamt-Mittel**. Diese Parameterisierung wird häufig abgekürzt geschrieben mit

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad i = 1, \dots, k, j = 1, \dots, n_i, \quad (15.1.4)$$

wobei i jetzt die Gruppen indexiert.

Anzahl Parameter. Das Modell hat zunächst zu viele Parameter, nämlich (neben σ^2) $1 + k$ Parameter für k Stichproben. Man macht das Modell *identifizierbar* mit einer Zusatzbedingung, z.B. $\sum_i \alpha_i = 0$ (*sum-to-zero* Kodierung). In R wird statt der Bedingung $\sum_i \alpha_i = 0$ standardmäßig $\alpha_1 = 0$ gesetzt (mit $\mu = \mu_1$) (*Referenzgruppe*-Kodierung). Es werden also die Effekte **relativ zu einer Referenzkategorie** (die alphabetisch erste, wenn man die Kategorien nicht sonst ordnet) als Parameter gesetzt. Das Modell sieht dann so aus, wie wir es von einem LM schon gewohnt sind,

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \dots, n. \quad (15.1.5)$$

Der erste Parameter steht dann für den Erwartungswert in der Referenzgruppe (Intercept), alle anderen Parameter für den erwarteten Unterschied zur Referenzgruppe:

Parameter	Interpretation
$\beta_1 = \mu_1$	Durchschnitt Referenzgruppe
$\beta_2 = \mu_2 - \mu_1$	Erwarteter Unterschied der zweiten Gruppe zur Referenz
$\beta_3 = \mu_3 - \mu_1$	Erwarteter Unterschied der dritten Gruppe zur Referenz
$\beta_k = \mu_k - \mu_1$	Erwarteter Unterschied der k -ten Gruppe zur Referenz

Tabelle 15.1: Effekt-Parameterisierung in R

15.1.2 Beispiel

Wir betrachten den eingebauten Datesatz `chickwts`.

```
str(chickwts)

## 'data.frame': 71 obs. of 2 variables:
## $ weight: num 179 160 136 227 217 168 108 124 143 140 ...
## $ feed   : Factor w/ 6 levels "casein","horsebean",...: 2 2 2 2 2 2 2 2 2 2 ...

by(chickwts$weight, chickwts$feed, psych::describe)

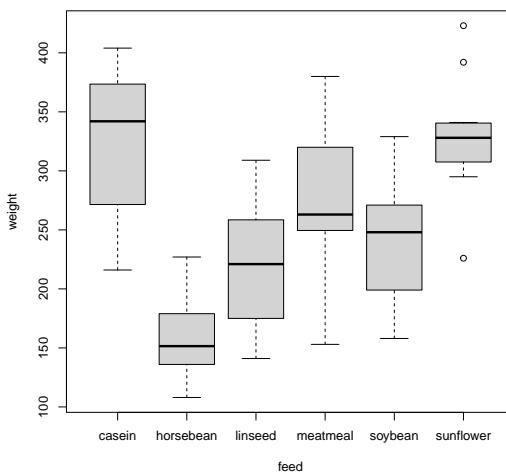
## chickwts$feed: casein
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 12 324 64.4    342     326  63 216 404    188 -0.46   -1.37 18.6
## -----
## chickwts$feed: horsebean
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 10 160 38.6    152     158 32.6 108 227    119 0.47   -1.19 12.2
## -----
## chickwts$feed: linseed
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 12 219 52.2    221     218 58.6 141 309    168 0.01   -1.33 15.1
## -----
## chickwts$feed: meatmeal
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 11 277 64.9    263     279 77.1 153 380    227 -0.25   -0.93 19.6
## -----
## chickwts$feed: soybean
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 14 246 54.1    248     247 53.4 158 329    171 0.03   -1.17 14.5
## -----
## chickwts$feed: sunflower
##   vars n mean sd median trimmed mad min max range skew kurtosis se
## X1    1 12 329 48.8    328     330 18.5 226 423    197 -0.05    0.06 14.1
```

Der Übersicht halber, und damit wir nachher mit den Parametern üben können, hier nur die geschätzten Mittelwerte $\hat{\mu}_i$ der sechs Gruppen sowie die geschätzten Abstände $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$ zum globalen Durchschnitt $\hat{\mu} = 261.31$.

	Durchschnitt	Effekt
1	323.58	62.27
2	160.20	-101.11
3	218.75	-42.56
4	276.91	15.60
5	246.43	-14.88
6	328.92	67.61

Tabelle 15.2: Durchschnitte und Effekte

```
boxplot(weight ~ feed, data = chickwts)
```



Modell anpassen. Wir passen jetzt ein Modell an für `weight`, die Eingangsgröße ist der Faktor `feed`.

```
modOne <- lm(weight ~ feed, data = chickwts)
summary(modOne)

##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -123.91  -34.41    1.57   38.17  103.09 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 323.58     15.83  20.44 < 2e-16 ***
## feedhorsebean -163.38    23.49  -6.96  2.1e-09 ***
##
```

```
## feedlinseed   -104.83    22.39   -4.68  1.5e-05
## feedmeatmeal   -46.67    22.90   -2.04  0.04557
## feedsoybean    -77.15    21.58   -3.58  0.00067
## feedsunflower     5.33    22.39    0.24  0.81249
##
## Residual standard error: 54.9 on 65 degrees of freedom
## Multiple R-squared:  0.542, Adjusted R-squared:  0.506
## F-statistic: 15.4 on 5 and 65 DF,  p-value: 5.94e-10
```

Wir können jetzt die oben eingeführte Parameterisierung erkennen, $\beta_1 = \mu_1 = \mu + \alpha_1$, $\beta_2 = \mu_2 - \mu_1 = \alpha_2 - \alpha_1$, etc:

- $\hat{\beta}_1$: Geschätzter Erwartungswert in der Gruppe `casein` (Referenzstufe)
- $\hat{\beta}_2$: Geschätzter Unterschied im Erwartungswert zwischen der Gruppe `horsebean` und der Gruppe `casein`
- $\hat{\beta}_3$: Geschätzter Unterschied im Erwartungswert zwischen der Gruppe `linseed` und der Gruppe `casein`
- $\hat{\beta}_4$: Geschätzter Unterschied im Erwartungswert zwischen der Gruppe `meatmeal` und der Gruppe `casein`
- $\hat{\beta}_5$: Geschätzter Unterschied im Erwartungswert zwischen der Gruppe `soybean` und der Gruppe `casein`
- $\hat{\beta}_6$: Geschätzter Unterschied im Erwartungswert zwischen der Gruppe `sunflower` und der Gruppe `casein`

Wenn eine Eingangsgröße und damit ein Term in der Modell-Formel einen Faktor beinhaltet (wie hier `feed`), sind die Tests für die einzelnen Koeffizienten nicht immer sinnvoll. Bei einer kategorialen Eingangsgröße ist nämlich zuerst die Frage, ob es überhaupt einen Effekt gibt von `feed` auf `weight`. Das hat zu tun mit dem Problem vom *Multiplen Testen*, auf das wir unten eingehen. Außerdem sehen wir "nur" die Effekte relativ zur Referenzkategorie. Das lösen wir unten, in dem wir *Kontraste* einführen.

Den *Globaltest* für den Effekt von `feed`: $H_0 : \beta_2 = \dots = \beta_k = 0$ machen wir mit `anova()` oder `drop1()`.

```
mod0 <- update(modOne, . ~ . - feed)
anova(mod0, modOne)

## Analysis of Variance Table
##
## Model 1: weight ~ 1
## Model 2: weight ~ feed
##   Res.Df   RSS Df Sum of Sq   F Pr(>F)
## 1     70 426685
## 2     65 195556  5    231129 15.4 5.9e-10
```

```
drop1(modOne, test = "F")

## Single term deletions
##
## Model:
## weight ~ feed
##   Df Sum of Sq   RSS AIC F value  Pr(>F)
## <none>          195556 574
## feed      5    231129 426685 620    15.4 5.9e-10
```

Da wir nur einen Faktor `feed` (kodiert mit 5 Indikatorvariablen plus ein Intercept für die Referenzkategorie) als Eingangsgrößen haben, entspricht der Test für `feed` dem F -Test im `summary()`-Output.

Wir haben also einen signifikanten Effekt von `feed` auf `weight`. Wir möchten jetzt *a posteriori* wissen, zwischen welchen Gruppen die entsprechenden Nullhypotesen verworfen werden können. Dazu brauchen wir den Begriff vom **Kontrast**.

Kontraste. Kontraste sind *Linearkombinationen von Parametern* mit der Bedingung, dass die Summe der Koeffizienten Null ist. Mit Konstanten a_1, a_2, \dots, a_t und Parametern $\theta_1, \theta_2, \dots, \theta_t$ ist ein Kontrast also

$$L = \sum_{i=1}^t a_i \theta_i, \quad \text{mit } \sum_{i=1}^t a_i = 0. \quad (15.1.6)$$

Dem Kontrast mit den Koeffizienten $a_1 = 0, a_2 = 0, a_3 = 0, a_4 = 0, a_5 = 1, a_6 = -1$ entspricht z.B. der erwartete Abstand von `soybean` auf `sunflower`, da $\hat{L} = \hat{\beta}_5 - \hat{\beta}_6 = -77.15 - 5.33 = -82.49$.

Kontraste werden konstruiert, um spezifische Forschungsfragen zu beantworten. Wir möchten jetzt alle *paarweisen* Kontraste berechnen. Das ist implementiert mit `emmeans()` aus dem gleichnamigen Packet. (“Estimated marginal means”). Wir bekommen dann alle paarweisen Kontraste mit Punktschätzung, Standardfehler, Intervallschätzung, t -Statistik und p -Werten. Die Berechnung von Standardfehlern und Konfidenzintervallen von Kontrasten ist z.T. komplex, das macht `emmeans()` für uns. Die Kontraste können dann graphisch dargestellt werden.

```
library(emmeans)
em <- emmeans(modOne, pairwise ~ feed)
summary(em, infer = c(TRUE, TRUE))

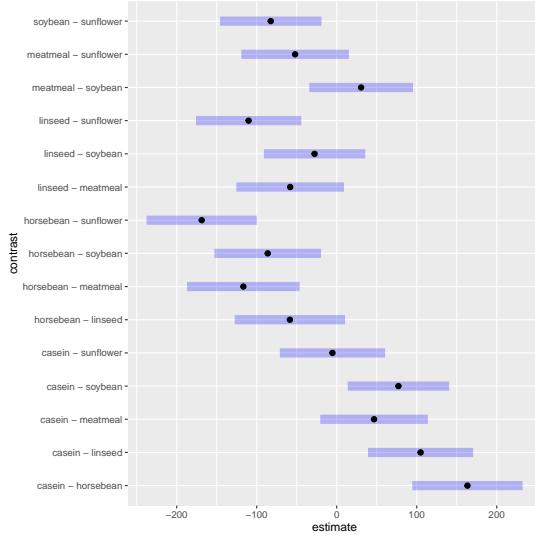
## $emmeans
##   feed    emmean    SE df lower.CL upper.CL t.ratio p.value
##   casein     324 15.8 65     292     355 20.440 <.0001
##   horsebean   160 17.4 65     126     195 9.240 <.0001
##   linseed    219 15.8 65     187     250 13.820 <.0001
##   meatmeal   277 16.5 65     244     310 16.740 <.0001
##   soybean    246 14.7 65     217     276 16.810 <.0001
##   sunflower   329 15.8 65     297     361 20.770 <.0001
##
## Confidence level used: 0.95
##
## $contrasts
##   contrast      estimate    SE df lower.CL upper.CL t.ratio p.value
##   casein - horsebean  163.4 23.5 65     94.4   232.3  6.960 <.0001
##   casein - linseed    104.8 22.4 65     39.1   170.6  4.680 <.0001
##   casein - meatmeal   46.7 22.9 65    -20.6   113.9  2.040  0.3320
##   casein - soybean    77.2 21.6 65     13.8   140.5  3.580  0.0080
##   casein - sunflower   -5.3 22.4 65    -71.1    60.4  -0.240  1.0000
##   horsebean - linseed -58.6 23.5 65   -127.5    10.4  -2.490  0.1410
##   horsebean - meatmeal -116.7 24.0 65   -187.1   -46.3  -4.870 <.0001
```

```

##  horsebean - soybean    -86.2 22.7 65   -152.9   -19.5  -3.800  0.0040
##  horsebean - sunflower  -168.7 23.5 65   -237.7   -99.8  -7.180  <.0001
##  linseed - meatmeal     -58.2 22.9 65   -125.4    9.1  -2.540  0.1280
##  linseed - soybean      -27.7 21.6 65   -91.0   35.7  -1.280  0.7930
##  linseed - sunflower    -110.2 22.4 65   -175.9   -44.4  -4.920  <.0001
##  meatmeal - soybean     30.5 22.1 65   -34.4   95.4  1.380  0.7390
##  meatmeal - sunflower   -52.0 22.9 65   -119.2   15.2  -2.270  0.2210
##  soybean - sunflower    -82.5 21.6 65   -145.9   -19.1  -3.820  0.0040
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 6 estimates
## P value adjustment: tukey method for comparing a family of 6 estimates

plot(em$contrasts)

```



Hier wird auch automatisch für *Multiples Testen* korrigiert (mit der *Tukey*-Methode). Auf dieses Problem gehen wir jetzt ein.

Multiples Testen. Es gibt einen wichtigen Grund, wieso man statt einem Globaltest nicht einfach zuerst immer je zwei Gruppen mit einem *t*-Test miteinander vergleicht. Bei einem Design mit einer dreiwertigen Eingangsgröße (3 Gruppen) würde man ja 3 Zwischengruppenvergleiche machen, nämlich die Vergleiche 1 – 2, 1 – 3 und 2 – 3. Bei k Gruppen hat man $m = \binom{k}{2}$ Vergleiche. Das Problem ist nun, dass bei vielen unabhängigen *t*-Tests der Typ-1-Fehler *kumuliert*.

Wir können das leicht erkennen, wenn wir uns die Falsch-Positiv-Raten beim Entscheiden gegen H_0 anschauen: Es sei $\Pr(T = 1 | H_0)$ die Wahrscheinlichkeit eines positiven Tests gegeben H_0 ($= \alpha$). Dann ist die Wahrscheinlichkeit, dass *irgendein* von m Tests positiv ist, gegeben H_0 , gerade das Komplement der Wahrscheinlichkeit, dass alle Tests

gleichzeitig negativ sind, nämlich

$$\alpha_{kum} = 1 - \Pr(T_1 = 0, \dots, T_m = 0 \mid H_0) = 1 - (1 - \alpha)^m. \quad (15.1.7)$$

Schon bei 4 Tests ist diese Wahrscheinlichkeit 0.19, bei 10 unabhängigen Tests 40%! Das Problem ist in Abbildung 15.1 dargestellt.

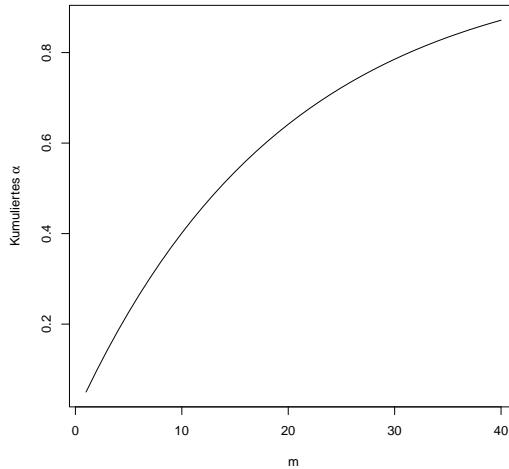


Abbildung 15.1: Kumulierung der Falsch-Positiv-Rate bei multiplem Testen

Allgemeiner – wenn also die Tests nicht unabhängig sind – kann man zeigen, dass – ausser wenn die Tests perfekt positiv korrelieren (also gleich sind), α immer noch mit der Zahl der Vergleiche wächst. Es gilt dann immer noch¹.

$$\alpha_{kum} \leq m\alpha \quad (15.1.8)$$

Die Wahrscheinlichkeit eines Typ-1 oder α -Fehlers steigt also mit der Anzahl Tests. Man sollte also nicht einfach statistische Tests kumulieren. Es ist auch unwissenschaftlich und sogar verwerflich, nach vielen statistischen Tests *a posteriori* gefundene Ergebnisse *im Nachhinein* zu *a priori* formulierten Hypothesen zu machen. Der Erkenntnisgewinn bei der Bestätigung einer *a priori* formulierten Hypothese ist viel grösser zu bewerten als *a posteriori* gefundene Ergebnisse ohne vorige Erwartungen.

Die einfachste Form, das α -Niveau anzupassen, ist die *Bonferroni*-Korrektur. Das Signifikanzniveau der einzelnen Tests α wird durch die Anzahl Tests geteilt, $\alpha^* = \frac{\alpha}{m}$, oder die p -Werte werden mit der Anzahl Tests multipliziert, $p^* = p \times m$.

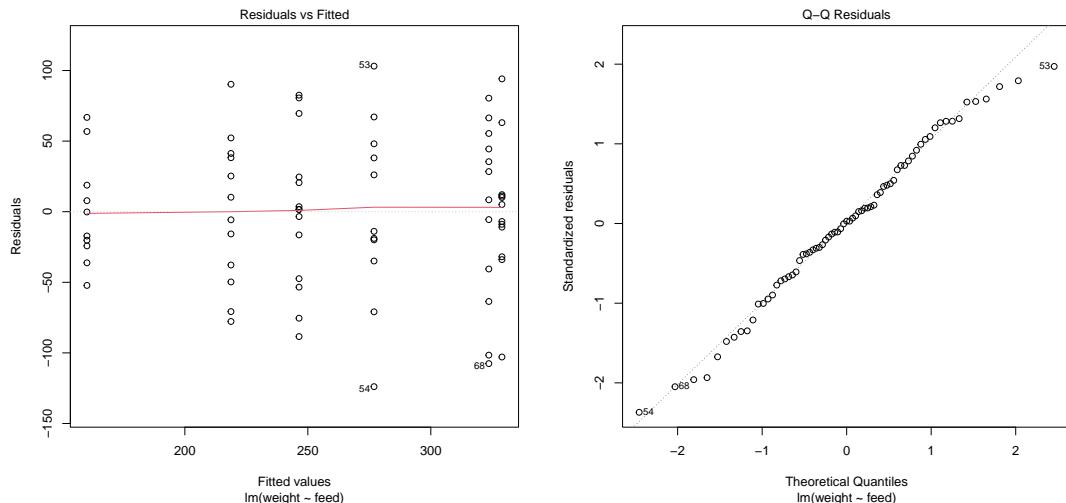
Diese Korrektur ist aber konservativ und es gibt viele andere, auf die wir hier nicht eingehen. Wir unterscheiden aber zwischen zwei Typen von Prozeduren für Multiples

¹Das folgt aus $\Pr(\cup_i A_i) \leq \sum_i \Pr(A_i)$.

Testen. Wenn die Vergleiche *geplant* sind (“*a priori*”), dann wird man nur für die geplanten Kontraste das α -Niveau korrigieren. Wenn man die Vergleiche im Voraus *nicht geplant* hat (“*a posteriori*”), dann werden alle möglichen Vergleiche korrigiert.

Residuenanalyse. Wir wollen noch die Modellannahmen testen, jetzt direkt mit dem implementierten `plot()` Befehl, wo wir das Modellobjekt als Argument übergeben (für uns sind im Moment nur die beiden ersten Plots wichtig, `which=c(1,2)`).

```
plot(modOne, which = c(1, 2))
```



Das sieht gut aus. Wir haben bereits früher gesagt, dass statistische Tests bezüglich homogenen Varianzen und Normalverteilung umstritten sind. Für die Interessierten: Ein bekannter Test für die Varianzhomogenität wäre der *Levene*-Test (im Packet `car`). Wir präferieren aber die visuelle Inspektion, vor allem vom TA-plot.

```
car::leveneTest(weight ~ feed, chickwts) ## 'test of homogeneity of variances'

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    5   0.75   0.59
##       65
```

15.2 Zwei kategoriale Eingangsgrößen

Wir haben jetzt *zwei* kategoriale Eingangsgrößen, einen Faktor A mit I Kategorien und einen Faktor B mit J Kategorien, wir haben also $I \times J$ Stichproben S_{ij} mit Grösse n_{ij} und $n = \sum_{i,j} n_{ij}$.

		Faktor B					
		B_1	B_2	\dots	B_j	\dots	B_J
Faktor A	A_1	S_{11}	S_{12}	\dots	S_{1j}	\dots	S_{1J}
	A_2	S_{21}	S_{22}	\dots	S_{2j}	\dots	S_{2J}
	\vdots					\vdots	
	A_i			\dots	S_{ij}	\dots	S_{iJ}
	\vdots					\vdots	
	A_I	S_{I1}	S_{I2}	\dots	S_{Ij}	\dots	S_{IJ}

Klassisch nennt man ein solches Design eine *zweifaktorielle* Varianzanalyse. Für uns ist es ein Spezialfall vom Allgemeinen Linearen Modell mit zwei kategorialen Eingangsgrößen A und B .

15.2.1 Modell

Means-Parameterisierung. Die abhängige Variable Y_{ijk} setzt sich zusammen aus dem linearen Prädiktor μ_{ij} und einem Fehler ϵ_{ijk} . Die μ_{ij} , die $I \times J$ Erwartungswerte sind dann die Parameter in der *Means*-Parameterisierung,

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (15.2.1)$$

Effekt-Parameterisierung. Für die *Effekt*-Parameterisierung ersetzen wir die μ_{ij} durch die Summe aus

- Gesamtmittel μ
- Haupteffekte von A , α_i
- Haupteffekte von B , β_j
- Interaktionseffekte $(\alpha\beta)_{ij}$

Das Modell schreiben wir dann

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (15.2.2)$$

$(\alpha\beta)_{ij}$ ist nicht als ein Produkt zu verstehen, sondern als ein Parameter für den Interaktionseffekt.

Anzahl Parameter. Das Modell hat $1 + I + J + (IJ)$ Parameter, also mehr Parameter als erlaubt; es gibt ja nur $I \times J$ Stichproben und damit $I \times J$ Erwartungswerte μ_{ij} . Wie beim einfaktoriellen Modell müssen wir das Modell identifizierbar machen, z.B. mit den *sum-to-zero*-Bedingungen. $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ für alle j und $\sum_j (\alpha\beta)_{ij} = 0$ für alle i . Durch die Zusatzbedingungen haben wir dann $1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = I \times J$ Parameter.

Effekt-Parameterisierung in R. Wie im Modell mit einer kategorialen Eingangsgröße ist in R standardmäßig nicht die *sum-to-zero* Parameterisierung eingestellt, sondern die *Referenzgruppe*-Parameterisierung:

- $\alpha_1 = 0$
- $\beta_1 = 0$
- $(\alpha\beta)_{1j} = 0$ für alle j und $(\alpha\beta)_{i1} = 0$ für alle i

Wir werden diese Parameterisierung unten bei einem Beispiel vertiefen.

Interaktionseffekte und Prinzip der Marginalität. In einem zweifaktoriellen Design haben Faktoren A und B einen Interaktionseffekt, wenn die erwartete Veränderung in der Zielgröße bei einer Veränderung der Stufe auf dem Faktor A von der Stufe des Faktors B abhängt, und umgekehrt. Wenn jede Veränderung der Stufe des Faktors A über alle Stufen vom Faktor B konstant ist, dann haben die beiden Faktoren keinen Interaktionseffekt, sie agieren *additiv*. Die Haupteffekte sind dann wohldefiniert und diese zu testen macht Sinn.

Wenn es jedoch einen Interaktionseffekt gibt, ist der Haupteffekt nicht von Interesse, da ja der Effekt der Veränderung der Stufe auf A auf die Zielgröße von der Stufe auf B abhängt.

Das Testen von Haupteffekten in der Gegenwart von Interaktionseffekten verletzt das *Prinzip der Marginalität*. Dieses Prinzip ist zwar nicht universell, aber in praktisch allen Anwendungen macht es Sinn, dieses zu respektieren. Meistens, und so werden wir es halten, ist es falsch, Haupteffekte zu interpretieren in der Gegenwart von Interaktionseffekten, oder Haupteffekte aus einem Modell zu entfernen, wenn Interaktionseffekte signifikant sind.

15.2.2 Beispiel.

Simulation*. Der folgende Code simuliert Daten aus einem zweifaktoriellen Design.

```
set.seed(22)
nage <- 3 ## Anzahl Kategorien Altersgruppe
ntherapy <- 2 ## Anzahl Kategorien Therapieart
n <- 200 ## Totale Sample Size
age <- factor(sample(c("child", "young", "old"), size = n, replace = TRUE, prob = c(1, 2, 3)), levels = c("child", "young", "old"))
```

```

therapy <- factor(sample(c("Ctrl", "Trt"), size = n, replace = TRUE))
beta1 <- 40 ## Referenz
betaAge <- c(10, 20) ## betaAge2 und betaAge3
betaTr <- c(10) ## betaTreat
alphabeta <- c(-0, 0) ## youngTrt und oltTrt, Interaktionseffekte=0
parameter <- c(beta1, betaAge, betaTr, alphabeta) ## Wahrer Parametervektor
sigma <- 12 ## Noise SD
epsilon <- rnorm(n, 0, sigma) ## Fehler
X <- model.matrix(~age * therapy) ## Design-Matrix
response <- as.numeric(X %*% parameter + epsilon) ## Y=Xbeta+epsilon, Daten ziehen aus Modell
d.cat2 <- data.frame(response, age, therapy)

```

Daten. Wir haben nun einen Datensatz `d.cat2` mit $n = 200$ Beobachtungen auf den Faktoren `age` (dreiwertig) und `therapy` (zweiwertig.)

```

str(d.cat2)

## 'data.frame': 200 obs. of 3 variables:
## $ response: num 78.5 45.6 11.1 51.5 20.6 ...
## $ age     : Factor w/ 3 levels "child","young",...: 3 3 1 2 1 2 2 2 3 3 ...
## $ therapy : Factor w/ 2 levels "Ctrl","Trt": 2 2 1 2 1 2 2 2 2 2 ...

head(d.cat2)

##   response age therapy
## 1    78.5   old     Trt
## 2    45.6   old     Trt
## 3    11.1 child    Ctrl
## 4    51.5 young    Trt
## 5    20.6 child    Ctrl
## 6    43.4 young    Trt

```

Ziel. Wir wollen den Effekt von `therapy` auf die Zielgröße quantifizieren, und für `age` kontrollieren, wenn nötig.

Beschreibende Statistik. Im folgenden schauen wir mit `table()` die Häufigkeiten aller Kombinationen an. Die deskriptive Statistik machen wir tabellarisch mit `table()`, `by()`, `aggregate()`, `tapply()` und graphisch mit `interaction.plot()` (Abbildung 15.2).

```

addmargins(table(therapy, age), 2)

##      age
## therapy child young old Sum
##   Ctrl     20     25  54  99
##   Trt      10     34  57 101

```

Wir sehen, dass die Altersgruppen innerhalb der Treatmentgruppen nicht gleich gross sind. Das Design ist dann *nicht balanciert*. Die graphische Darstellung deutet auf keinen Interaktionseffekt hin.

```
interaction.plot(x.factor = age, trace.factor = therapy, response = response, trace.label = "treatment",
  xlab = "age group", ylab = "mean of response")
```

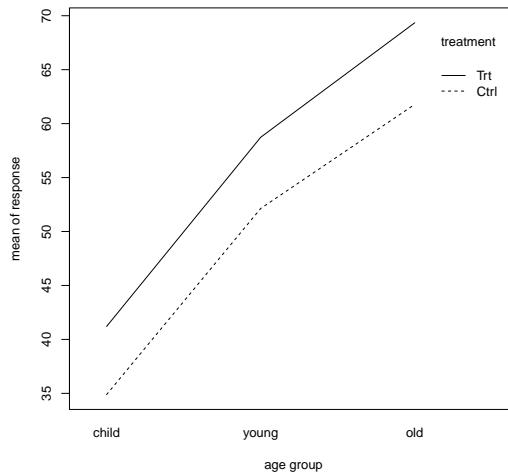


Abbildung 15.2: Beobachtete Durchschnitte bezüglich Faktoren age und therapy.

Wir beschreiben die Daten pro Gruppe auf `therapy` und dann pro Kombination auf `therapy` und `age`

```
psych::describeBy(response, list(therapy), mat = TRUE)

##      item group1 vars   n  mean    sd median trimmed mad  min  max range skew kurtosis   se
## X11     1   Ctrl   1 99 53.9 16.3   54.4    54.9 15.6 -4.0 99.7 103.7 -0.617   1.386 1.64
## X12     2     Trt   1 101 63.0 14.3   64.2    63.4 15.8 21.6 93.3  71.6 -0.326  -0.273 1.43

psych::describeBy(response, list(therapy, age), mat = TRUE)

##      item group1 group2 vars   n  mean    sd median trimmed mad  min  max range skew kurtosis   se
## X11     1   Ctrl  child   1 20 34.9 16.5   38.0    36.3 16.4 -4.0 56.6  60.6 -0.6681  -
## 0.584 3.69
## X12     2     Trt  child   1 10 41.2 11.8   44.4    41.5 13.7 21.6 58.4  36.8 -0.1773  -
## 1.375 3.72
## X13     3   Ctrl  young   1 25 52.1 11.7   50.9    51.9 15.2 34.1 71.6  37.5  0.2074  -
## 1.326 2.33
## X14     4     Trt  young   1 34 58.7 12.4   55.2    58.3 13.9 40.9 81.5  40.6  0.3188  -
## 1.273 2.12
## X15     5   Ctrl   old    1 54 61.8 11.5   60.9    61.4 11.6 38.6 99.7  61.2  0.5691  0.929 1.56
## X16     6     Trt   old    1 57 69.3 10.9   69.3    69.5 11.2 45.6 93.3  47.7 -0.0521  -
## 0.341 1.44
```

Tabelle 15.3 fasst die 6 Durchschnitte $\hat{\mu}_{ij}$ zusammen und zeigt die *ungewichteten* Mittel (Durchschnitte der Durchschnitte) und die mit der jeweiligen Gruppengröße *gewichteten* Mittel (Durchschnitte in Therapiegruppen unabhängig von `age`).

	child	young	old	unweighted.mean	weighted.mean
Ctrl	34.89	52.13	61.78	49.60	53.91
Trt	41.20	58.74	69.35	56.43	62.99

Tabelle 15.3: Ungewichtete und Gewichtete Mittel

“Durchschnitte”. Die beiden Durchschnitte (je für `Ctrl` und `Trt`) in den beiden letzten Spalten sind nur gleich, wenn das Design *balanciert* ist, wenn die Altersgruppen gleich gross sind, im Allgemeinen ist das aber – wie hier – nicht der Fall. Es wird sich zeigen, dass die Frage welche Durchschnitte (ungewichtet oder gewichtet) relevant sind – dasselbe ist wie die Frage nach der getesteten Hypothese.

Wenn der Durchschnitt von allen in `Ctrl` (`Trt`) die Kontrollgruppe (Treatmentgruppe) repräsentiert (das gewichtete Mittel), dann interessiert uns der Effekt von `therapy` auf den Stufen von `age`, wie sie in unserer Stichprobe sind, d.h. wir ignorieren die Variable `age` und beschreiben einfach die abhängige Variable in beiden Gruppen von `therapy`.

Wenn wir die Mittelwerte der Mittelwerte brauchen, dann interessiert uns der Effekt von `therapy` *kontrolliert* für `age`. Das erlaubt uns, sinnvolle Schlüsse zu ziehen für den Unterschied zwischen den Therapiegruppen mit *gleicher* Anzahl in jeder Altersgruppe.

Modell anpassen. Wir passen jetzt ein Modell an mit Haupt- und Interaktionseffekten. Die Modell-Formel kann man auf zwei verschiedene Arten eingeben, mit `A+B+A:B` oder mit `A*B`. In einer Modellformel stellt `A:B` einen Interaktionseffekt dar.

```
modelInt <- lm(response ~ therapy * age, data = d.cat2)
## modelInt <- lm(response~therapy+age+therapy:age,data=d.cat2)
summary(modelInt)

##
## Call:
## lm(formula = response ~ therapy * age, data = d.cat2)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -38.89  -8.48  -0.72   8.07  37.93 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 34.889     2.702   12.91 < 2e-16 ***
## therapyTrt  6.314     4.681    1.35    0.18    
## ageyoung    17.245    3.626    4.76   3.8e-06 ***
## ageold      26.893    3.164    8.50   5.0e-15 ***
## therapyTrt:ageyoung  0.288    5.661    0.05    0.96    
## therapyTrt:ageold   1.251    5.213    0.24    0.81    
##
```

```
##
## Residual standard error: 12.1 on 194 degrees of freedom
## Multiple R-squared:  0.442, Adjusted R-squared:  0.428
## F-statistic: 30.7 on 5 and 194 DF,  p-value: <2e-16
```

Interpretation Parameter. Es werden dann wie oben beschrieben $I \times J = 6$ Parameter geschätzt, aber jetzt wieder mit der Default-Effekt-Parameterisierung in R. child und Ctrl sind die beiden **Referenzkategorien**. Relativ zu diesen sind dann alle Effekte zu interpretieren.

	child	young	old
Control	μ_{11}	μ_{12}	μ_{13}
Treatment	μ_{21}	μ_{22}	μ_{23}

- (Intercept): Geschätzter Erwartungswert für child in Ctrl: $\hat{\mu}_{11}$
- therapyTrt: Geschätzter erwarteter Abstand von Trt zu Ctrl für child: $\hat{\mu}_{21} - \hat{\mu}_{11}$
- ageyoung: Geschätzter erwarteter Abstand von young zu child für Ctrl: $\hat{\mu}_{12} - \hat{\mu}_{11}$
- ageold: Geschätzter erwarteter Abstand von old zu child für Ctrl: $\hat{\mu}_{13} - \hat{\mu}_{11}$
- therapyTrt:ageyoung: Geschätzter erwarteter Unterschied im Unterschied Trt-Ctrl für young versus child: $\hat{\mu}_{22} - \hat{\mu}_{12} - (\hat{\mu}_{21} - \hat{\mu}_{11})$
- therapyTrt:ageold: Geschätzter erwarteter Unterschied im Unterschied Trt-Ctrl für old versus child: $\hat{\mu}_{23} - \hat{\mu}_{13} - (\hat{\mu}_{21} - \hat{\mu}_{11})$

Wichtig. Interaktionen sind **Unterschiede von Unterschieden** und drücken sich in der Abbildung 15.2, wenn vorhanden, durch entsprechend *nicht-parallele* Linien aus.

Wenn wir im Moment nur die Punktschätzungen anschauen, dann hängt der geschätzte Therapieeffekt vom Alter ab (Wir werden unten aber sehen, dass diese Interaktion nicht statistisch signifikant ist). Dieser ist

- 6.314 für child
- 6.314 + 0.288 für young
- 6.314 + 1.251 für old

Der Durchschnitt von diesen drei Effekten ist gleich dem Unterschied der ungewichteten Mittel in Tabelle 15.3. Man kann den Therapieeffekt (Punktschätzung) auch kurz mit Indikatorvariablen schreiben:

$$\text{Therapieeffekt} = 6.314 + 0.288 \times I_{\text{young}} + 1.251 \times I_{\text{old}}$$

Testen von Hypothesen. Wir passen nun auch ein Modell an ohne den Interaktionseffekt und die beiden Modelle mit nur einer Eingangsgröße:

```
modelMain <- update(modelInt, ~. - therapy:age) ##Nur Haupteffekte
modelTherapy <- update(modelMain, ~. - age) ##Nur therapy
modelAge <- update(modelMain, ~. - therapy) ##Nur Age
```

Ist der Effekt von `therapy` abhängig von der Altersgruppe? Ist der Interaktionseffekt signifikant? Wie immer machen wir das wieder über `anova()` oder `drop1()`.

```
anova(modelMain, modelInt)

## Analysis of Variance Table
##
## Model 1: response ~ therapy + age
## Model 2: response ~ therapy * age
##   Res.Df   RSS Df Sum of Sq   F Pr(>F)
## 1     196 28351
## 2     194 28337  2      13.7 0.05  0.95

anova(modelInt) ## geht auch, da sequentiell und INTERAKTIONSEFFEKT AM SCHLUSS

## Analysis of Variance Table
##
## Response: response
##             Df Sum Sq Mean Sq F value Pr(>F)
## therapy      1   4117   4117  28.19  3e-07
## age          2  18313   9157  62.69 <2e-16
## therapy:age  2     14      7  0.05  0.95
## Residuals   194 28337    146

drop1(modelInt, test = "F") ## drop1 respektiert das Prinzip der Marginalität, Haupteffekte werden nicht getestet.

## Single term deletions
##
## Model:
## response ~ therapy * age
##             Df Sum of Sq   RSS AIC F value Pr(>F)
## <none>            28337 1003
## therapy:age  2      13.7 28351  999  0.05  0.95
```

Der Interaktionseffekt ist nicht signifikant, d.h. wir können das Modell ohne Interaktion nicht verwerfen. Wir können jetzt Hypothesen testen bezüglich den Haupteffekten. Wir testen den Effekt von `therapy`, kontrolliert für `age`:

```
anova(modelAge, modelMain)

## Analysis of Variance Table
##
## Model 1: response ~ age
## Model 2: response ~ therapy + age
##   Res.Df   RSS Df Sum of Sq   F   Pr(>F)
## 1     197 30817
## 2     196 28351  1      2467 17.1 0.000054
```

Achtung, die Reihenfolge ist wichtig, **wenn man nur das grosse Modell eingibt** in `anova()`. Dazu das Modell ohne Interaktion mit je anderer Reihenfolge:

```
anova(lm(response ~ age + therapy, data = d.cat2))

## Analysis of Variance Table
##
## Response: response
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age        2 19964   9982   69.0 < 2e-16
## therapy    1  2467   2467   17.1 0.000054
## Residuals 196 28351    145

## Was wir wollen. Effekt von Therapy kontrolliert für Alter
```

```
anova(lm(response ~ therapy + age, data = d.cat2))

## Analysis of Variance Table
##
## Response: response
##           Df Sum Sq Mean Sq F value    Pr(>F)
## therapy     1  4117   4117   28.5 2.6e-07
## age         2 18313   9157   63.3 < 2e-16
## Residuals 196 28351    145

## Effekt von Therapy nicht kontrolliert. Effekt von Alter kontrolliert für Therapy
```

Die Resultate sind verschieden! Im ersten Modell ist der Effekt von `therapy` kontrolliert für `age`, im zweiten Modell ist es umgekehrt. `anova()` macht sequentielle Tests. Es ist immer am sichersten, Hypothesen über den expliziten Vergleich der entsprechenden Modelle zu machen. Wenn man nur das grosse Modell als Argument in `anova()` übergibt, kann man schnell falsche Schlüsse ziehen, wenn man vergisst, dass sequentiell getestet wird.

Da der Interaktionseffekt nicht signifikant ist, sollten wir auch das Haupteffektmodell brauchen, um den für Alter kontrollierten Therapieeffekt zu testen. Wenn wir den für Alter kontrollierten Therapieeffekt im Interaktionsmodell testen, sind die F - und p -Werte verschieden (wenn hier auch nur leicht), da man dann die Residualvarianz vom Interaktionsmodell braucht:

```
anova(lm(response ~ age * therapy, data = d.cat2))

## Analysis of Variance Table
##
## Response: response
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age        2 19964   9982   68.34 < 2e-16
## therapy    1  2467   2467   16.89 0.000059
## age:therapy 2    14      7   0.05     0.95
## Residuals 194 28337    146
```

Die Interaktion war aber nicht signifikant. Es ist daher angebracht, als Residualvarianz diejenige vom Haupteffektmodell zu nehmen, da bei Abwesenheit von Interaktion die Varianz, die zum Interaktionseffekt gehört dem Zufall zugeordnet werden sollte.

Wir haben gesehen, dass es keinen signifikanten `therapy:age` Interaktionseffekt auf die Zielgröße gibt, der Therapieeffekt ist also unabhängig von der Altersgruppe. Das war auch in der Abbildung 15.2 ersichtlich. Die Profile sind annähernd parallel. `age` bleibt aber vorerst im Modell als potentieller Confounder. Der für Alter kontrollierte Effekt von `therapy` (der über den Effekt von `age` hinausgehende Effekt) ist 7.109 (Punktschätzung):

```
summary(modelMain)$coef

##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.62      2.27   15.26 3.73e-35
## therapyTrt  7.11      1.72    4.13 5.38e-05
## ageyoung    17.22     2.73    6.31 1.83e-09
## ageold      27.39     2.49   10.98 3.55e-22
```

Der nicht für `age` kontrollierte Effekt von `therapy` ist 9.08 (Punktschätzung) und entspricht dem Unterschied der *gewichteten Durchschnitte* in Tabelle 15.3. Hier wird die Variable `age` in der Analyse ignoriert:

```
summary(modelTherapy)$coef

##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.91      1.54   34.94 1.29e-86
## therapyTrt  9.08      2.17    4.18 4.38e-05
```

was äquivalent wäre mit

```
t.test(response ~ therapy, var.equal = TRUE)
```

Kontraste. Kontraste sind – wie bereits gesehen – Linearkombinationen von Parametern mit der Bedingung, dass die Summe der Koeffizienten Null ist. Mit Kontrasten werden aus den geschätzten Parametern eines Modells die Funktionen berechnet, die die Quantitäten von Interesse repräsentieren.

Würde man das Interaktionsmodell beibehalten, kann man sich die Therapieeffekte innerhalb der Altersgruppen von `emmeans()` herausgeben, jetzt aber mit Konfidenzintervallen und Tests (Die Berechnung von Konfidenzintervallen von Kontrasten ist z.T. komplex, das macht `emmeans()` für uns).

```
emInt <- emmeans(modelInt, revpairwise ~ therapy | age) #revpairwise: damit Trt-Ctrl statt Ctrl-Trt gerechnet wird
summary(emInt, infer = c(TRUE, TRUE))

## $emmeans
```

```

## age = child:
##   therapy emmean   SE df lower.CL upper.CL t.ratio p.value
##   Ctrl      34.9 2.70 194     29.6     40.2 12.900 <.0001
##   Trt       41.2 3.82 194     33.7     48.7 10.800 <.0001
##
## age = young:
##   therapy emmean   SE df lower.CL upper.CL t.ratio p.value
##   Ctrl      52.1 2.42 194     47.4     56.9 21.600 <.0001
##   Trt       58.7 2.07 194     54.6     62.8 28.300 <.0001
##
## age = old:
##   therapy emmean   SE df lower.CL upper.CL t.ratio p.value
##   Ctrl      61.8 1.64 194     58.5     65.0 37.600 <.0001
##   Trt       69.3 1.60 194     66.2     72.5 43.300 <.0001
##
## Confidence level used: 0.95
##
## $contrasts
## age = child:
##   contrast estimate   SE df lower.CL upper.CL t.ratio p.value
##   Trt - Ctrl    6.31 4.68 194    -2.918     15.6    1.350 0.1790
##
## age = young:
##   contrast estimate   SE df lower.CL upper.CL t.ratio p.value
##   Trt - Ctrl    6.60 3.18 194     0.321     12.9    2.070 0.0395
##
## age = old:
##   contrast estimate   SE df lower.CL upper(CL) t.ratio p.value
##   Trt - Ctrl    7.56 2.30 194     3.038     12.1    3.300 0.0012
##
## Confidence level used: 0.95

```

Eine **Mittelung** dieser drei Effekte im Interaktionsmodell über alle Altersgruppen (indem wir das `|age` in `revpairwise` `therapy|age` weglassen) gäbe dann

```

emIntMean <- emmeans(modelInt, revpairwise ~ therapy)

## NOTE: Results may be misleading due to involvement in interactions

summary(emIntMean, infer = c(TRUE, TRUE))

## $emmeans
##   therapy emmean   SE df lower.CL upper(CL) t.ratio p.value
##   Ctrl      49.6 1.33 194     47.0     52.2 37.400 <.0001
##   Trt       56.4 1.54 194     53.4     59.5 36.500 <.0001
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95
##
## $contrasts
##   contrast estimate   SE df lower(CL) upper(CL) t.ratio p.value
##   Trt - Ctrl    6.83 2.04 194     2.81     10.8    3.350 0.0010
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95

```

Achtung: Wir sehen den Hinweis, dass dieses Resultat **irreführend** sein kann. Wir

haben gesehen, dass eine Interpretation der Haupteffekte in einem Interaktionsmodell dem Prinzip der Marginalität widerspricht.

Nun war aber der Interaktionseffekt nicht statistisch signifikant. Bei Abwesenheit von Interaktionseffekten sind dann die Therapieeffekte für alle Altersgruppen gleich. Das Resultat für den für `age` kontrollierten Effekt – jetzt im Haupteffektmodell – ist

```
emMain <- emmeans(modelMain, revpairwise ~ therapy)
summary(emMain, infer = c(TRUE, TRUE))

## $emmeans
##   therapy emmean    SE df lower.CL upper.CL t.ratio p.value
##   Ctrl      49.5 1.27 196      47     52.0 38.900 <.0001
##   Trt       56.6 1.33 196      54     59.2 42.500 <.0001
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95
##
## $contrasts
##   contrast   estimate    SE df lower.CL upper.CL t.ratio p.value
##   Trt - Ctrl    7.11 1.72 196     3.71    10.5  4.130  0.0001
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95
```

Das Resultat für den nicht für `age` kontrollierten Effekt – also im Modell mit nur `therapy` als Eingangsgröße – ist

```
emTher <- emmeans(modelTherapy, revpairwise ~ therapy)
summary(emTher, infer = c(TRUE, TRUE))

## $emmeans
##   therapy emmean    SE df lower.CL upper.CL t.ratio p.value
##   Ctrl      53.9 1.54 198     50.9     57 34.900 <.0001
##   Trt       63.0 1.53 198     60.0     66 41.200 <.0001
##
## Confidence level used: 0.95
##
## $contrasts
##   contrast   estimate    SE df lower.CL upper.CL t.ratio p.value
##   Trt - Ctrl   9.08 2.17 198     4.79    13.4  4.180 <.0001
##
## Confidence level used: 0.95
```

Balancierte versus nicht-balancierte Daten. Wir hatten oben von den verschiedenen Durchschnitten gesprochen, hier nochmals die Tabelle:

	child	young	old	unweighted.mean	weighted.mean
Ctrl	34.89	52.13	61.78	49.60	53.91
Trt	41.20	58.74	69.35	56.43	62.99

Zur Wiederholung:

- Die beiden Durchschnitte (je für `Ctrl` und `Trt`) in den beiden letzten Spalten sind nur gleich, wenn das Design balanciert ist, wenn die Altersgruppen gleich gross sind.
- Die Frage welche Durchschnitte (ungewichtet oder gewichtet) relevant sind – ist dasselbe wie die Frage nach der getesteten Hypothese.
- Wenn die Daten balanciert wären (was hier nicht der Fall ist), dann gäbe es keinen Unterschied zwischen dem für `age` kontrollierten Therapieeffekt und dem nicht für `age` kontrollierten Therapieeffekt.

15.3 Eine kategoriale und eine kontinuierliche Eingangsgröße

Im vorigen Kapitel hatten wir zwei kategoriale Eingangsgrößen im Modell. Ein lineares Modell kann aber auch kategoriale *und* kontinuierliche Eingangsgrößen haben.

Wir betrachten dazu direkt eine konkrete Situation. Im Folgenden haben wir eine kategoriale Eingangsgröße `group` (mit zwei Kategorien *A* und *B*) und eine kontinuierliche Eingangsgröße `height`. Die Zielgröße, die abhängige Variable, sei `weight`.

Ziel. Uns interessiert der Effekt von `group` auf `weight`, und wenn nötig, soll dieser für `height` kontrolliert werden.

Klassisch werden solche Situationen *Kovarianzanalysen* genannt.

15.3.1 Modell

Means-Parameterisierung. Die Beobachtung Y_i (`weight`) der Person i in der Gruppe j (`group`) mit Kovariablen x_i (`height`) schreiben wir dann

$$Y_i = \alpha_{j(i)} + \beta_{j(i)} x_i + \epsilon_i, \quad j = 1, 2, \quad i = 1, \dots, n. \quad (15.3.1)$$

In 15.3.1 sind die $\alpha_{j(i)}$ das Intercept und $\beta_{j(i)}$ die Steigung für den `height-weight` Zusammenhang in Gruppe j . x_i ist die Körpergröße der Person i und ϵ_i beschreibt den kombinierten Effekt aller unbekannten Einflüsse auf das Körpergewicht von Person i und wird dann meistens wieder als Zufallsvariable mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ angenommen. Das Modell hat also (neben σ^2) vier Parameter (2 Gruppen mal 2 Parameter). Wenn die Steigungen in den zwei Gruppen verschieden sind, haben wir einen Interaktionseffekt von `height` und `group` auf `weight`.

Effekt-Parameterisierung. Die Effekt-Parameterisierung für das Modell mit Interaktion ist

$$Y_i = \beta_1 + \beta_2 I_{B(i)} + \beta_3 x_i + \beta_4 x_i I_{B(i)} + \epsilon_i. \quad (15.3.2)$$

Auch dieses Modell hat vier Parameter (was sich ändert, ist die Parameterisierung, sonst ändert sich am Modell grundsätzlich nichts). Die ersten zwei Parameter sind analog zu einem einfaktoriellen Design (bei zwei Gruppen also ein Zweistichproben-*t*-Test). β_1 ist das erwartete Gewicht für eine Person in Gruppe *A* und Körpergrösse Null. β_2 ist der erwartete Unterschied im Gewicht für eine Person in Gruppe *B* relativ zu Gruppe *A* bei Körpergrösse Null. $I_{B(i)}$ ist eine Indikatorvariable für die Zugehörigkeit zu Gruppe *B*

$$I_{B(i)} = \begin{cases} 1 & \text{group } B \\ 0 & \text{group } \neq B. \end{cases}$$

β_3 ist die erwartete Steigung der Regression von `weight` auf `height` in Gruppe *A*. β_4 ist der erwartete Unterschied in den Steigungen in der Gruppe *B* relativ zur Gruppe *A*. Die Erwartungswerte in den zwei Gruppen sind

$$\begin{aligned} E(Y_i; \text{group} = A) &= \beta_1 + \beta_3 x_i \\ E(Y_i; \text{group} = B) &= \beta_1 + \beta_2 + \beta_3 x_i + \beta_4 x_i \end{aligned} \quad (15.3.3)$$

15.3.2 Beispiel

Simulation*. Folgender Code simuliert Daten aus dem beschriebenen Modell.

```
set.seed(10)
n.groups <- 2
n.sample <- 50
n <- n.groups * n.sample  ##sample size
ind <- rep(1:n.groups, each = n.sample)  ##Indicator for group
group <- factor(ind, labels = c("A", "B"))
height <- rnorm(n, mean = 165, sd = 11.4)
covariates <- data.frame(group, height)
Xeffects <- model.matrix(~group * height)
Xmeans <- model.matrix(~group * height - 1)
sigma <- 2
betaM <- c(muA <- -36.475, muB <- -45.5, slopeA <- 0.615, slopeB <- 0.7)  ##Means-Param.
betaE <- c(muA, muB - muA, slopeA, slopeB - slopeA)  ##Effekt-Parm
lin.pred <- Xeffects %*% betaE
lin.pred2 <- Xmeans %*% betaM
# all.equal(lin.pred, lin.pred2) ## ist dasselbe
eps <- rnorm(n = n, mean = 0, sd = sigma)  ## add noise
weight <- lin.pred + eps  ## Zielgrösse
d.catcont <- data.frame(group, height, weight)
```

Daten. Wir haben nun einen Datensatz `d.catcont` mit $n = 100$ Beobachtungen auf `weight` und mit unabhängigen Variablen `height` und `group` (zweiwertig)

```

str(d.catcont)

## 'data.frame': 100 obs. of  3 variables:
## $ group : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 ...
## $ height: num  165 163 149 158 168 ...
## $ weight: num  63.6 64.5 53.3 62.2 65.8 ...

head(d.catcont)

##   group height weight
## 1     A    165   63.6
## 2     A    163   64.5
## 3     A    149   53.3
## 4     A    158   62.2
## 5     A    168   65.8
## 6     A    169   68.9

```

Beschreibende Statistik.

Zuerst die beschreibende Statistik:

```

psych::describe(d.catcont[, -1])

##           vars n  mean   sd median trimmed   mad   min   max range skew kurtosis   se
## height      1 100 163.4 10.73 162.8   163.4 12.13 140.1 190.3 50.2 0.03   -0.56 1.07
## weight      2 100  66.4  8.08  65.8    66.2  8.17  47.8  88.3 40.5 0.22   -0.26 0.81

by(d.catcont[, -1], d.catcont$group, psych::describe)

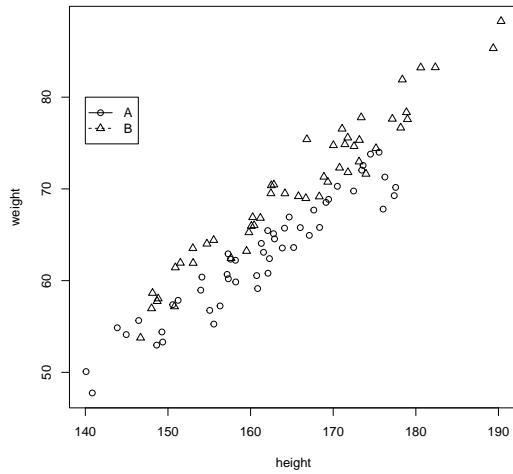
## d.catcont$group: A
##           vars n  mean   sd median trimmed   mad   min   max range skew kurtosis   se
## height      1 50 161.1 9.88 162    161.4 9.86 140.1 178 37.5 -0.2   -0.75 1.40
## weight      2 50  62.7 6.32  63    62.8 7.02  47.8  74 26.2 -0.2   -0.70 0.89
## -----
## d.catcont$group: B
##           vars n  mean   sd median trimmed   mad   min   max range skew kurtosis   se
## height      1 50 165.8 11.13 166.8   166 10.50 146.7 190.3 43.6 0.06   -0.81 1.57
## weight      2 50  70.1  7.96  70.4    70  8.28  53.8  88.3 34.5 0.06   -0.61 1.13

```

```

plot(weight ~ height, data = d.catcont, pch = as.numeric(group))
legend(140, 80, legend = levels(group), lty = c(1, 2), pch = c(1, 2))

```



Modell anpassen. Die Größen β_1 und $\beta_1 + \beta_2$ stellen das Intercept dar für Gruppe A beziehungsweise Gruppe B , das erwartete Gewicht bei Körpergrösse Null. Damit das Intercept besser interpretierbar ist, macht es Sinn, die Variable `height` vorab zu zentrieren. Das heisst, dass man von jedem Wert den Mittelwert subtrahiert. Das wird nur eine Auswirkung haben auf die Interpretation des Intercept.

```
d.catcont$heightCent <- scale(d.catcont$height, scale = FALSE)
```

Wir passen das Modell an, zuerst mit nicht-, dann mit zentrierter Körpergrösse. Wir sehen, dass sich nur die geschätzten Intercepts unterscheiden, nicht aber die Steigungen.

```
modraw <- lm(weight ~ group * height, d.catcont)
summary(modraw)$coef

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.5185    4.5592  -7.79 7.86e-12
## groupB       -9.3987    6.1742  -1.52 1.31e-01
## height        0.6094    0.0282  21.57 1.36e-38
## groupB:height  0.0845    0.0378   2.24 2.75e-02
```

```
mod <- lm(weight ~ group * heightCent, d.catcont)
summary(mod)$coef

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0776    0.2839 225.67 1.22e-132
## groupB       4.4191    0.4004  11.04 9.01e-19
## heightCent    0.6094    0.0282  21.57 1.36e-38
## groupB:heightCent  0.0845    0.0378   2.24 2.75e-02
```

In der Abbildung 15.3 sind die Daten *und* das Modell dargestellt.

```

plot(weight ~ height, data = d.catcont, pch = as.numeric(group))
legend(140, 80, legend = levels(group), lty = c(1, 2), pch = c(1, 2))
abline(a = coef(moddraw)[1], b = coef(moddraw)[3], lty = 1)
abline(a = coef(moddraw)[1] + coef(moddraw)[2], b = coef(moddraw)[3] + coef(moddraw)[4], lty = 2)

plot(weight ~ heightCent, data = d.catcont, pch = as.numeric(group))
legend(-20, 80, legend = levels(group), lty = c(1, 2), pch = c(1, 2))
abline(a = coef(mod)[1], b = coef(mod)[3], lty = 1)
abline(a = coef(mod)[1] + coef(mod)[2], b = coef(mod)[3] + coef(mod)[4], lty = 2)

```

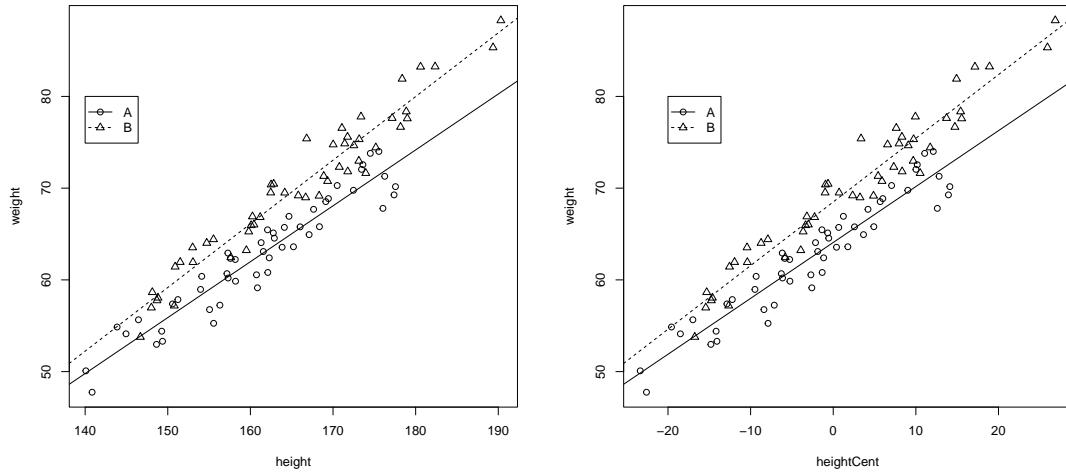


Abbildung 15.3: Angepasstes Modell. Rechts: mit zentrierter Körpergröße

Interpretation.

- $\hat{\beta}_1 = 64.078$ ist jetzt die Punktschätzung für das erwartete Gewicht bei Durchschnittsgröße für eine Person in Gruppe A.
- $\hat{\beta}_2 = 4.419$ ist die Punktschätzung für die Differenz im erwarteten Gewicht bei Durchschnittsgröße zwischen Gruppe B und Gruppe A.
- $\hat{\beta}_3 = 0.609$ ist die Punktschätzung für die erwartete Steigung für eine Person in Gruppe A
- $\hat{\beta}_4 = 0.085$ ist die Punktschätzung für den Unterschied in der erwarteten Steigung für eine Person aus der Gruppe B relativ zu einer Person in der Gruppe A.

Hypothesen testen. Wir wollen jetzt testen, ob der Interaktionseffekt signifikant ist. Wir sehen das eigentlich schon im `summary()`-Output, da ja t-Tests bezüglich individuellen Parametern immer den Effekt quantifizieren, der über den Effekt der

anderen Variablen im Modell hinausgeht. Wir machen aber diesen Test wieder explizit über einen Modellvergleich.

```
modMain <- update(mod, . ~ . - group:heightCent)
anova(modMain, mod)

## Analysis of Variance Table
##
## Model 1: weight ~ group + heightCent
## Model 2: weight ~ group * heightCent
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1     97 385
## 2     96 366  1      19.1 5.01  0.027
```

oder mit `drop1()`

```
drop1(mod, test = "F")

## Single term deletions
##
## Model:
## weight ~ group * heightCent
##                   Df Sum of Sq RSS AIC F value    Pr(>F)
## <none>                 366 138
## group:heightCent 1     19.1 385 141     5.01  0.027
```

Das heisst, der Effekt von `group` auf `weight` hängt vom Wert auf `heightCent` ab (oder, was quantitativ äquivalent ist, aber nicht unsere Frage: Der Effekt von `heightCent` auf `weight` hängt von `group` ab). Das Modell ohne Interaktionseffekt wird verworfen.

Der Interaktionseffekt ist damit signifikant. Es gibt nun Varianzanalysen mit sogenannten *Type III*-Quadratsummen, die Haupteffekte testen, **kontrolliert für Interaktionseffekte**. Das sind “durchschnittliche Haupteffekte”, wie wir sie schon weiter oben besprochen haben. Das sind nicht-sequentielle Tests, die umstritten sind und zu viel Verwirrung führen, da sie Hypothesen testen, die eigentlich nicht von Interesse sind. Wir folgen weiter dem Prinzip der Marginalität und gehen nicht weiter darauf ein.

Wir können aber immer noch die Frage stellen, ob der Effekt von `group`, kontrolliert für `heightCent`, signifikant ist. Dazu müssen wir aber – im Gegensatz zum Beispiel in vorigem Abschnitt – beim Interaktionsmodell bleiben, da ja dieser signifikant ist. Da `anova()` sequentielle Tests, müssen wir die Reihenfolge beachten und `group` nach `heightCent` eingeben.

```
anova(lm(weight ~ heightCent * group, d.catcont))

## Analysis of Variance Table
##
## Response: weight
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## heightCent        1   5612   5612 1471.39 <2e-16
```

```
## group           1     460      460  120.59 <2e-16
## heightCent:group 1      19       19    5.01  0.027
## Residuals       96     366       4
```

Nicht angemessen wäre, aus den Gründen, die beim zweifaktoriellen Design erläutert wurden:

```
anova(lm(weight ~ heightCent + group, d.catcont))

## Analysis of Variance Table
##
## Response: weight
##             Df Sum Sq Mean Sq F value Pr(>F)
## heightCent  1   5612   5612   1413 <2e-16
## group       1     460     460    116 <2e-16
## Residuals  97    385      4
```

Wir sehen, dass die F -Werte von `group` verschieden sind (wenn auch minimal und hier ohne Konsequenzen). Der Grund ist, dass in den beiden Fällen eine andere Residualvarianz gebraucht wird.

Konfidenzintervalle für Kontraste von Interesse. Zu den Intervallschätzungen für die Parameter kommen wir mit `confint()`:

```
cbind(summary(mod)$coef, confint(mod)) #ein beliebter Output

##               Estimate Std. Error t value Pr(>|t|)    2.5 % 97.5 %
## (Intercept) 64.0776   0.2839 225.67 1.22e-132 63.51394 64.641
## groupB      4.4191   0.4004 11.04  9.01e-19  3.62425  5.214
## heightCent   0.6094   0.0282 21.57  1.36e-38  0.55329  0.665
## groupB:heightCent 0.0845   0.0378  2.24  2.75e-02  0.00959  0.159
```

Damit wir auch eine Intervallschätzung für die Steigung in der Gruppe B haben (Punktschätzung $0.609+0.085$), können wir die Funktion `emtrends()` aus dem Packet `emmeans` brauchen:

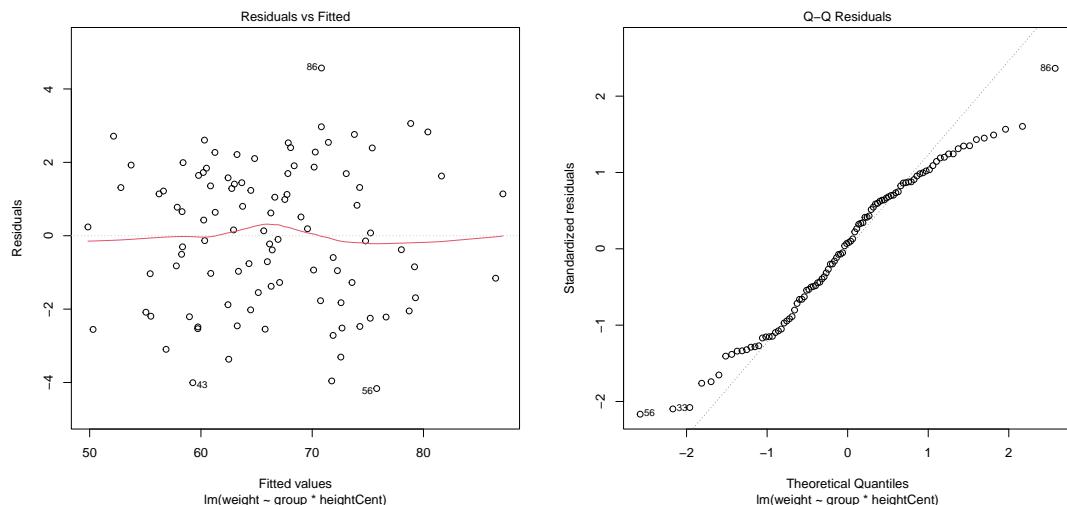
```
emt <- emtrends(mod, rvpairwise ~ group, var = "heightCent", infer = c(TRUE, TRUE))
emt

## $emtrends
##   group heightCent.trend     SE df lower.CL upper.CL t.ratio p.value
##   A          0.609 0.0283 96    0.553    0.665  21.570 <.0001
##   B          0.694 0.0251 96    0.644    0.744  27.690 <.0001
##
## Confidence level used: 0.95
##
## $contrasts
##   contrast estimate     SE df lower.CL upper.CL t.ratio p.value
##   B - A      0.0845 0.0378 96    0.00959     0.16    2.239  0.0275
##
## Confidence level used: 0.95
```

Vergleichen wir diesen Output mit dem `summary()`-output: Wir haben hier dieselbe Information wie im `summary()`-Output, eine Schätzung bezüglich der Steigung in der Gruppe A, eine Schätzung des Unterschieds der Steigungen in Gruppe B versus A, aber zusätzlich jetzt noch eine Schätzung für die Steigung in der Gruppe B.

Residuenanalyse. Der TA-plot sieht gut aus (weil man ja aus einem Modell mit entsprechenden Annahmen simuliert hat). Der Quantil-Quantil-Plot sieht nicht perfekt aus.

```
plot(mod, which = c(1, 2))
```



Ein statistischer Test würde hier sogar Normalverteilung – zu Unrecht – verwerfen:

```
shapiro.test(resid(mod))

##
## Shapiro-Wilk normality test
##
## data: resid(mod)
## W = 1, p-value = 0.04
```

n ist aber gross genug, damit wir immer noch eine approximative Normalverteilung für die Teststatistiken voraussetzen könnten (denn in Wahrheit würden wir natürlich nicht wissen, dass die Fehler wirklich – wie in diesen aus dem Modell simulierten Daten – normalverteilt sind). Streng genommen müssten wir dann die p -Werte bezüglich normalverteilter statt t -verteilter Teststatistik berechnen.

Vorhersage. Wir haben bis jetzt vor allem Parameter geschätzt und Hypothesen diesbezüglich getestet. In der Wissenschaft will man aber häufig ein angepasstes

Modell dahingehend benutzen, um mit *neuen Werten* auf den Prädiktoren neue Werte auf der abhängigen Variablen vorherzusagen.

Das ist in R implementiert mit der `predict()`-Funktion. Es gibt zwei Arten von Vorhersagen, *mittlere* Vorhersagen oder *individuelle* Vorhersagen. Achtung: Es geht jetzt nicht mehr um Unsicherheit(en) für einzelne Parameter, sondern für erwartete oder individuelle Y 's.

Wir möchten nun mit unserem angepassten Modell `mod` für ausgewählte Körpergrößen und Gruppe neue Beobachtungen vorhersagen.

Dazu generieren wir ein data frame mit den *neuen* Eingangsgrößen, für die wir eine Vorhersage gemäss Modell haben möchten:

```
new <- data.frame(group = c("A", "B", "A"), height = c(170, 180, 190))
new

##   group height
## 1      A     170
## 2      B     180
## 3      A     190
```

Punktvorhersagen wären dann zu haben mit

```
predict(moddraw, newdata = new)

##    1    2    3
## 68.1 80.0 80.3
```

Wir wollen aber natürlich die Unsicherheit der Vorhersage einbeziehen. Zuerst machen wir eine Vorhersage für eine *individuelle* Beobachtung $Y_{new} | X_{new} = x_{new}$. Das machen wir mit dem Argument `intervall="prediction"`:

```
pred <- predict(moddraw, newdata = new, interval = "prediction")
cbind(new, pred)

##   group height fit  lwr  upr
## 1      A     170 68.1 64.1 72.0
## 2      B     180 80.0 76.0 84.0
## 3      A     190 80.3 76.0 84.5
```

Jetzt machen wir eine Vorhersage für eine durchschnittliche Beobachtung, also für den Erwartungswert $E(Y_{new} | X_{new} = x_{new})$. Das machen wir mit dem Argument `intervall="confidence"`:

```
pred2 <- predict(moddraw, newdata = new, interval = "confidence")
cbind(new, pred2)

##   group height fit  lwr  upr
## 1      A     170 68.1 67.3 68.8
## 2      B     180 80.0 79.1 80.9
## 3      A     190 80.3 78.6 82.0
```

Die Unsicherheit für die Vorhersage einer individuellen Beobachtung ist immer grösser als die Unsicherheit für die Vorhersage eines Erwartungswertes (des linearen Prädiktors). Im Gegensatz zu letzterem kommt bei der individuellen Vorhersage immer noch die Fehlervarianz σ^2 dazu. Der Unterschied zwischen den Konfidenzgrenzen für den Erwartungswert bei gegebenen Prädiktorwerten versus einem Vorhersageintervall ist visualisiert in <https://rstudio.zhaw.ch/rsconnect/content/84>.

Wir können für unser angepasstes Modell Konfidenz- und Vorhersageintervalle für alle Körpergrössen und jede Gruppe graphisch darstellen (Abbildung 15.4, Code für die Interessierten).

```

pred.frame <- data.frame(group = "A", height = seq(min(d.catcont$height), max(d.catcont$height)))
pc <- data.frame(predict(moddraw, newdata = pred.frame, interval = "confidence"))
pp <- data.frame(predict(moddraw, newdata = pred.frame, interval = "prediction"))
plot(pred.frame$height, pc[, 1], col = 2, type = "l", xlab = "height", ylab = "predicted weight", main = "Group A")
points(height[group == "A"], weight[group == "A"])
lines(pred.frame$height, pc[, 2])
lines(pred.frame$height, pc[, 3])
lines(pred.frame$height, pp[, 2], lty = 2)
lines(pred.frame$height, pp[, 3], lty = 2)
grid()

pred.frame <- data.frame(group = "B", height = seq(min(d.catcont$height), max(d.catcont$height)))
pc <- data.frame(predict(moddraw, newdata = pred.frame, interval = "confidence"))
pp <- data.frame(predict(moddraw, newdata = pred.frame, interval = "prediction"))
plot(pred.frame$height, pc[, 1], col = 2, type = "l", xlab = "height", ylab = "predicted weight", main = "Group B")
points(height[group == "B"], weight[group == "B"])
lines(pred.frame$height, pc[, 2])
lines(pred.frame$height, pc[, 3])
lines(pred.frame$height, pp[, 2], lty = 2)
lines(pred.frame$height, pp[, 3], lty = 2)
grid()

```

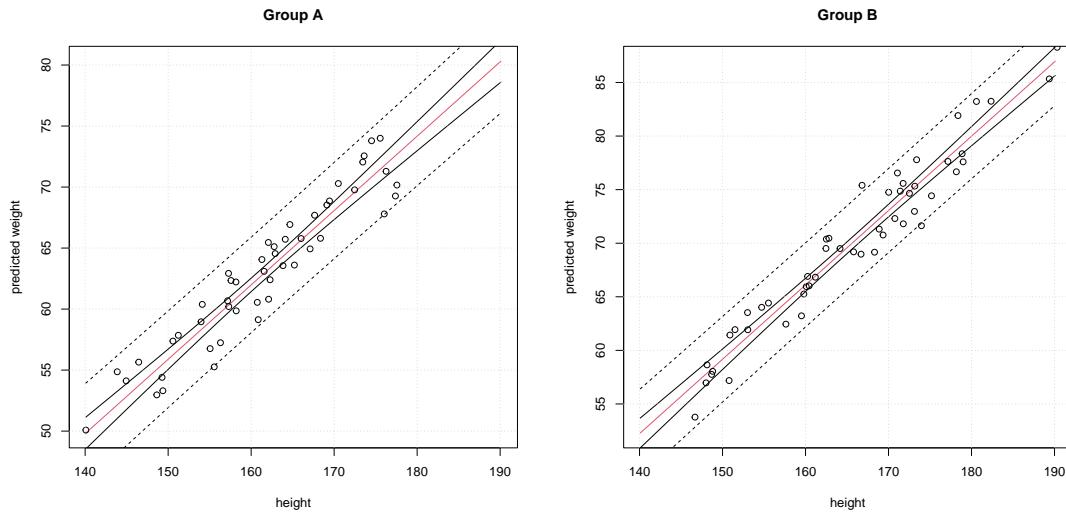
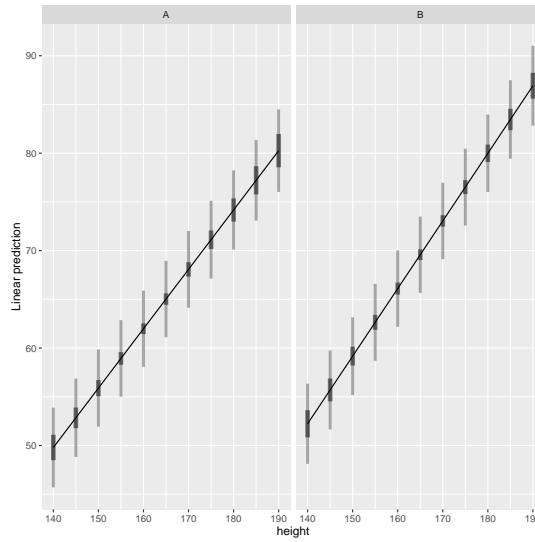


Abbildung 15.4: Vorhersage: 95% Konfidenzgrenzen für Erwartungswerte (durchgezogen) und 95% Vorhersagegrenzen für individuelle Beobachtungen (gestrichelt).

Mit ein bisschen weniger Aufwand kann man mit der Funktion `emmpip()` aus dem Paket `emmeans` ebenfalls graphisch Vorhersagen mit Konfidenz- und Vorhersageintervallen darstellen:

```
emmpip(modraw, group~height|group, cov.reduce=function(x){seq(140,190,by=5)}, CIs=TRUE, PIs=TRUE)
```



Mit `plotit=FALSE` bekommt man die Konfidenz- und Vorhersageintervalle numerisch:

```

emmmip(modraw,group~height,cov.reduce=function(x){seq(140,190,by=5)},CIs=TRUE,PIs=TRUE,plotit=FALSE)

##   group height yvar      SE df  LCL  UCL  LPL  UPL tvar xvar
##   A       140 49.8 0.657 96 48.5 51.1 45.7 53.9 A     140
##   B       140 52.2 0.703 96 50.8 53.6 48.1 56.3 B     140
##   A       145 52.8 0.532 96 51.8 53.9 48.8 56.9 A     145
##   B       145 55.7 0.589 96 54.5 56.9 51.6 59.7 B     145
##   A       150 55.9 0.418 96 55.1 56.7 51.9 59.9 A     150
##   B       150 59.2 0.482 96 58.2 60.1 55.2 63.2 B     150
##   A       155 58.9 0.326 96 58.3 59.6 55.0 62.9 A     155
##   B       155 62.6 0.386 96 61.9 63.4 58.7 66.6 B     155
##   A       160 62.0 0.278 96 61.4 62.5 58.1 65.9 A     160
##   B       160 66.1 0.312 96 65.5 66.7 62.2 70.0 B     160
##   A       165 65.0 0.297 96 64.4 65.6 61.1 68.9 A     165
##   B       165 69.6 0.277 96 69.0 70.1 65.7 73.5 B     165
##   A       170 68.1 0.373 96 67.3 68.8 64.1 72.0 A     170
##   B       170 73.0 0.296 96 72.5 73.6 69.1 77.0 B     170
##   A       175 71.1 0.480 96 70.2 72.1 67.1 75.1 A     175
##   B       175 76.5 0.360 96 75.8 77.2 72.6 80.5 B     175
##   A       180 74.2 0.601 96 73.0 75.4 70.1 78.2 A     180
##   B       180 80.0 0.451 96 79.1 80.9 76.0 84.0 B     180
##   A       185 77.2 0.729 96 75.8 78.7 73.1 81.4 A     185
##   B       185 83.5 0.555 96 82.4 84.6 79.4 87.5 B     185
##   A       190 80.3 0.862 96 78.6 82.0 76.0 84.5 A     190
##   B       190 86.9 0.667 96 85.6 88.2 82.8 91.0 B     190
##
##   ## Confidence level used: 0.95

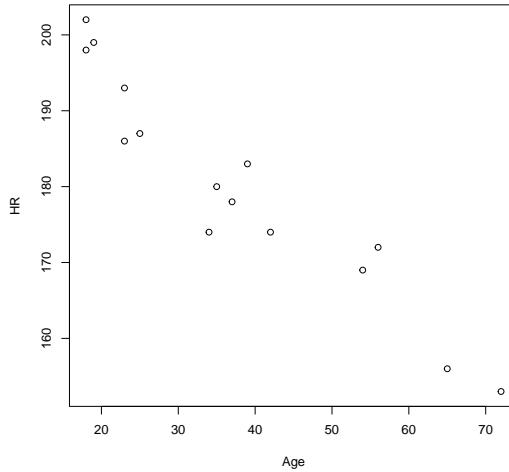
```

Wichtigkeit der Unsicherheit der Vorhersage. Die Unsicherheit von Vorhersagen kann sehr gross sein und wird oft ungenügend beachtet. Als Beispiel nehmen wir das Problem der Vorhersage von Maximaler Herzfrequenz (HR) durch das Alter. Nehmen wir an, eine kleine “Studie” mit Werten auf Age und HR gäbe die Daten:

```
Age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
HR <- c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183, 178)
```

Wir passen das lineare Modell an:

```
modHR <- lm(HR ~ Age)
plot(Age, HR)
```

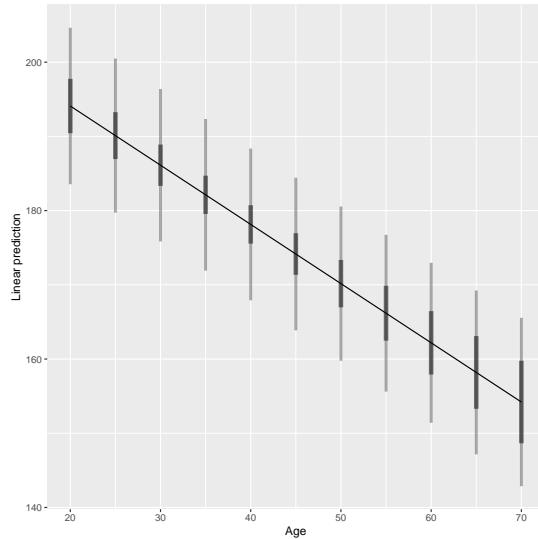


```
summary(modHR)

##
## Call:
## lm(formula = HR ~ Age)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -8.926 -2.538  0.388  3.187  6.624 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 210.048    2.867   73.3 < 2e-16 ***
## Age         -0.798     0.070  -11.4  3.8e-08 ***
## 
## Residual standard error: 4.58 on 13 degrees of freedom
## Multiple R-squared:  0.909, Adjusted R-squared:  0.902 
## F-statistic: 130 on 1 and 13 DF,  p-value: 3.85e-08
```

Wir wollen nun die Vorhersage gemäss Modell graphisch darstellen:

```
emmmip(modHR, ~Age, cov.reduce=function(x){seq(20,70,5)}, PIs=TRUE, CIs=TRUE)
```



```
emmmip(modHR, ~Age, cov.reduce=function(x){seq(20,70,5)}, PIs=TRUE, CIs=TRUE, plotit=FALSE)
```

```
##   Age yvar    SE df LCL UCL LPL UPL tvar xvar
## 20 194 1.69 13 190 198 184 205 1     20
## 25 190 1.46 13 187 193 180 200 1     25
## 30 186 1.29 13 183 189 176 196 1     30
## 35 182 1.19 13 180 185 172 192 1     35
## 40 178 1.20 13 176 181 168 188 1     40
## 45 174 1.30 13 171 177 164 184 1     45
## 50 170 1.48 13 167 173 160 181 1     50
## 55 166 1.71 13 162 170 156 177 1     55
## 60 162 1.98 13 158 166 151 173 1     60
## 65 158 2.27 13 153 163 147 169 1     65
## 70 154 2.57 13 149 160 143 166 1     70
##
## Confidence level used: 0.95
```

Man sieht schön, dass einfache Formeln wie 220-Alter usw. für die meisten Menschen nicht funktionieren. Neben der Unsicherheit der Schätzung der Regressionsgerade (dargestellt durch die dicken Linien) fliesst bei der Unsicherheit der individuellen Vorhersage noch die Unsicherheit einer einzelnen Beobachtung hinein, die Standardabweichung der Fehler, 4.578 Herzschläge.

Kapitel 16

Generalisierte Lineare Modelle (GLM)*

Wenn wir diskrete oder kategoriale Größen modellieren wollen, ist das allgemeine lineare Modell nicht geeignet, da dieses eine kontinuierliche abhängige Variable voraussetzt. Oft haben wir zum Beispiel eine zweiseitige kategoriale abhängige Variable (mit den Werten Event und kein Event). Gibt es eine Möglichkeit, diesen Wertebereich auf ein Kontinuum abzubilden? Kann man das Allgemeine Lineare Modell (LM) verallgemeinern und damit Modelle für diskrete und kontinuierliche abhängige Variable in einer einzigen statistischen Methodologie zusammenführen?

Die Antwort darauf hat den Namen *Generalisierte Lineare Modelle (GLM)* (also nicht zu verwechseln mit dem Allgemeinen Linearen Modell (LM), das wir in Kapitel ?? eingeführt haben). Dazu gehören zum Beispiel die *Poisson-Regression* und die *logistische Regression*, aber eben auch die bereits eingeführte multiple Regression und Varianzanalyse als Vertreter der Linearen Modelle.

16.1 Allgemeiner Fall

Generalisierte lineare Modelle (GLM) stellen eine Verallgemeinerung dar der Theorie von linearen Modellen. Diese Generalisierung findet auf drei Arten statt:

1. *Systematischer Teil:* Der *Erwartungswert* der Zielgröße Y_i , also $\mu_i = E(Y_i)$, wird mit einer *Linkfunktion* geeignet transformiert; und diese Größe ist dann eine lineare Funktion der Parameter β_j , genannt der *lineare Prädiktor* mit der Link-Funktion $h(\cdot)$,

$$h(E(Y_i)) = h(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (16.1.1)$$

2. *Zufälliger Teil:* Die Varianz $Var(Y_i)$ ist eine Funktion des Erwartungswertes,

$$Var(Y_i) = v_i = \phi v(\mu_i), \quad (16.1.2)$$

wobei $v(\cdot)$ eine *Varianzfunktion* und ϕ einen *Dispersionsparameter* darstellt, der geschätzt werden muss oder nicht.

3. Jede Klasse eines GLM's folgt einem Modell mit der Dichte aus der *exponentiellen*

Familie. Für die Interessierten: Verteilungen der Exponentialfamilie haben Dichte $f(y_i) = \exp\{(y_i\theta_i - A(\theta_i))/\phi + B(y_i, \phi)\}$. Dabei ist θ_i der *kanonische Parameter*, θ_i kann durch μ_i ausgedrückt werden. Man kann zeigen, dass $E(Y_i) = \mu_i = A'(\theta_i)$ und $\text{Var}(Y_i) = A''(\theta_i)\phi$. Die Funktion $A(\cdot)$ legt fest, um welche Exponentialfamilie es sich handelt. Die Funktion $B(\cdot)$ ist eine Normierungsfunktion.

Die *Normalverteilung*, die *Binomial-* und die *Poissonverteilung* sind drei häufige Spezialfälle der Klasse der Exponentialfamilien. Diese Fälle werden im Folgenden dargestellt.

16.2 Spezialfälle

Normalverteilung, *Bernoulli*-, *Binomial*- und *Poissonverteilung* sind Spezialfälle der Klasse der Exponentialfamilien. Diese führen zur bereits bekannten multiplen linearen Regression, zur Poisson-Regression und zur logistischen Regression.

16.2.1 Lineares Modell

Wir haben bereits lineare Modelle in Kapitel ?? eingeführt. Dort haben wir normalverteilte Fehler vorausgesetzt, $Y_i \sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$. Das lineare Regressionsmodell kann geschrieben werden als $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$. Der Erwartungswert μ_i von Y_i wird modelliert mit

$$\mu_i = E(Y_i) = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (16.2.1)$$

Das bedeutet, dass die *Linkfunktion* $h(\cdot)$ die Identität ist und die Varianzfunktion ist $v(\mu_i) = 1$, zudem ist uns der Dispersionparameter bereits bekannt, $\phi = \sigma^2$. Das Allgemeine Lineare Modell (LM) ist also ein einfacher Spezialfall eines GLM.

Interpretation. In einem linearen Modell ist β_j offensichtlich der Unterschied der Erwartungswerte zwischen zwei Subpopulationen, die sich auf x_j um *eine Einheit* unterscheiden. Diese Grösse kann also als Steigung interpretiert werden.

16.2.2 Logistische Regression

Bei zweiwertiger dichotomer abhängiger Variable ist ein sehr beliebtes Modell das *logistische* Regressionsmodell, z.B. bei retrospektiven *Case-Control*-Studien. Dieses Modell ist sehr häufig anzutreffen in der angewandten Forschung.

Die Verteilung der Y_i ist in diesem Fall *binomial*¹, also $Y_i \sim \text{Bin}(\mu_i = \pi_i, n = 1)$ (??) und das Modell für den Erwartungswert $\mu_i = \pi_i$ wird geschrieben als

$$\text{logit}(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (16.2.2)$$

¹Oder *Bernoulli*, das ist der Spezialfall einer Binomialverteilung für Parameter $n = 1$.

für die Linkfunktion haben wir dann $h(\pi_i) = \text{logit}(\pi_i) = \log(\pi_i/(1 - \pi_i)) = \log \text{odds}$, die Varianzfunktion ist $v(\pi) = \pi(1 - \pi)$ und $\phi = 1$.

Interpretation. In einem logistischen Regressionmodell ist β_j der Unterschied der logarithmierten Chance zwischen zwei Subpopulationen, die sich auf x_j um *eine Einheit* unterscheiden; $\exp(\beta_j)$ (ausser beim Intercept) kann interpretiert werden als Chancenverhältnis, *odds ratio OR*, welches ein häufiges Effektmass bei Studien im Fall-Kontroll Design darstellt.

Beweis. $\beta_j = \log \text{odds}_{x_j} - \log \text{odds}_{x_j-1} = \log \frac{\text{odds}_{x_j}}{\text{odds}_{x_j-1}} = \log OR$, also $\exp(\beta_j) = OR$. \square

16.2.3 Poisson Regression

Die Poisson-Regression kommt oft bei der Modellierung von (positiven) *Zähldaten*, z.B. in prospektiven *Kohortenstudien* zum Zuge. Die Verteilung von Y_i ist dann *Poisson* $Y_i \sim \text{Pois}(\mu_i)$ (??) mit Erwartungswert $\mu_i = \lambda_i T$ mit λ_i als der Rate und T als der Beobachtungsdauer (z.B. Personenjahre).

Das Modell für die Rate λ_i wird dann geschrieben

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \log(T), \quad (16.2.3)$$

man modelliert also den Logarithmus der *Rate*, $\log \frac{\mu_i}{T} = \log(\mu_i) - \log(T)$.

Wenn $T = n$ ist, also wenn die Beobachtungsdauer für alle statistischen Einheiten eine Einheitsdauer ist (n „Personeneinheiten“), modelliert man den Logarithmus vom *Risiko*, $\log \frac{\mu_i}{n} = \log(\mu_i) - \log(n)$.

Die Linkfunktion ist also der Logarithmus, $h(\mu_i) = \log(\mu_i)$, der positive Zahlen auf den Bereich der reellen Zahlen abbildet (analog der logit-Funktion bei dichotomen Daten in der logistischen Regression), und die Varianzfunktion ist $v(\mu_i) = \mu_i$ und $\phi = 1$.

Interpretation. In Poisson-Modellen ist β_j der Unterschied in der logarithmierten Rate (im logarithmierten Risiko) zwischen zwei Subpopulationen, die sich auf x_j um *eine Einheit* unterscheiden; $\exp(\beta_j)$ (ausser beim Intercept) kann in diesem Falle interpretiert werden als Risiko- oder Ratenverhältnis *RR*, welches ein häufiges Effektmass in Kohortenstudien darstellt.

Beweis. $\beta_j = \log \lambda_{x_j} - \log \lambda_{x_j-1} = \log \frac{\lambda_{x_j}}{\lambda_{x_j-1}} = \log RR$, also $\exp(\beta_j) = RR$. \square

Kapitel 17

Gemischte Modelle (LMM, GLMM)*

Die Modelle, die bis jetzt betrachtet wurden, gingen meistens davon aus, dass Beobachtungen unabhängig voneinander waren (unabhängige ϵ_i , $i = 1, \dots, n$ und damit unabhängige Y_i). Oft sind aber Beobachtungen nicht unabhängig voneinander, dann muss der *Korrelation* Rechnung getragen werden.

17.1 Korrelierte Daten

Wenn Daten aus Beobachtungen bestehen, die nicht unabhängig voneinander sind (also der erste Aspekt von i.i.d.¹ nicht gegeben ist), so z.B. bei wiederholten Messungen an Personen, oder wenn Beobachtungen in Clustern auftreten (Patienten aus verschiedenen Spitätern), dürfen diese Beobachtungen natürlich nicht als unabhängige Informationen behandelt werden. Die Messungen *innerhalb* einer Person (oder innerhalb eines Clusters) werden ähnlicher sein als die Messungen *zwischen* Personen (zwischen Clustern). Die Korrelation der Beobachtungen innerhalb einer Person muss also berücksichtigt werden. Im Extremfall korrelieren Beobachtungen an einem Probanden perfekt, was bedeuten würde, dass diese nur eine Information „wert“ sind; alle anderen Beobachtungen also redundant wären. Wird die Abhängigkeit von Beobachtungen nicht berücksichtigt, überschätzt man z.T. massiv die Präzision von Statistiken (oder – äquivalent – unterschätzt die Standardfehler). Man wendet dann – fälschlicherweise – im Falle der einfachen Schätzung eines Durchschnitts – (7.1.6) an, man teilt dann also durch ein zu grosses n . Im Kapitel ?? haben wir eine solche Situation bereits betrachtet. Dort wurden mehrere Beobachtungen pro statistische Einheit modelliert, die durch eine *within-subject*-Variable, z.B. den Zeitpunkt, definiert wurden.

17.2 Notation für wiederholte Messungen

Im Folgenden stehen Grossbuchstaben für Zufallsvariablen und Matrizen, Kleinbuchstaben für Beobachtungen. Vektoren werden fett notiert, Skalare und

¹Independent and identically distributed.

Matrizen normal.

Es sei Y_{ij} das *Outcome* (die abhängige Variable), und \mathbf{x}_{ij} der Vektor der Länge p mit den *Eingangsgrößen*, die zum Zeitpunkt t_{ij} beobachtet wurden, für j -te Beobachtung $j = 1, \dots, n_i$ auf der i -ten Person $i = 1, \dots, m$. Der (wahre) Durchschnitt von Y_{ij} wird notiert mit $E(Y_{ij}) = \mu_{ij}$ und die Varianz von Y_{ij} wird notiert mit $\text{Var}(Y_{ij}) = v_{ij}$.

$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ stellt den Vektor von wiederholten Messungen bei Person i dar, mit Durchschnitt $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ und mit der $n_i \times n_i$ Kovarianzmatrix $\Sigma_i = \text{Var}(\mathbf{Y}_i)$, wobei das j -kte Element die Kovarianz zwischen Y_{ij} und Y_{ik} darstellt, notiert mit $\text{Cov}(Y_{ij}, Y_{ik}) = v_{ijk}$. R_i ist die $n_i \times n_i$ Korrelationsmatrix von \mathbf{Y}_i .

$\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_m)^T$ ist der Vektor aller Beobachtungen für alle statistischen Einheiten, mit Länge $N = \sum_{i=1}^m n_i$. In longitudinalen Studien stellt nun die Sequenz \mathbf{Y}_i die statistische Einheit dar, das heisst, Replikation bezieht sich auf die Anzahl Personen m , nicht auf die Anzahl Messungen N .

Ein lineares Modell für Beobachtung Y_{ij} kann man schreiben als

$$Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \dots + \beta_p x_{ijp} + \epsilon_{ij} \quad i = 1, \dots, m, \quad j = 1, \dots, n_i, \quad (17.2.1)$$

oder in Vektornotation

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij} \quad i = 1, \dots, m, \quad j = 1, \dots, n_i, \quad (17.2.2)$$

wobei $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ den p -Vektor der unbekannten Parameter und ϵ_{ij} eine Zufallsgröße darstellt mit um Null verteilten Fehler um die Vorhersage $\mathbf{x}_{ij}^T \boldsymbol{\beta}$.

Für die i -te Person nimmt das die Form an

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad i = 1, \dots, m,$$

mit X_i als der $n_i \times p$ Matrix mit \mathbf{x}_{ij} in der j -ten Zeile und $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^T$. Wenn wir den Vektor \mathbf{Y} der Länge $N = \sum_{i=1}^m n_i$ brauchen, können wir das kompakt schreiben als

$$\mathbf{Y} = X \boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (17.2.3)$$

dabei haben \mathbf{Y} und $\boldsymbol{\epsilon}$ die Längen N und X ist eine $N \times p$ Matrix. Im Gegensatz zu Querschnittsstudien sind die Fehler ϵ_{ij} und ϵ_{ik} (und damit Y_{ij} und Y_{ik}) in longitudinalen Studien typischerweise *nicht unabhängig*.

Daten sind *balanciert*, wenn die gleiche Anzahl Messungen $n_i = n$ gemacht wurde für jede statistische Einheit $i = 1, \dots, m$ an denselben Zeitpunkten $t_{ij} = t_j, j = 1, \dots, n_j$. Longitudinale Daten haben *gleiche Abstände* wenn $t_{j+1} - t_j = d$ für alle $j = 1, \dots, n-1$. Wir nehmen typischerweise an, dass $\Sigma_i = \Sigma_0$ für $i = 1, \dots, m$, das heisst, Σ ist block-diagonal mit $n \times n$ Einträgen Σ_0 .

17.3 Gemischte Modelle

Link zu Varianzanalyse mit Messwiederholungen. Eine Möglichkeit, den Korrelationen von wiederholten Beobachtungen an einer Person (oder an Messungen in einem Cluster) Rechnung zu tragen, sind sogenannte *Mixed Models*. In solchen Modellen gibt es zusätzlich zu ϵ noch andere zufällige Größen. Ein Beispiel für eine Analyse mit wiederholten Messungen haben wir bereits im Kapitel ?? gesehen. Die ANOVA mit Messwiederholung (??) haben wir dort aufgeschrieben mit

$$Y_{im} = \mu + \alpha_i + \pi_m + Res_{im}, \quad i = 1, \dots, I, \quad m = 1, \dots, n. \quad (17.3.1)$$

Der Personeneffekt π_m wird nun als zusätzlicher *zufälliger Effekt* behandelt (neben dem zufälligen Residualterm Res_{im} mit Varianz $\sigma_{Res_{im}}^2$), das heisst, wir nehmen eine Normalverteilung für die π_m an mit Varianz $\sigma_{\pi_m}^2$, $\pi_m \sim \mathcal{N}(0, \sigma_{\pi_m}^2)$. Wir nehmen also an, dass die Personen zufällig aus der Population gezogen wurde, die π_m können dann als Populationsparameter interpretiert werden. Bei *fixiertem* π_m (wenn wir π_m kennen) sind je zwei wiederholte Messungen $Y_{km} \mid \pi_m$ und $Y_{lm} \mid \pi_m$ – also bedingt auf das Personenniveau π_m , voneinander unabhängig. Wir haben also *bedingte Unabhängigkeit*. Marginal aber induziert dieses Modell eine Korrelation zwischen den wiederholten Messungen, diese ist

$$\rho(Y_{km}, Y_{lm}) = \frac{\sigma_{\pi_m}^2}{\sigma_{\pi_m}^2 + \sigma_{Res_{im}}^2}. \quad (17.3.2)$$

Diese Grösse haben wir bereits als Intraklassenkorrelation – also als Korrelation innerhalb einer Klasse (hier Personen) – kennengelernt (??). Das Modell (17.3.1) werden wir unten als *random intercept model* wiedererkennen.

Zuerst schauen wir uns den allgemeinen Fall von einem gemischten Modell an.

Allgemeiner Fall: Generalized linear mixed model GLMM*. Gegeben die Zufallseffekte \mathbf{U}_i sind Y_{i1}, \dots, Y_{in_i} wechselseitig unabhängig und folgen einem Modell mit der Dichte der Exponentialfamilien (z.B. Normal-, Binomial-, Poissonverteilung, usw.)²

$$h(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i, \quad v_{ij} = v(\mu_{ij})\phi, \quad (17.3.4)$$

wobei $h(\cdot)$ die Linkfunktion, $v(\cdot)$ die Varianzfunktion und ϕ den Dispersionsparameter darstellt (siehe Kapitel 16) und \mathbf{d}_{ij} ($q \times 1$) ein Teilmenge von \mathbf{x}_{ij} ($p \times 1$) ist. Die Zufallseffekte $\mathbf{U}_i, i = 1, \dots, m$, sind wechselseitig unabhängig und folgen einer multivariaten Verteilung F mit Parametern $\boldsymbol{\alpha}$.

²mit Parameter θ_{ij} und Dispersionsparameter ϕ ,

$$f(y_{ij} \mid \mathbf{U}_i; \theta, \phi) = \exp[\{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))\}/\phi + c(y_{ij}, \phi)]. \quad (17.3.3)$$

Die bedingten Momente sind $\mu_{ij} = E(Y_{ij} \mid \mathbf{U}_i) = \psi'(\theta_{ij})$ und $v_{ij} = \text{Var}(Y_{ij} \mid \mathbf{U}_i) = \psi''(\theta_{ij})\phi$.

Linearer Fall: Linear mixed model LMM*. Im linearen Fall reduziert sich das Modell (17.3.4) auf ein *lineares gemischtes Modell*

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i + \epsilon_{ij}, \quad \mathbf{U}_i \sim \mathcal{N}_q(\mathbf{0}, G), \quad \epsilon_{ij} \sim \mathcal{N}(0, \tau^2), \quad (17.3.5)$$

wobei G die Kovarianzmatrix der Zufallseffekte \mathbf{U}_i darstellt. Bedingter Erwartungswert und bedingte Varianz sind

$$\mathbb{E}(\mathbf{Y}_i | \mathbf{U}_i) = X_i \boldsymbol{\beta} + D_i \mathbf{U}_i, \quad \text{Cov}(\mathbf{Y}_i | \mathbf{U}_i) = \tau^2 I_{n_i}, \quad (17.3.6)$$

wobei D_i eine $n_i \times q$ Matrix darstellt. $(\mathbf{d}_{i1}, \dots, \mathbf{d}_{in_i})^T$. Marginaler Erwartungswert und Varianz sind

$$\mathbb{E}(\mathbf{Y}_i) = X_i \boldsymbol{\beta}, \quad \text{Cov}(\mathbf{Y}_i) = D_i G D_i^T + \tau^2 I_{n_i}. \quad (17.3.7)$$

Random intercept model. Ein sehr häufiger und einfacher Spezialfall von (17.3.5) ist das *random intercept model*,

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + U_i + \epsilon_{ij}, \quad U_i \sim N(0, \nu^2), \quad \text{Cov}(\mathbf{Y}_i) = \nu^2 J_{n_i} + \tau^2 I_{n_i}, \quad (17.3.8)$$

wobei J_{n_i} eine Matrix mit nur Einsen der Dimension $n_i \times n_i$ darstellt. Es folgt, dass $\text{Var}(Y_{ij}) = \nu^2 + \tau^2$ und $\text{Cov}(Y_{ij}, Y_{ik}) = \nu^2$ für $j \neq k$. Diese Modell ist äquivalent mit einem Modell mit stationärer Korrelation, $\rho = \nu^2 / (\nu^2 + \tau^2)$.

Falls wir nur eine kategoriale Eingangsgröße X haben, entspricht dieses Modell demjenigen einer *Varianzanalyse mit Messwiederholungen*, wie oben bereits erläutert wurde (17.3.1). Wir hatten dieses Modell bereits im Kapitel Varianzanalysen eingeführt (??).

Random slope model. Ein anderer Spezialfall von (17.3.5) ist das *random slope model* (Wachstumsmodell); dabei wird zusätzlich eine zufällige Steigung, z.B. mit $d_{ij} = t_{ij}$ als der Eingangsgröße Zeit, eingebaut,

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + U_{i0} + d_{ij} U_{i1} + \epsilon_{ij}. \quad (17.3.9)$$

Die Korrelationen zwischen wiederholten Messungen sind bei diesem Modell nicht mehr stationär.

Interpretation der Parameter. $\boldsymbol{\beta}$ stellt die erwartete Differenz im Response dar für zwei Individuen mit identischen Eingangsgrößen und identischen Zufallseffekten, die sich auf X um eine Einheit unterscheiden.

Anhang A

Notation

Einige Bemerkungen zur Notation. *Grosse lateinische Buchstaben* X, Y, Z usw. stellen meistens *Variablen* oder *Zufallsgrössen* dar. *Kleine lateinische Buchstaben* x, y, z, s, r usw. bedeuten *empirische Realisationen* oder *Beobachtungen* dieser Zufallsgrössen, einzelne Messwerte, Kennwerte oder allgemein *fixe Grössen, also Zahlen*.

Griechische Buchstaben π, μ, σ, ρ usw. stellen i.A. *unbekannte, wahre Grössen* wie z.B. einen wahren Mittelwert μ oder eine wahre Korrelation ρ zwischen zwei Merkmalen dar. Da sie *unbekannt* sind, werden in der Statistik Schätzungen bezüglich ihnen gemacht oder es werden Hypothesen über sie getestet. Schätzungen von diesen unbekannten Grössen kennzeichnen wir mit einem *Hut* ($\hat{\cdot}$): $\hat{\pi}, \hat{\mu}, \hat{\sigma}, \hat{\rho}$.

X_1, \dots, X_n	Stichprobe der Grösse n
x_1, \dots, x_n	Beobachtungen der Variable X
X	Zufallsvariable X
\bar{x}	Empirischer Mittelwert
\bar{X}	Variable Durchschnitt
s, s^2	Empirische Standardabweichung/Varianz einer Variablen X
$\mu = E(X)$	Wahrer Mittelwert (Erwartungswert von X)
$\sigma^2 = \text{Var}(X)$	Wahre Varianz
$\sigma = \sqrt{\text{Var}(X)}$	Wahre Standardabweichung
α	Irrtums-Wahrscheinlichkeit
T	Allgemein für Teststatistik
Z	Standardnormalverteilte Variable
$\text{se}(T)$	Standardfehler einer Statistik T
$\text{se}(\bar{X})$	Standardfehler des Durschschnitts ($= s/\sqrt{n}$)
p, P, \Pr	Wahrscheinlichkeit
$z, t, F, \chi^2, W, U \dots$	Spezifische Teststatistiken
$X \sim \mathcal{N}(\mu, \sigma^2)$	X ist normalverteilt mit Mittelwert μ und Varianz σ^2
$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$	\bar{X} ist normalverteilt mit Mittelwert μ und Varianz σ^2/n
$Z \sim \mathcal{N}(0, 1)$	Z ist (standard)normalverteilt mit Mittelwert 0 und Varianz 1
ρ, r	Wahre und empirische Korrelation
β, b	Wahrer und empirischer Regressionskoeffizient
$z_{1-\alpha/2}$	($1 - \alpha/2$)-Quantil der z -Verteilung

$t_{1-\alpha/2,n-1}$	(1 - $\alpha/2$)-Quantil der t -Verteilung mit $n - 1$ Freiheitsgraden (df)
$F_{1-\alpha,I-1,n-I}$	(1 - α)-Quantil der F -Verteilung mit $df_1 = I - 1$ und $df_2 = n - I$
$\chi^2_{1-\alpha,p}$	(1 - α)-Quantil einer χ^2 -Verteilung mit p Freiheitsgraden
$\Pr(A B)$	Wahrscheinlichkeit von Ereignis A gegeben Ereignis B
H_0, H_1	Null- und Alternativhypothese
$\Pr(T \geq t_{emp} H_0)$	Wahrscheinlichkeit, dass $T \geq$ empirischer Wert der Statistik (= p -Wert)
$\Pr(T \geq t_{krit} H_0)$	Wahrscheinlichkeit, dass $T \geq$ kritischer Wert der Statistik (= α)
$\epsilon \sim \mathcal{N}(0, \sigma^2)$	Um Null normalverteilter Fehler
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	Modell für einfache lineare Regression von Y auf X ($i = 1, \dots, n$)
$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$	Modell einer einfaktoriellen Varianzanalyse ($i = 1, \dots, k$, $j = 1, \dots, n_i$)

Anhang B

Testübersicht und Tabellen

Stetige Daten. Die Abbildung B.1 gibt einen Überblick über klassische Testprobleme für stetige Daten:

	Daten <i>normalverteilt?</i>	
	ja	nein
1 Stichprobe	Einstichproben <i>t</i> -Test	Vorzeichen-Rangtest
2 Stichproben ungepaart	ungepaarter <i>t</i> -Test	Rangsummentest
2 Stichproben gepaart	gepaarter <i>t</i> -Test	Vorzeichen-Rangtest
> 2 Stichproben ungepaart	ANOVA	Kruskal-Wallis-Test
> 2 Stichproben gepaart	ANOVA für wiederholte Messungen	Friedman-Test

Tabelle B.1: Spezifische Tests für stetige Daten.

Diskrete Daten. Bei binären und nominalen Daten: χ^2 -Verfahren.

Tabellen. Auf den folgenden Seiten sind relevante Verteilungen tabellarisch dargestellt, für die Statistiken z,t,χ^2,F,U und W . Die Tabellen wurden erstellt mit `qnorm()`, `qt()`, usw.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999
3	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabelle B.2: Standardnormalverteilung, Verteilungsfunktion von z .

df	$t_{0.6}$	$t_{0.7}$	$t_{0.8}$	$t_{0.9}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
Inf	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

 Tabelle B.3: t -Tabelle. Quantile der t -Verteilung für verschiedene Freiheitsgrade.

df	$\chi^2_{0.6}$	$\chi^2_{0.7}$	$\chi^2_{0.8}$	$\chi^2_{0.9}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
1	0.71	1.07	1.64	2.71	3.84	5.02	6.63	7.88
2	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.60
3	2.95	3.66	4.64	6.25	7.81	9.35	11.34	12.84
4	4.04	4.88	5.99	7.78	9.49	11.14	13.28	14.86
5	5.13	6.06	7.29	9.24	11.07	12.83	15.09	16.75
6	6.21	7.23	8.56	10.64	12.59	14.45	16.81	18.55
7	7.28	8.38	9.80	12.02	14.07	16.01	18.48	20.28
8	8.35	9.52	11.03	13.36	15.51	17.53	20.09	21.95
9	9.41	10.66	12.24	14.68	16.92	19.02	21.67	23.59
10	10.47	11.78	13.44	15.99	18.31	20.48	23.21	25.19
11	11.53	12.90	14.63	17.28	19.68	21.92	24.72	26.76
12	12.58	14.01	15.81	18.55	21.03	23.34	26.22	28.30
13	13.64	15.12	16.98	19.81	22.36	24.74	27.69	29.82
14	14.69	16.22	18.15	21.06	23.68	26.12	29.14	31.32
15	15.73	17.32	19.31	22.31	25.00	27.49	30.58	32.80
16	16.78	18.42	20.47	23.54	26.30	28.85	32.00	34.27
17	17.82	19.51	21.61	24.77	27.59	30.19	33.41	35.72
18	18.87	20.60	22.76	25.99	28.87	31.53	34.81	37.16
19	19.91	21.69	23.90	27.20	30.14	32.85	36.19	38.58
20	20.95	22.77	25.04	28.41	31.41	34.17	37.57	40.00
21	21.99	23.86	26.17	29.62	32.67	35.48	38.93	41.40
22	23.03	24.94	27.30	30.81	33.92	36.78	40.29	42.80
23	24.07	26.02	28.43	32.01	35.17	38.08	41.64	44.18
24	25.11	27.10	29.55	33.20	36.42	39.36	42.98	45.56
25	26.14	28.17	30.68	34.38	37.65	40.65	44.31	46.93
26	27.18	29.25	31.79	35.56	38.89	41.92	45.64	48.29
27	28.21	30.32	32.91	36.74	40.11	43.19	46.96	49.64
28	29.25	31.39	34.03	37.92	41.34	44.46	48.28	50.99
29	30.28	32.46	35.14	39.09	42.56	45.72	49.59	52.34
30	31.32	33.53	36.25	40.26	43.77	46.98	50.89	53.67
40	41.62	44.16	47.27	51.81	55.76	59.34	63.69	66.77
50	51.89	54.72	58.16	63.17	67.50	71.42	76.15	79.49
60	62.13	65.23	68.97	74.40	79.08	83.30	88.38	91.95
70	72.36	75.69	79.71	85.53	90.53	95.02	100.43	104.21
80	82.57	86.12	90.41	96.58	101.88	106.63	112.33	116.32
90	92.76	96.52	101.05	107.57	113.15	118.14	124.12	128.30
100	102.95	106.91	111.67	118.50	124.34	129.56	135.81	140.17

 Tabelle B.4: χ^2 -Tabelle. Quantile der χ^2 -Verteilung für verschiedene Freiheitsgrade.

n_1, n_2	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0	0	0	0	1	1	1	1	1	2	2	2	2
3	2	2	3	3	4	4	5	5	6	6	7	7	8
4	4	4	5	6	7	8	9	10	11	11	12	13	14
5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	10	12	14	16	18	20	22	24	26	28	30	32	34
8	13	15	17	19	22	24	26	29	31	34	36	38	41
9	15	17	20	23	26	28	31	34	37	39	42	45	48
10	17	20	23	26	29	33	36	39	42	45	48	52	55
11	19	23	26	30	33	37	40	44	47	51	55	58	62
12	22	26	29	33	37	41	45	49	53	57	61	65	69
13	24	28	33	37	41	45	50	54	59	63	67	72	76
14	26	31	36	40	45	50	55	59	64	69	74	78	83
15	29	34	39	44	49	54	59	64	70	75	80	85	90
16	31	37	42	47	53	59	64	70	75	81	86	92	98
17	34	39	45	51	57	63	69	75	81	87	93	99	105
18	36	42	48	55	61	67	74	80	86	93	99	106	112
19	38	45	52	58	65	72	78	85	92	99	106	113	119
20	41	48	55	62	69	76	83	90	98	105	112	119	127

Tabelle B.5: Kritische Werte für den Mann-Whitney- U -Test für $\alpha = 0.05$, zweiseitiger Test.

n	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
7	2	0	
8	3	1	0
9	5	3	1
10	8	5	3
11	10	7	5
12	13	9	7
13	17	12	9
14	21	15	12
15	25	19	15
16	29	23	19
17	34	27	23
18	40	32	27
19	46	37	32
20	52	43	37
21	58	49	42
22	65	55	48
23	73	62	54
24	81	69	61
25	89	76	68

Tabelle B.6: Kritische Werte für den Wilcoxon-Test für $\alpha = 0.05, 0.02, 0.01$, zweiseitiger Test.

All analyses were performed using the R statistical software R version 4.5.1 (2025-06-13) [40].

- R version 4.5.1 (2025-06-13), x86_64-pc-linux-gnu
- Running under: Ubuntu 22.04.5 LTS
- Matrix products: default
- BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.10.0
- LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.10.0
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: BSDA 1.2.2, carData 3.0-5, emmeans 1.10.1, epiR 2.0.19, knitr 1.50, lattice 0.22-6, psych 2.5.3, scatterplot3d 0.3-44, survival 3.8-3, TOSTER 0.8.2
- Loaded via a namespace (and not attached): abind 1.4-5, admisc 0.35, barsurf 0.7.0, base64enc 0.1-3, BiasedUrn 2.0.11, bitops 1.0-9, bivariate 0.7.0, broman 0.80, car 3.1-3, caTools 1.18.3, class 7.3-22, cli 3.6.5, coda 0.19-4.1, codetools 0.2-20, colorspace 2.1-0, compiler 4.5.1, cowplot 1.1.3, dichromat 2.0-0.1, digest 0.6.37, distributional 0.4.0, dplyr 1.1.4, e1071 1.7-14, estimability 1.5, evaluate 1.0.3, exactRankTests 0.8-35, farver 2.1.2, fastmap 1.2.0, formatR 1.14, Formula 1.2-5, generics 0.1.3, ggdist 3.3.2, ggplot2 3.5.2, glue 1.8.0, gplots 3.1.3.1, grid 4.5.1, gtable 0.3.6, gtools 3.9.5, highr 0.11, htmltools 0.5.8.1, htmlwidgets 1.6.4, igraph 2.0.3, jsonlite 2.0.0, KernSmooth 2.23-22, kubik 0.3.0, labeling 0.4.3, lifecycle 1.0.4, lubridate 1.9.3, magrittr 2.0.3, MASS 7.3-60.2, matlib 0.9.6, Matrix 1.7-0, mnormt 2.1.1, multcomp 1.4-25, mvtnorm 1.2-4, nlme 3.1-164, pander 0.6.5, parallel 4.5.1, pillar 1.11.0, pkgconfig 2.0.3, plyr 1.8.9, proxy 0.4-27, purrr 1.0.4, R6 2.6.1, RColorBrewer 1.1-3, Rcpp 1.1.0, reshape2 1.4.4, rgl 1.3.1, rlang 1.1.6, sandwich 3.1-0, scales 1.4.0, splines 4.5.1, stringi 1.8.7, stringr 1.5.1, TH.data 1.1-2, tibble 3.3.0, tidyr 1.3.1, tidyselect 1.2.1, timechange 0.3.0, tools 4.5.1, vctrs 0.6.5, venn 1.12, withr 3.0.2, xfun 0.52, xtable 1.8-4, zoo 1.8-12

Literaturverzeichnis

- [1] Aaron, R., Caldwell, and aut. Exploring equivalence testing with the updated toster r package. *PsyArXiv*, 2022.
- [2] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.
- [3] Alan T. Arnholt and Ben Evans. *BSDA: Basic Statistics and Data Analysis*, 2023. R package version 1.2.2.
- [4] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, 53:370–418, 1763.
- [5] J. Bortz. *Statistik für Human- und Sozialwissenschaftler*. Springer, Berlin; Heidelberg; New York, 2005.
- [6] J. Bortz and N. Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Springer, Heidelberg, 2006.
- [7] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, 1962.
- [8] Ronald Christensen. Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59(2):121–126, 2005.
- [9] Daniel, Lakens, and aut. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 1:1–8, 2017.
- [10] C.S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. Springer Texts in Statistics. Springer, 2002.
- [11] B. De Finetti. *Funzione caratteristica di un fenomeno aleatorio*. Mem. della R. accad. naz. dei Lincei, classe di sci. fis., vi, 4. Società anonima tipografica, 1930.
- [12] H.C.W. de Vet, C.B. Terwee, L.B. Mokkink, and D.L. Knol. *Measurement in Medicine: A Practical Guide*. Practical Guides to Biostatistics and Epidemiology. Cambridge University Press, 2011.
- [13] PJ Diggle, P. Heagerty, KY Liang, and SL Zeger. *Analysis of Longitudinal Data*. Oxford University Press, second edition, 2002.

- [14] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik, Der Weg zur Datenanalyse*. Springer-Lehrbuch. Springer, Heidelberg, 6rd edition, 2007.
- [15] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.
- [16] John Fox, Sanford Weisberg, and Brad Price. *carData: Companion to Applied Regression Data Sets*, 2022. R package version 3.0-5.
- [17] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.
- [18] S. Greenland and K.J. Rothman. *Modern Epidemiology*. Lippincott-Raven, 1998.
- [19] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.
- [20] L. Held. *Methoden der statistischen Inferenz: Likelihood und Bayes*. Springer, Heidelberg, 2008.
- [21] H. Jeffreys. *Theory of Probability*. Oxford, Oxford, England, third edition, 1961.
- [22] Donald E. Knuth. Literate programming. *Comput. J.*, 27(2):97–111, May 1984.
- [23] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 1st edition, 2010.
- [24] John K. Kruschke and Mike Meredith. *BEST: Bayesian Estimation Supersedes the t-Test*, 2015. R package version 0.3.0.
- [25] Daniel Lakens and Aaron Caldwell. *TOSTER: Two One-Sided Tests (TOST) Equivalence Testing*, 2024. R package version 0.8.2.
- [26] Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. R package version 1.10.1.
- [27] Uwe Ligges and Martin Mächler. Scatterplot3d - an r package for visualizing multivariate data. *Journal of Statistical Software*, 8(11):1–20, 2003.
- [28] Uwe Ligges, Martin Maechler, and Sarah Schnackenberg. *scatterplot3d: 3D Scatter Plot*, 2023. R package version 0.3-44.
- [29] Paul E. Meehl. Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, 34(2):103–115, 1967.
- [30] Paul E Meehl. Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology*, 46(4):806, 1978.

- [31] A. Meichtry. *Statistik: Handbuch für Therapeuten*. Thieme, 2017.
- [32] A. Meichtry, J. Kool, A. Schämann, and R. Hilfiker. Therapieeffekte: Beurteilung der empirischen Evidenz. *Physioscience*, 4:184–193, 2008.
- [33] S. Morkved, K. Bo, B. Schei, and K. A. Salvesen. Pelvic floor muscle training during pregnancy to prevent urinary incontinence: a single-blind randomized controlled trial. *Obstet Gynecol*, 101(2):313–319, Feb 2003.
- [34] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231:289–337, 1933.
- [35] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge Univ Press, 2000.
- [36] C. S. Peirce. *Collected Papers*. Harvard University Press, Cambridge, 1931–1935.
- [37] K.R. Popper. *The Logic of Scientific Discovery*. Routledge Classics. Routledge, 2002.
- [38] K.R. Popper and T.E. Hansen. *Die beiden Grundprobleme der Erkenntnistheorie*. Die Einheit der Gesellschaftswissenschaften. Mohr, 1979.
- [39] L.G. Portney and M.P. Watkins. *Foundations of Clinical Research: Applications to Practice*. Pearson/Prentice Hall, 2009.
- [40] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.
- [41] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*, 2025. R package version 2.5.3.
- [42] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008.
- [43] Deepayan Sarkar. *lattice: Trellis Graphics for R*, 2024. R package version 0.22-6.
- [44] S. Senn. *Dicing with Death: Chance, Risk and Health*. Cambridge University Press, 2003.
- [45] David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester, 2004.
- [46] W. Stegmüller. *Jenseits von Popper und Carnap*. Personelle und Statistische Wahrscheinlichkeit. Springer, 1973.
- [47] Mark Stevenson, Evan Sergeant with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jeno Reiczigel, Jim Robison-Cox, Paola Sebastiani, Peter Solymos, Kazuki Yoshida, Geoff Jones, Sarah Pirikahu,

- Simon Firestone, Ryan Kyle, Johann Popp, Mathew Jay, and Charles Reynard. *epiR: Tools for the Analysis of Epidemiological Data*, 2021. R package version 2.0.19.
- [48] D.L. Streiner, G.R. Norman, and J. Cairney. *Health Measurement Scales: A practical guide to their development and use*. OUP Oxford, 2014.
 - [49] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.
 - [50] Terry M Therneau. *survival: Survival Analysis*, 2024. R package version 3.8-3.
 - [51] Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.
 - [52] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. ISBN 978-1498716963.
 - [53] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2025. R package version 1.50.
 - [54] Stephen T. Ziliak and Deirdre N. McCloskey. *The Cult of Statistical Significance*. University of Michigan Press, 1st edition, 2 2008.