University of Applied Sciences

HOCHSCHULE
EMDEN·LEER

Introduction to Data Science

Prof. Dr. Joachim Schwarz

Licence

Disclaimer

- The contents and all materials of this course are protected by copyright and are only intended for use in the course by the students.
- Copies and other reproductions in any form may only be made if and insofar as this has been expressly permitted. The content of the module and all materials may not be changed.
- Public presentations and reproductions in any form are not permitted.
- Downloading the module is at your own risk. The university accepts no liability for damage to the user's computer system, unless the liability is based on an intentional or grossly negligent breach of duty by the university.

Who is…

… Joachim Schwarz?

- Born and grown up in Emden
- Highschool in Emden
- Diploma degree in Mathematics
- Ph. D. in Business Administration
- Data mining team lead with the Deutsche Telekom Group
- Professor for Market Research and Quantitative Methods

Contact:
Room G 104, Emden
Email: joachim.schwarz@hs-emden-leer.de

… and what about you?

What experience have you already had with empirical research and statistics?

What expectations and, if applicable, wishes do you have for the course?

Teaching goals

At the end of this lecture, you should be able to…

- …define and differentiate between key terms and concepts of descriptive and inferential statistics.
- …know the basics of probability theory and the normal distribution.
- …independently select and apply suitable statistical methods depending on the research question and data.
- …apply the common data science methods covered in this lecture and interpret the results.
- …use an advanced statistical program package (R) for analyses.
- …interpret and critically classify statistical results in science and practice.

Curriculum

- Data Science basics including an introduction to R.
- Univariate and bivariate descriptive statistics.
- Probability theory and the normal distribution.
- Inferential statistics basics: Point estimation, confidence intervals and hypothesis testing.
- Data preprocessing and data cleaning.
- Linear regression.
- Methods for classification: Logistic regression.
- Methods for segmentation: Cluster analysis.
- Methods for dimension reduction: Principal component analysis.
- Text mining (optional).

Organisation of this course

- All relevant information and additional documents will be made available in Moodle.
- Please sign up for this course in Moodle!
- The exam consists of a written assignment .

Literature (latest edition)

- Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R.: Multivariate Analysis, Berlin.
- Field, A.; Miles, J.; Field, Z.: Discovering Statistics Using R, London.
- Hosmer, D. W.; Lemeshow, S.: Applied Logistic Regression, New York.
- James, G.; Witten, D.; Hastie, T.; Tibshirani, R.: An Introduction to Statistical Learning with Applications in R, New York, NY.
- Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W.: Applied Linear Statistical Models, Boston.
- Menard, S.: Logistic Regression: From Introductory to Advanced Concepts and Applications, Thousand Oaks.

1. Introduction

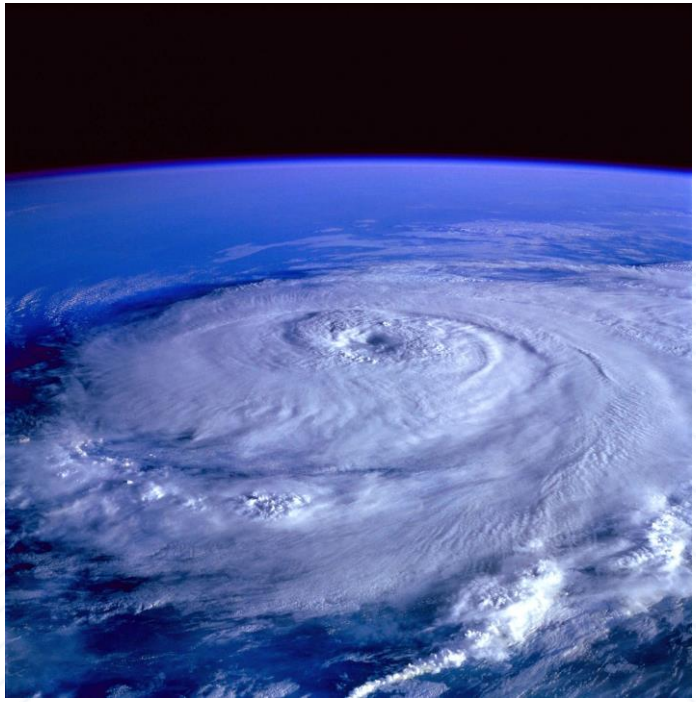
Data Scientist: The Sexiest Job of the 21st Century

Davenport / Patil, HBR, October 2012

1.1 Basic concepts and examples

Example – Hurricane Frances

- In 2004, hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast.
- Executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology.
- Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier.
- Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004).



Provost / Fawcett (2013), Data Science for Business, p. 3; Pixabay, accessed 25.11.2024

Example – Hurricane Frances

- Why is data-driven prediction useful in this scenario?
- One could predict that people in the path of the hurricane would buy more bottled water – but this is obvious, there is no data-driven prediction necessary for that conclusion.
- It is more valuable to discover patterns due to the hurricane that were not obvious.
- To do this, analysts examined the huge volume of Wal-Mart data from prior, similar situations (such as Hurricane Charley) to identify unusual local demand for products.
- And this it what the experts found (among others): Strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane, and the pre-hurricane top-selling item was beer

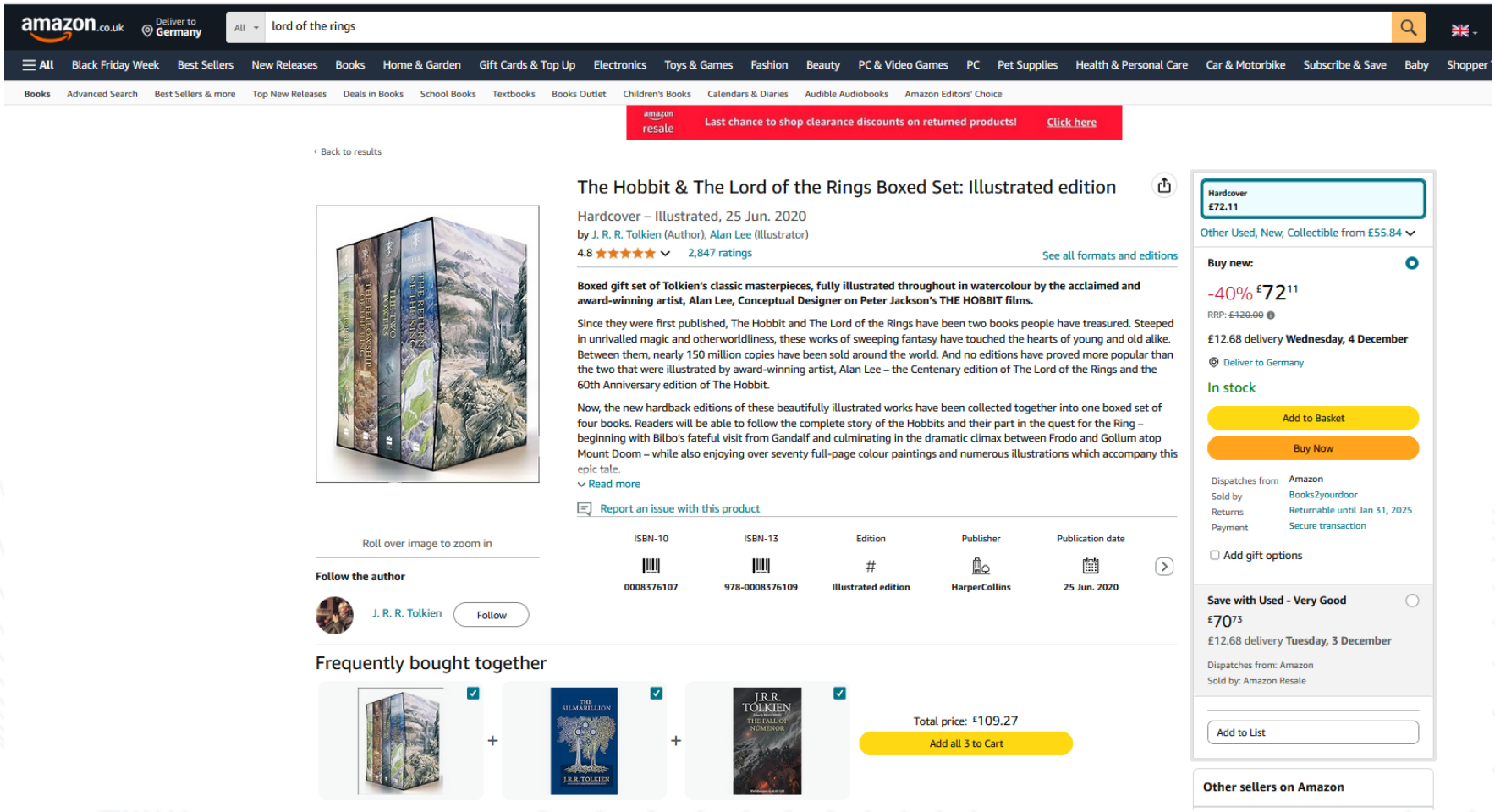
Provost / Fawcett (2013), Data Science for Business, p. 3; Pixabay, accessed 25.11.2024

Example – Diapers and Beer

And there are other examples for data-driven buying behaviour analyses…

https://goodscarlett.blogspot.com/, accessed 25.11.2024
https://www.bissantz.de/bella-consults/deceiving-with-diapers/?lang=en, accessed 25.11.2024

Example – Amazon



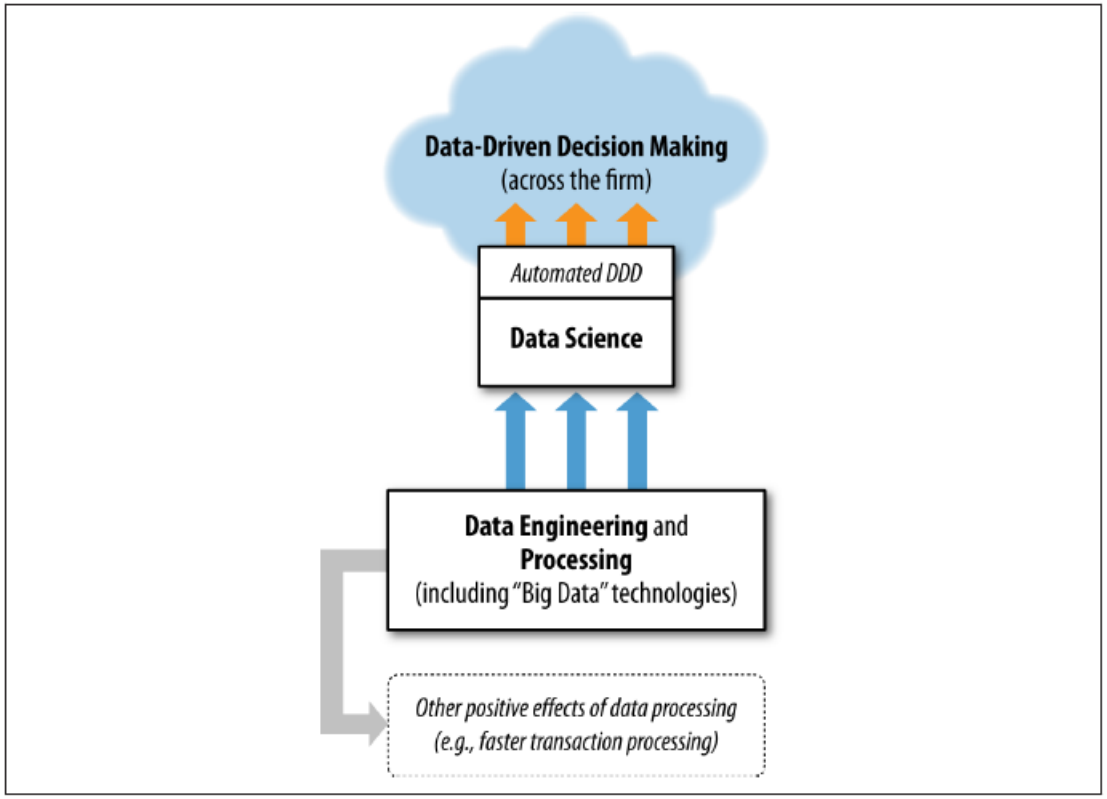
Example – Churn prediction

- Consider e.g. a telco company. It is not uncommon, that 20% of cell phone customers leave when their contracts expire.
- It is getting increasingly difficult to acquire new customers, since the cell phone market is now saturated.
- Customers' switching from one company to another is called churn, and it is expensive: One company must spend on incentives to attract a customer while another company loses revenue when the customer departs.
- Retaining existing customers is cheaper, but who are the customers to be offered the special retention deal prior to the expiration of their contracts?
- The analysis task is to look into the data of churns in the past and to build a model to predict churns in the future.
- In reality, customer retention has been a major use of data mining technologies – especially in telecommunications and finance businesses.

Provost / Fawcett (2013), Data Science for Business, p. 4.

Data Science and Data Driven Decisions

What is Data Science? – And what is it not?



Data Science:

Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.

Data Driven Decision (DDD):

Data-driven decision-making (DDD) refers to the practice of basing decisions on the analysis of data.

Provost / Fawcett (2013), Data Science for Business, pp. 4.

Data Science and Data Driven Decisions

Data Science and Data Driven Decisions are usually applied to two kinds of problems:

1. Decisions for which “discoveries” need to be made within data.

2. Decisions that repeat, especially at massive scale, and so decision-making can benefit from even small increases in decision-making accuracy based on data analysis.

→ The capability to extract useful knowledge from data is one of the key strategic assets of a company.

→ For this, a company needs both: The right data and suitable data science talents.

Provost / Fawcett (2013), Data Science for Business, pp. 6.

Unsupervised and supervised data mining problems

Consider two similar questions we might ask about a customer population.

The first is:

> "Do our customers naturally fall into different groups?"

Here no specific purpose or target has been specified for the grouping. When there is no such target, the data mining problem is referred to as **unsupervised**.

Contrast this with a slightly different question:

> "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire?"

Here there is a specific target defined: Will a customer leave when her contract expires? In this case, segmentation is being done for a specific reason: to take action based on likelihood of churn. This is called a **supervised** data mining problem. Supervised problems require the target information in the data, e.g. information on earlier service cancellations.

Provost / Fawcett (2013), Data Science for Business, p. 24.

Unsupervised and supervised data mining problems

Supervised data mining problem:

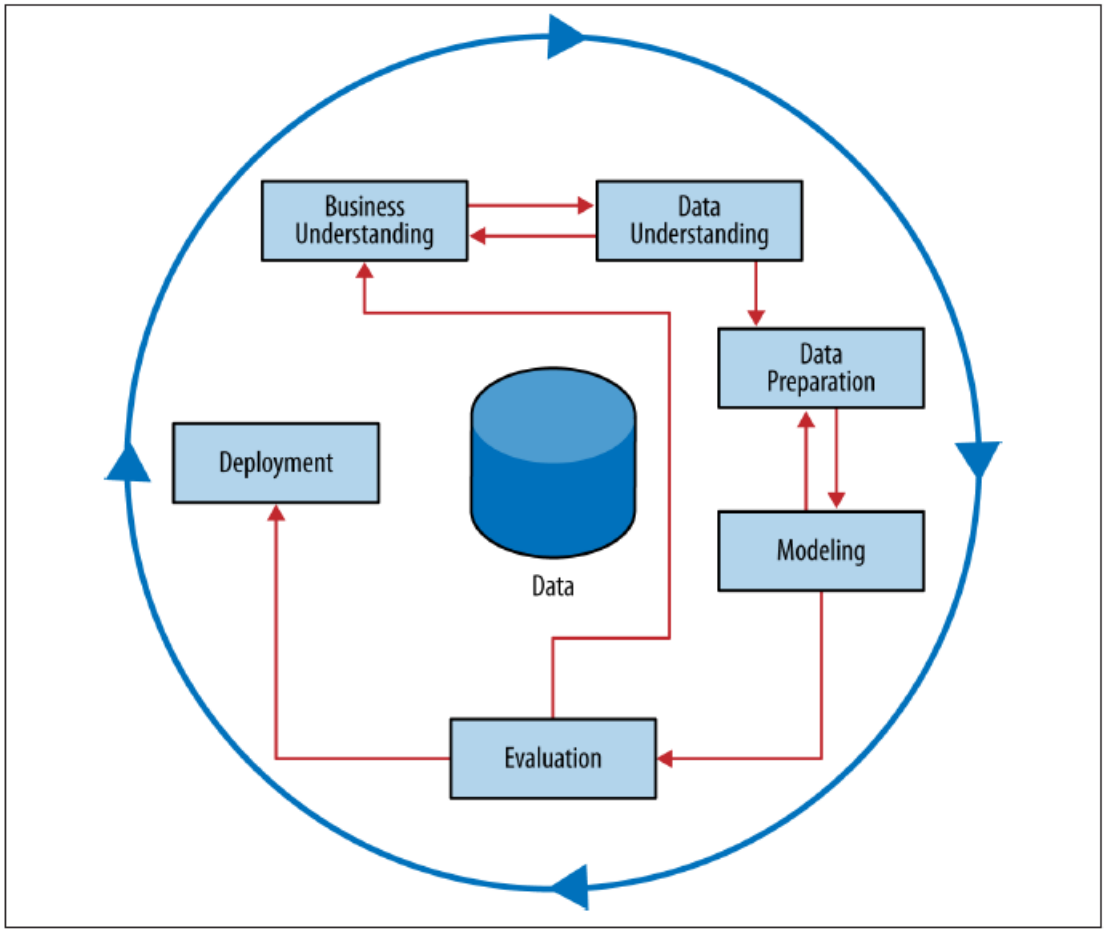
- For each observation of the predictor measurement(s) x_i, i = 1, . . . ,n there is an associated response measurement y_i.
- The task is to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).
- Typical methods: Linear regression, logistic regression.

Unsupervised data mining problem:

- For every observation i = 1, . . . ,n, we observe a vector of measurements x_i but no associated response y_i.
- The task is to understand the relationships between the variables or between the observations.
- Typical methods: Clustering.

James et al. (2023), An Introduction into Statistical Learning with Applications in R, pp. 26.

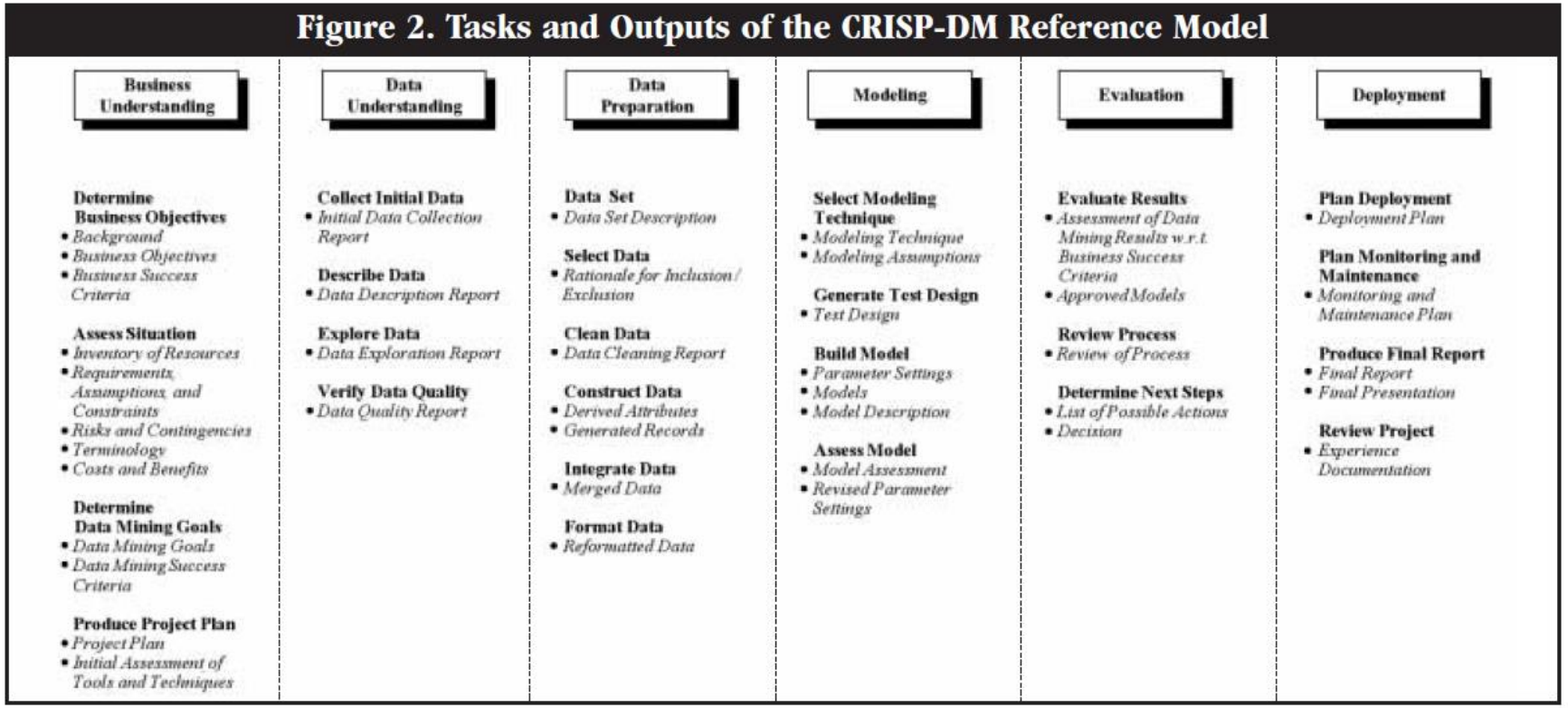
CRISP-DM – CRoss-Industry Standard Process for Data Mining



- Many data mining problems in industry follow a similar structure.
- A standard process supports project planning and problem structuring.
- A standard process allows projects to be replicated.

Shearer, C. (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, pp. 13-22.
Provost / Fawcett (2013), Data Science for Business, p. 27.

CRISP-DM – CRoss-Industry Standard Process for Data Mining



Provost / Fawcett (2013), Data Science for Business, pp. 26.
Shearer, C. (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5, p. 14.

The part of statistics within data science

- The field of Statistics provides us with a huge amount of knowledge that underlies analytics, and can be thought of as a component of the larger field of Data Science.
- Statistics helps us to understand different data distributions and what statistics are appropriate to summarize each.
- Statistics helps us understand how to use data to test hypotheses and to estimate the uncertainty of conclusions.
- In relation to data mining, hypothesis testing can help determine whether an observed pattern is likely to be a valid, general regularity as opposed to a chance occurrence in some particular dataset.
- Many of the techniques for extracting models or patterns from data have their roots in Statistics.

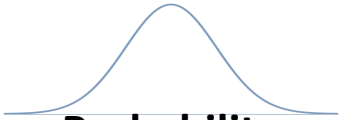
Provost / Fawcett (2013), Data Science for Business, p. 36.

Descriptive and inferential statistics

The two main objectives of statistics

Descriptive Statistics

Goal:
To summarize and describe data without drawing conclusions.



Probability Theory

Inferential Statistics

Goal:
To generalize and to draw conclusions from a sample to a larger underlying population.

1.2 Data Science as a profession

Exercise

Research some current job openings for data scientists (enter the search term “jobs for data scientists” in any search engine) and identify typical requirements and job profiles of data scientists.

What are typical fields of work for data scientists?

What are typical requirements for data scientists?

1.3 Introduction to 

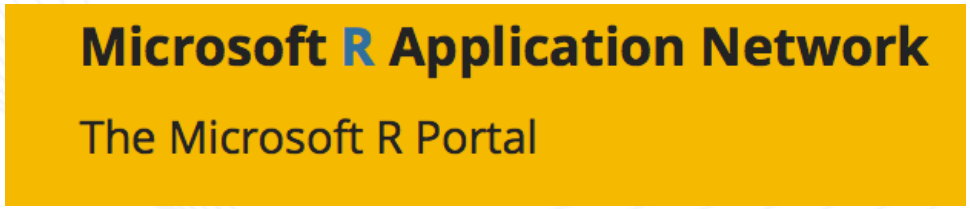
What is R?

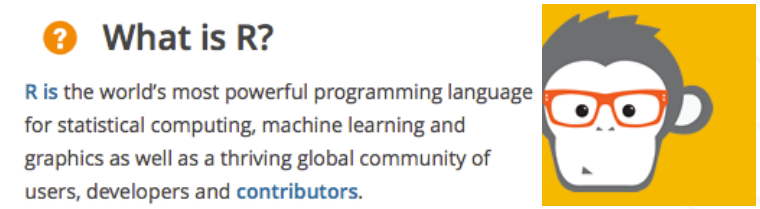
- R is a programming language for statistical graphics and statistical computing.
- R was developed in the 1990s in the Department of Statistics at the University of Auckland, New Zealand, and is based on the statistical programming language S - hence the name.
- Since 1997, the R Development Core Team has been maintaining the R software and integrating contributions from the constantly growing R community, which develops its own R source codes.
- Official website of the R project: https://www.r-project.org/.
- The commercial software R Studio is used as the user interface for R, which is free for academic purposes: https://rstudio.com/products/rstudio/.

R in real life

"I believe that the ability to program will become one of the basic skills of young people, alongside reading, writing and arithmetic. These will not disappear. But programming will be added". Angela Merkel

Companies who seriously analyse data use R quite often:





Advantages of R

- Very large variety of methods and applications (finance, marketing, HR, psychology, . . .).
- New methods of data analysis are often developed in R (including big data, AI, etc.).
- Free and open; free of charge.
- Interfaces to many data sources (including social media, etc.).
- Extensions for Microsoft, Oracle, SAP products, but also SPSS, SAS, among others.
- Countless users worldwide in companies and science.
- Options for reporting, apps, etc.
- Numerical stability / accuracy.
- Large developer community with a long history since 1993; R consortium, including IBM, Microsoft, TIPCO, Google, . . .

R is state of the art in statistical data analysis

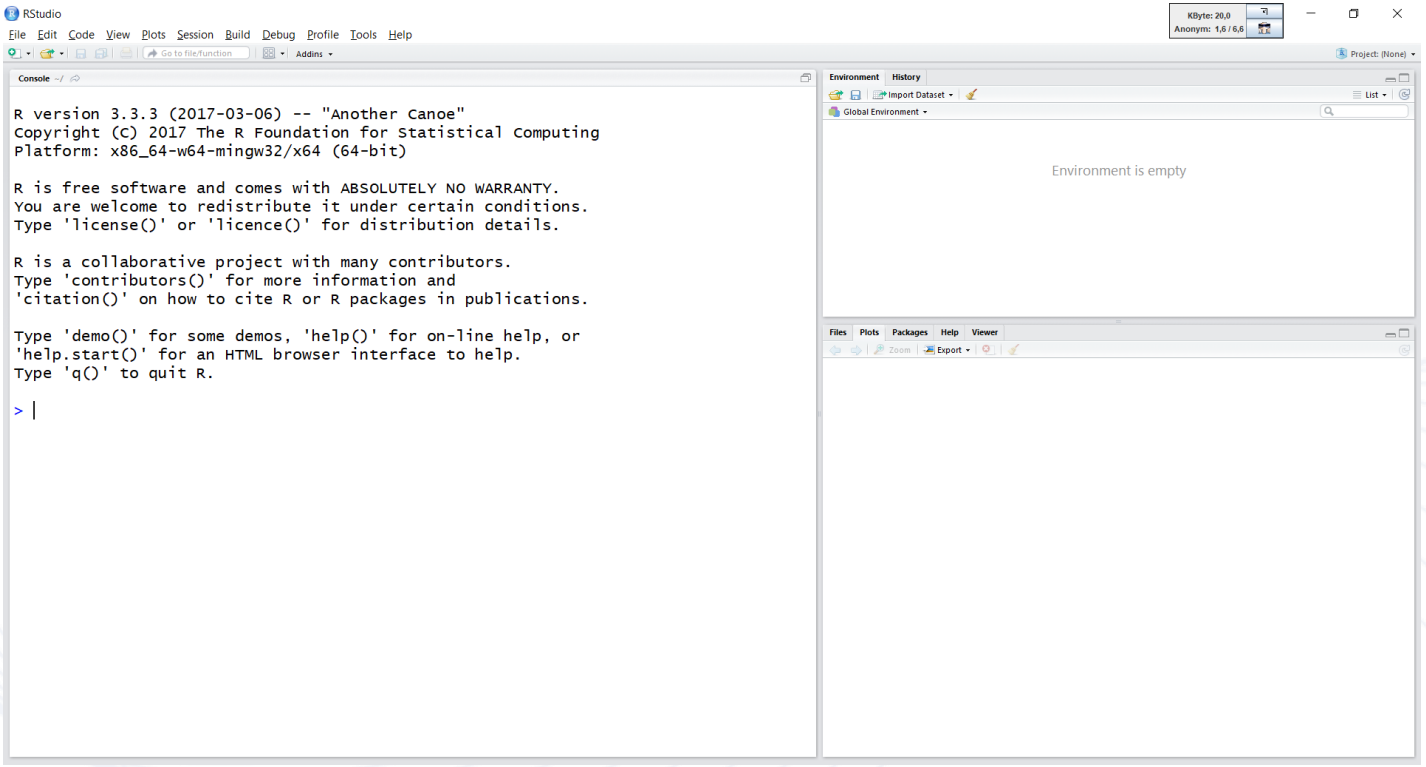
Working with R

This symbol indicates that we are going to start to work with R now. Please start R on your computer.

The first launch

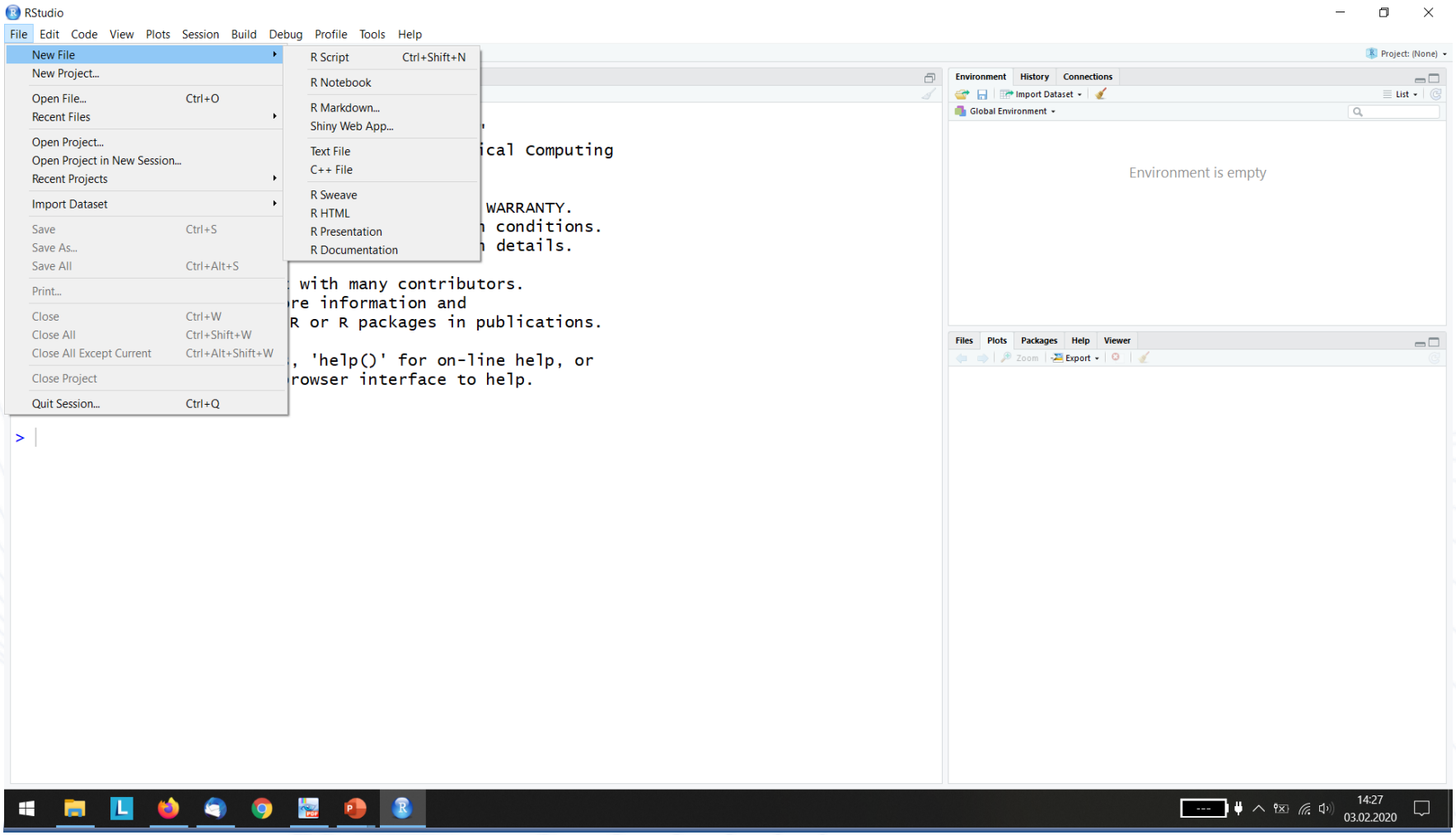
Start R by clicking on the icon with the R-Studio logo (or on the executable file in the `bin` folder of the directory where R Studio is installed).

You should see the following image on your screen:

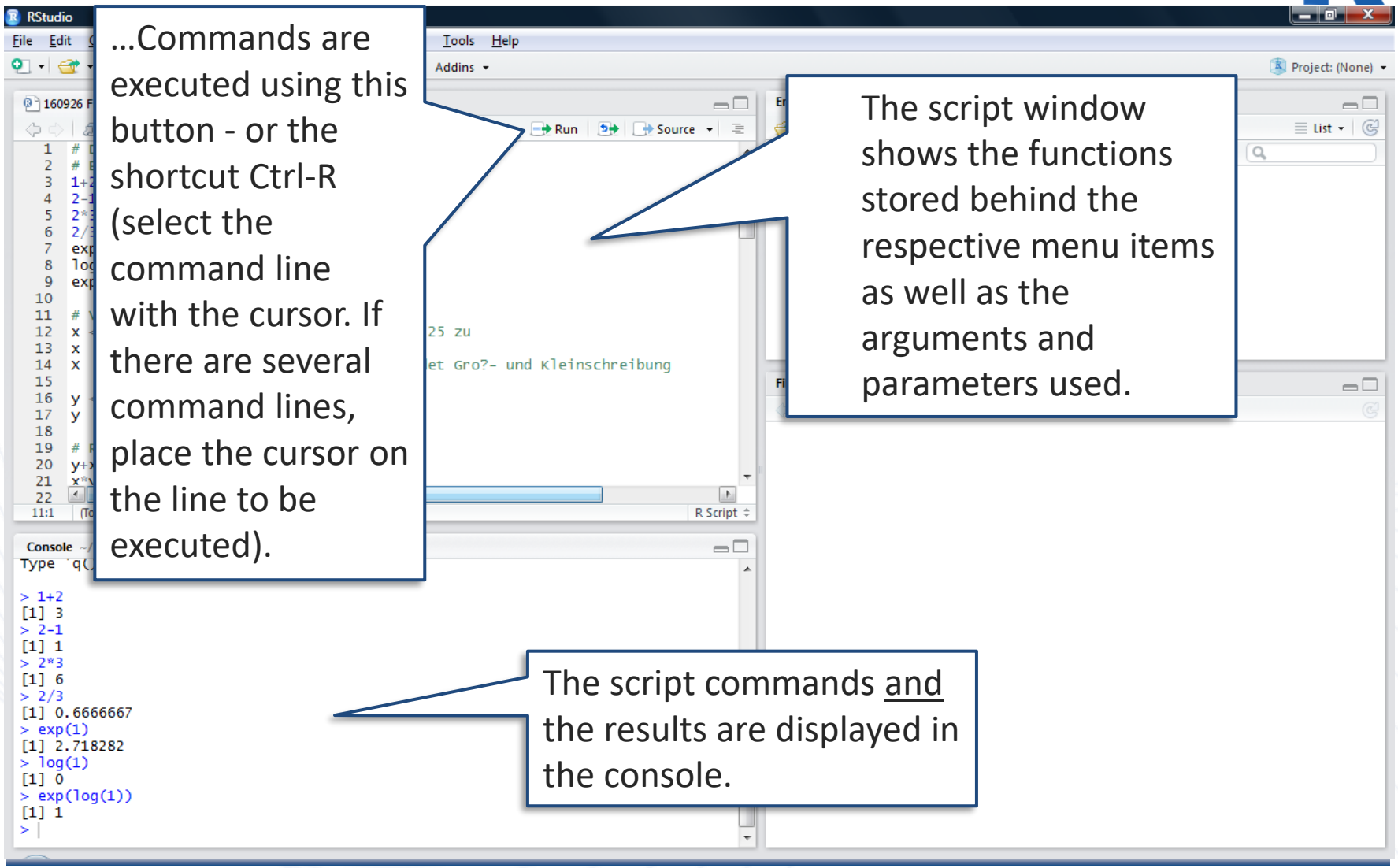


New script window

Please open a new R script window: File → New File → R Script :



R Studio: User interface for R

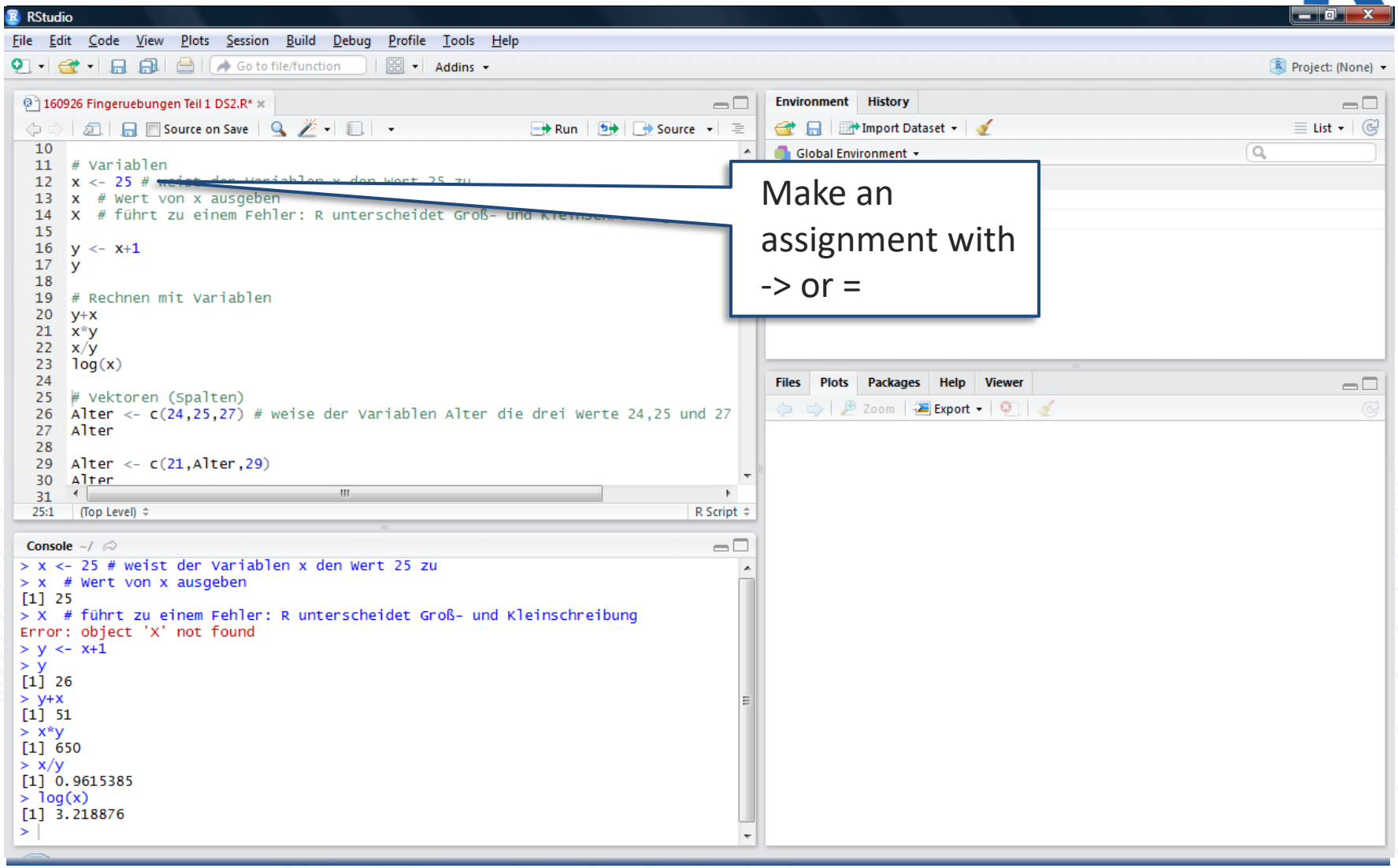


…Commands are executed using this button - or the shortcut Ctrl-R (select the command line with the cursor. If there are several command lines, place the cursor on the line to be executed).

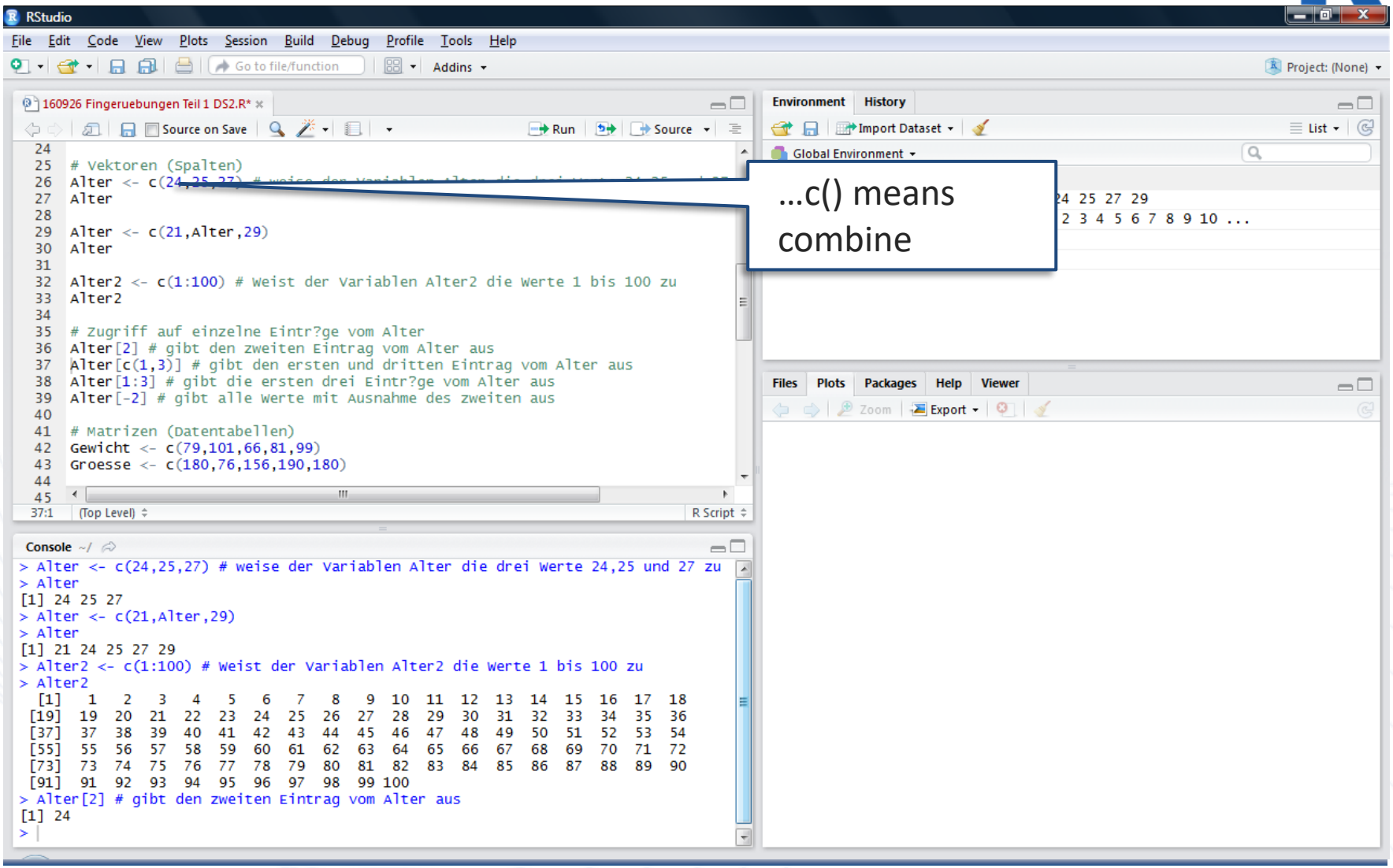
The script window shows the functions stored behind the respective menu items as well as the arguments and parameters used.

The script commands and the results are displayed in the console.

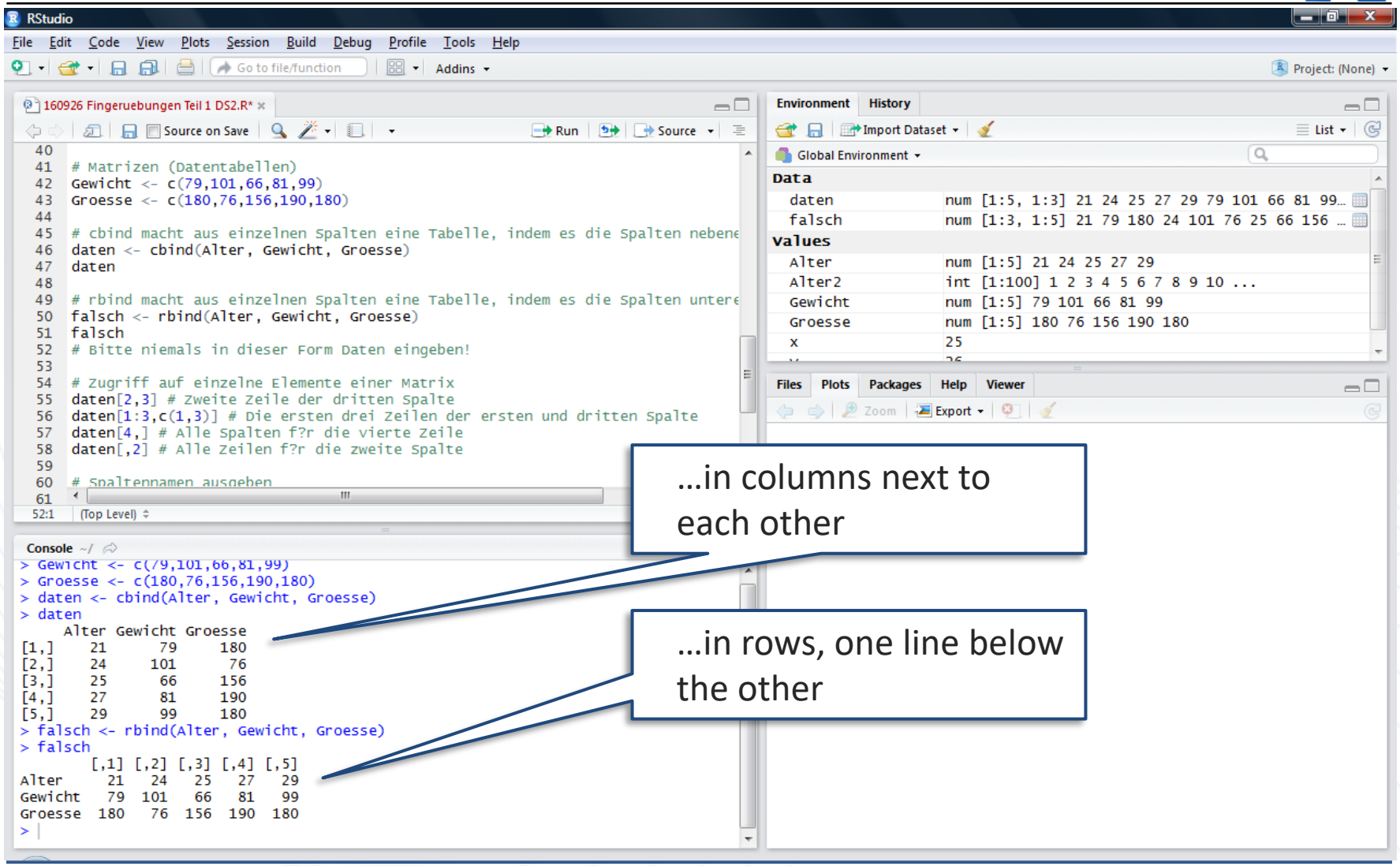
Variables


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
160926 Fingeruebungen Teil 1 DS2.R*
Source on Save
Run
Source
10
11 # Variablen
12 x <- 25 # weist der Variablen x den Wert 25 zu
13 x # Wert von x ausgeben
14 X # führt zu einem Fehler: R unterscheidet Groß- und Kleinschr
15
16 y <- x+1
17 y
18
19 # Rechnen mit Variablen
20 y+x
21 x*y
22 x/y
23 log(x)
24
25 # Vektoren (Spalten)
26 Alter <- c(24,25,27) # weise der Variablen Alter die drei Werte 24,25 und 27
27 Alter
28
29 Alter <- c(21,Alter,29)
30 Alter
31
25:1 (Top Level)
R Script
Console ~/
> x <- 25 # weist der Variablen x den Wert 25 zu
> x # Wert von x ausgeben
[1] 25
> X # führt zu einem Fehler: R unterscheidet Groß- und Kleinschreibung
Error: object 'X' not found
> y <- x+1
> y
[1] 26
> y+x
[1] 51
> x*y
[1] 650
> x/y
[1] 0.9615385
> log(x)
[1] 3.218876
>
Environment History
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
Zoom
Export
Make an assignment with -> or =

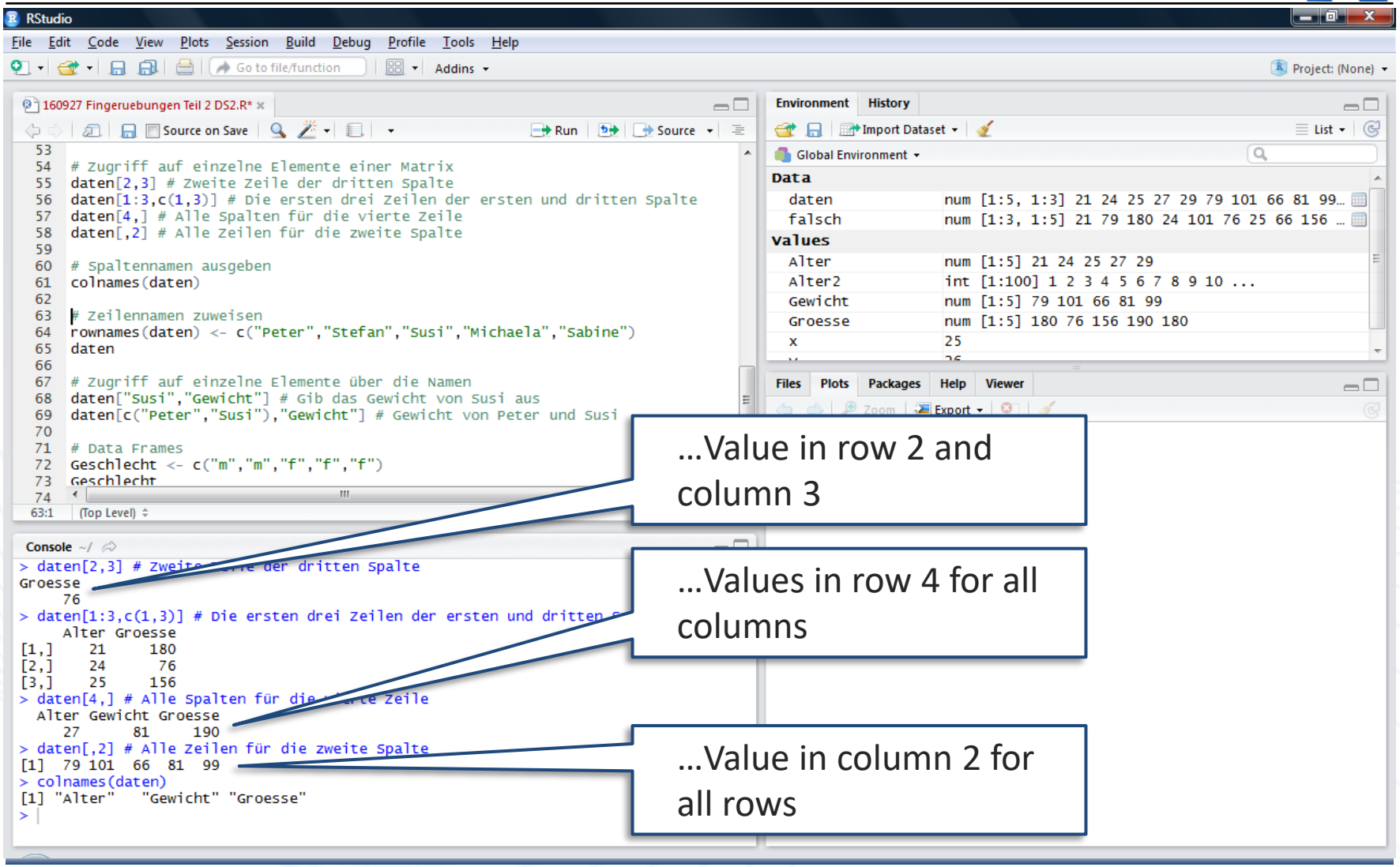
Vektors


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
160926 Fingeruebungen Teil 1 DS2.R*
Source on Save
Run
Source
24
25 # Vektoren (Spalten)
26 Alter <- c(24,25,27)
27 Alter
28
29 Alter <- c(21,Alter,29)
30 Alter
31
32 Alter2 <- c(1:100) # Weist der Variablen Alter2 die Werte 1 bis 100 zu
33 Alter2
34
35 # Zugriff auf einzelne Eintr?ge vom Alter
36 Alter[2] # gibt den zweiten Eintrag vom Alter aus
37 Alter[c(1,3)] # gibt den ersten und dritten Eintrag vom Alter aus
38 Alter[1:3] # gibt die ersten drei Eintr?ge vom Alter aus
39 Alter[-2] # gibt alle Werte mit Ausnahme des zweiten aus
40
41 # Matrizen (Datentabellen)
42 Gewicht <- c(79,101,66,81,99)
43 Groesse <- c(180,76,156,190,180)
44
45
37:1 (Top Level) R Script
Console ~/
> Alter <- c(24,25,27) # weise der Variablen Alter die drei Werte 24,25 und 27 zu
> Alter
[1] 24 25 27
> Alter <- c(21,Alter,29)
> Alter
[1] 21 24 25 27 29
> Alter2 <- c(1:100) # Weist der Variablen Alter2 die Werte 1 bis 100 zu
> Alter2
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
[19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
[37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
[55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
[73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
[91] 91 92 93 94 95 96 97 98 99 100
> Alter[2] # gibt den zweiten Eintrag vom Alter aus
[1] 24
>
Environment History
Import Dataset
List
Global Environment
24 25 27 29
2 3 4 5 6 7 8 9 10 ...
Files Plots Packages Help Viewer
Zoom
Export
...c() means combine

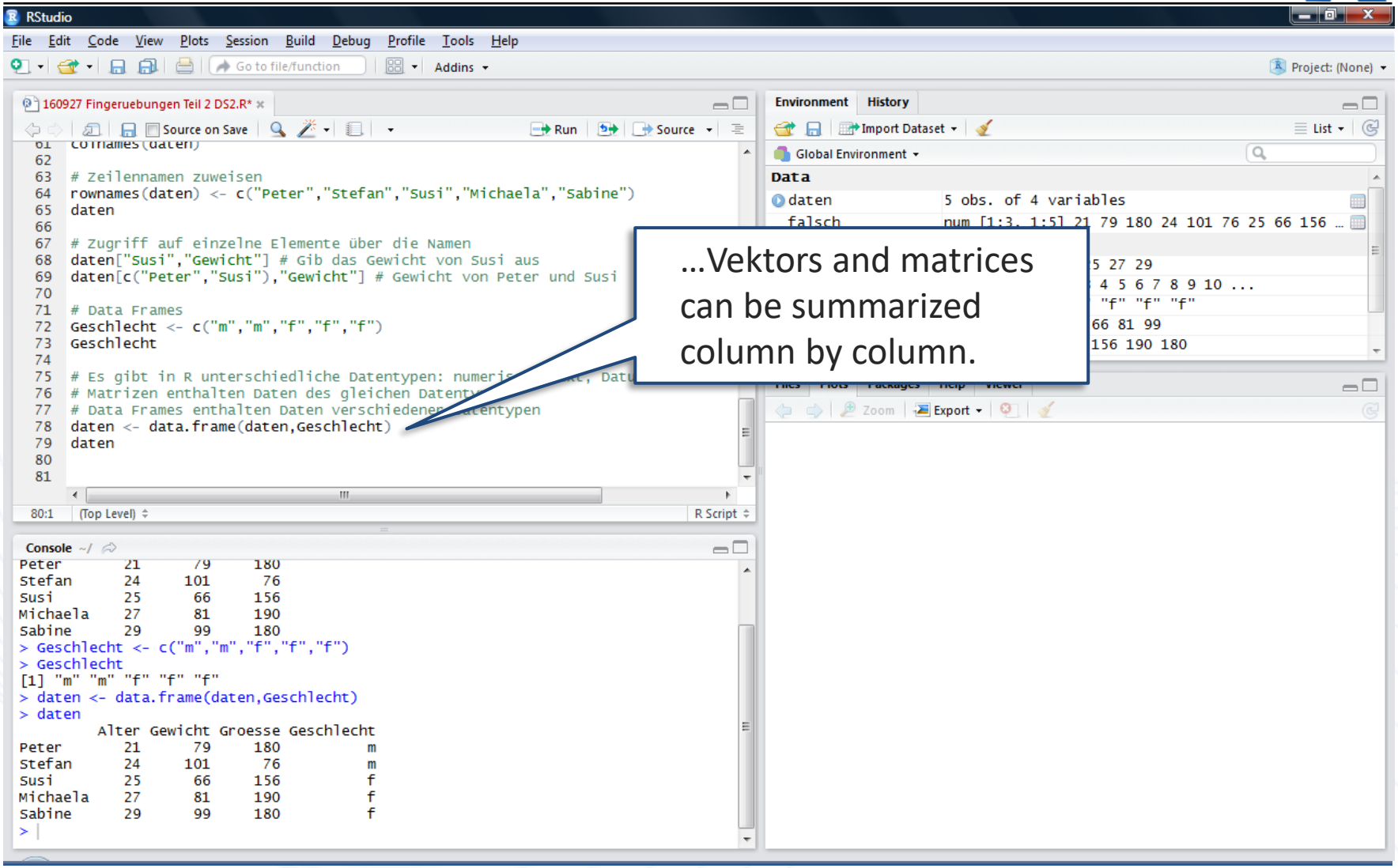
Matrices


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
160926 Fingeruebungen Teil 1 DS2.R*
Source on Save
Run
Source
40
41 # Matrizen (Datentabellen)
42 Gewicht <- c(79,101,66,81,99)
43 Groesse <- c(180,76,156,190,180)
44
45 # cbind macht aus einzelnen Spalten eine Tabelle, indem es die Spalten nebene
46 daten <- cbind(Alter, Gewicht, Groesse)
47 daten
48
49 # rbind macht aus einzelnen Spalten eine Tabelle, indem es die Spalten untere
50 falsch <- rbind(Alter, Gewicht, Groesse)
51 falsch
52 # Bitte niemals in dieser Form Daten eingeben!
53
54 # Zugriff auf einzelne Elemente einer Matrix
55 daten[2,3] # Zweite Zeile der dritten Spalte
56 daten[1:3,c(1,3)] # Die ersten drei Zeilen der ersten und dritten Spalte
57 daten[4,] # Alle Spalten f?r die vierte Zeile
58 daten[,2] # Alle Zeilen f?r die zweite Spalte
59
60 # Spaltennamen ausgeben
61
52:1 (Top Level)
Console ~/
> Gewicht <- c(79,101,66,81,99)
> Groesse <- c(180,76,156,190,180)
> daten <- cbind(Alter, Gewicht, Groesse)
> daten
Alter Gewicht Groesse
[1,] 21 79 180
[2,] 24 101 76
[3,] 25 66 156
[4,] 27 81 190
[5,] 29 99 180
> falsch <- rbind(Alter, Gewicht, Groesse)
> falsch
[,1] [,2] [,3] [,4] [,5]
Alter 21 24 25 27 29
Gewicht 79 101 66 81 99
Groesse 180 76 156 190 180
Environment History
Import Dataset
List
Global Environment
Data
daten num [1:5, 1:3] 21 24 25 27 29 79 101 66 81 99...
falsch num [1:3, 1:5] 21 79 180 24 101 76 25 66 156 ...
Values
Alter num [1:5] 21 24 25 27 29
Alter2 int [1:100] 1 2 3 4 5 6 7 8 9 10 ...
Gewicht num [1:5] 79 101 66 81 99
Groesse num [1:5] 180 76 156 190 180
x 25
Files Plots Packages Help Viewer
Zoom
Export
...in columns next to each other
...in rows, one line below the other

Matrices


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
160927 Fingeruebungen Teil 2 DS2.R*
Source on Save
Run
Source
53
54 # Zugriff auf einzelne Elemente einer Matrix
55 daten[2,3] # Zweite Zeile der dritten Spalte
56 daten[1:3,c(1,3)] # Die ersten drei Zeilen der ersten und dritten Spalte
57 daten[4,] # Alle Spalten für die vierte Zeile
58 daten[,2] # Alle Zeilen für die zweite Spalte
59
60 # Spaltennamen ausgeben
61 colnames(daten)
62
63 # Zeilennamen zuweisen
64 rownames(daten) <- c("Peter","Stefan","Susi","Michaela","Sabine")
65 daten
66
67 # Zugriff auf einzelne Elemente über die Namen
68 daten["Susi","Gewicht"] # Gib das Gewicht von Susi aus
69 daten[c("Peter","Susi"),"Gewicht"] # Gewicht von Peter und Susi
70
71 # Data Frames
72 Geschlecht <- c("m","m","f","f","f")
73 Geschlecht
74
63:1 (Top Level)
Console ~/
> daten[2,3] # Zweite Zeile der dritten Spalte
Groesse
76
> daten[1:3,c(1,3)] # Die ersten drei Zeilen der ersten und dritten
Alter Groesse
[1,] 21 180
[2,] 24 76
[3,] 25 156
> daten[4,] # Alle Spalten für die vierte Zeile
Alter Gewicht Groesse
27 81 190
> daten[,2] # Alle Zeilen für die zweite Spalte
[1] 79 101 66 81 99
> colnames(daten)
[1] "Alter" "Gewicht" "Groesse"
>
Environment History
Import Dataset
List
Global Environment
Data
daten num [1:5, 1:3] 21 24 25 27 29 79 101 66 81 99...
falsch num [1:3, 1:5] 21 79 180 24 101 76 25 66 156 ...
Values
Alter num [1:5] 21 24 25 27 29
Alter2 int [1:100] 1 2 3 4 5 6 7 8 9 10 ...
Gewicht num [1:5] 79 101 66 81 99
Groesse num [1:5] 180 76 156 190 180
x 25
Files Plots Packages Help Viewer
Zoom
Export
...Value in row 2 and column 3
...Values in row 4 for all columns
...Value in column 2 for all rows

Data frames


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
160927 Fingeruebungen Teil 2 DS2.R*
Source on Save
Run
Source
63 # Zeilennamen zuweisen
64 rownames(daten) <- c("Peter","Stefan","Susi","Michaela","Sabine")
65 daten
67 # Zugriff auf einzelne Elemente über die Namen
68 daten["Susi","Gewicht"] # Gib das Gewicht von Susi aus
69 daten[c("Peter","Susi"),"Gewicht"] # Gewicht von Peter und Susi
71 # Data Frames
72 Geschlecht <- c("m","m","f","f","f")
73 Geschlecht
75 # Es gibt in R unterschiedliche Datentypen: numeris
76 # Matrizen enthalten Daten des gleichen Datenty
77 # Data Frames enthalten Daten verschiedener
78 daten <- data.frame(daten,Geschlecht)
79 daten
80:1 (Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
Data
daten 5 obs. of 4 variables
falsch num [1:3, 1:5] 21 79 180 24 101 76 25 66 156 ...
5 27 29
4 5 6 7 8 9 10 ...
"f" "f" "f"
66 81 99
156 190 180
Zoom
Export
Console
Peter 21 79 180
Stefan 24 101 76
Susi 25 66 156
Michaela 27 81 190
Sabine 29 99 180
> Geschlecht <- c("m","m","f","f","f")
> Geschlecht
[1] "m" "m" "f" "f" "f"
> daten <- data.frame(daten,Geschlecht)
> daten
Alter Gewicht Groesse Geschlecht
Peter 21 79 180 m
Stefan 24 101 76 m
Susi 25 66 156 f
Michaela 27 81 190 f
Sabine 29 99 180 f
…Vektors and matrices can be summarized column by column.

Data import in R, install additional packages, activate data sets

Data import in R

1. Save or export data in Excel as csv-file
2. Read a data set: data <- read.csv2(file.choose())

This assigns the imported data set to an object named data.

Install and activate additional packages

Additional packages need be installed only once and then activated in each session as required. A stable Internet connection is required for installation.

- Only once: install.packages("reshape2")
- In every R session: library(reshape2)

Activate data sets contained in R

Tips data set: data(tips, package="reshape2")

The main sample data set for this lecture

- The tips data set contains data records from 244 restaurant visits.
- This data set serves as sample data set for our exercises.
- You find this data set in the reshape2-package: data(tips, package="reshape2")

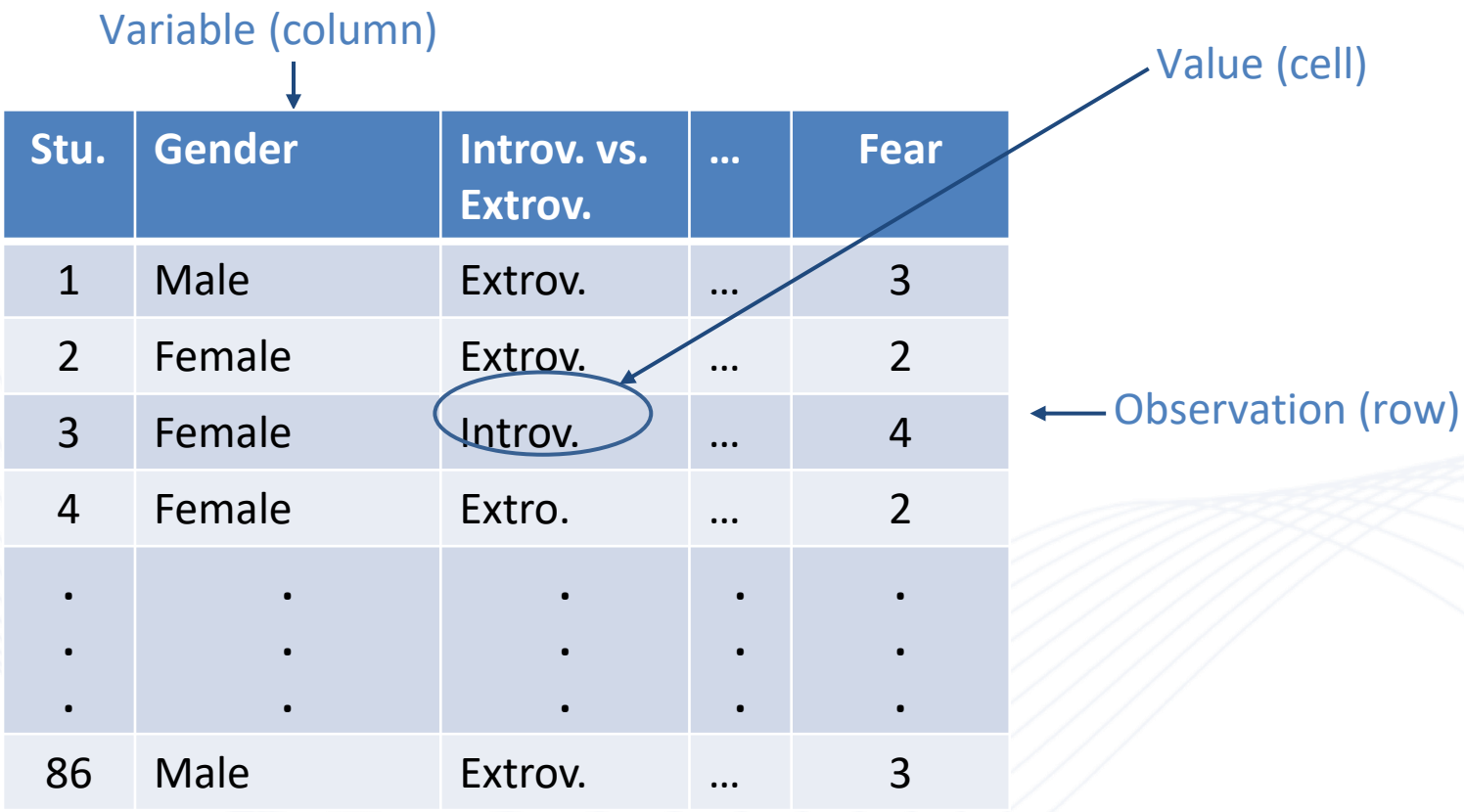
Variable	Beschreibung
total_bill	Restaurant bill ($)
tip	Tip ($)
sex	Gender of the tipper
day	Day of the week
time	Lunch or dinner
size	Number of people on the bill
smoker	Is a smoker present (yes/no)

2. Univariate and multivariate statistics

2.1 Basic terms and basic concepts

Typical structure of a data matrix for statistical analyses

Example: Data collection from students in a statistics lecture on a large number of variables.



Stu.	Gender	Introv. vs. Extrov.	...	Fear
1	Male	Extrov.	...	3
2	Female	Extrov.	...	2
3	Female	Introv.	...	4
4	Female	Extro.	...	2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
86	Male	Extrov.	...	3

Classification of variables: Continuous and discrete

Discrete

- The number of different values is finite or at least countable. Nominal-scaled variables are discrete, many ordinal variables as well as metric variables where something is counted (e.g. number of children, customers, cars produced).

Continuous

- At least theoretically, there are an infinite number of different intermediate values (real numbers, e.g. revenues, size, temperature, age) between two values of a variable.

Classification of variables: Measurement scales

high

Information and interpretation quality

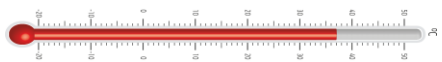
low

metric (numeric)

Ratio scale

- Definition of the distances between two values
- Absolute zero point
- **Examples:** Size, Age

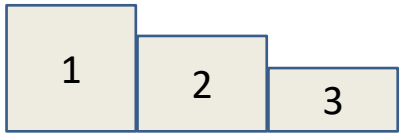
Interval scale



- Same as ratio scale, but:
- Arbitrary zero point
- **Examples:** Intelligence ratio, temperature

categorical

Ordinal scale



- There exists a natural order of the values
- No definition of the distances between the ranks
- **Examples:** Podium places, school marks

Nominal scale

- Purely nominal differences between the values
- No natural order of the values
- **Examples:** Blood group, gender

https://pixabay.com/de/vectors/thermometer-temperatur-grad-celsius-1134182/, accessed 03.01.2025

Measurement scales and feasible calculations

Measurement scale	order – distance – calculations – counting			
Nominal scale e.g. color	no	no	no	yes
Ordinal scale e.g. school marks	yes	no	no	yes
Numerical scales e.g. income	yes	yes	yes	yes

An example for the classification of variables

	Gender	Sleep	Bedtime	Countries	Fear
1	Male	5	Before 10 p.m.	13	3
2	Female	7	10-12 p.m.	7	2
3	Female	5,5	0-2 a.m.	1	4
4	Female	7	0-2 a.m.	-	2
5	Male	3	Before 10 p.m.	1	3
6	Female	3	10-12 p.m.	9	4

Gender – nominal / discrete

Sleeping time – metric (numerical) / continuous

Bedtime – ordinal / continuous

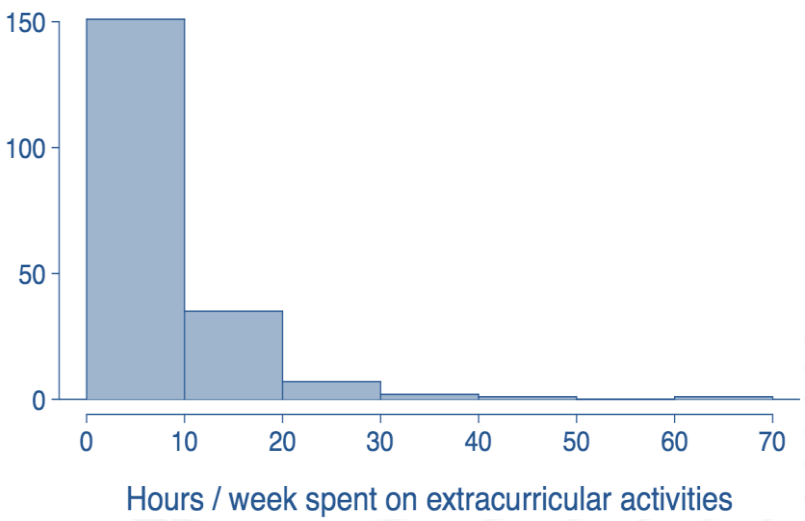
Visited countries in the world – metric / discrete

Fear – numerical / discrete

2.2 Univariate descriptive statistic for numerical data

Histograms

- A **histogram** displays grouped data graphically and provides a view of the **shape of the data distribution**.
- The size of the rectangle is proportional to the frequency. The larger the rectangles, the higher the frequency.
- If the groups / classes are the same width, then the height of the column is proportional to the frequency.



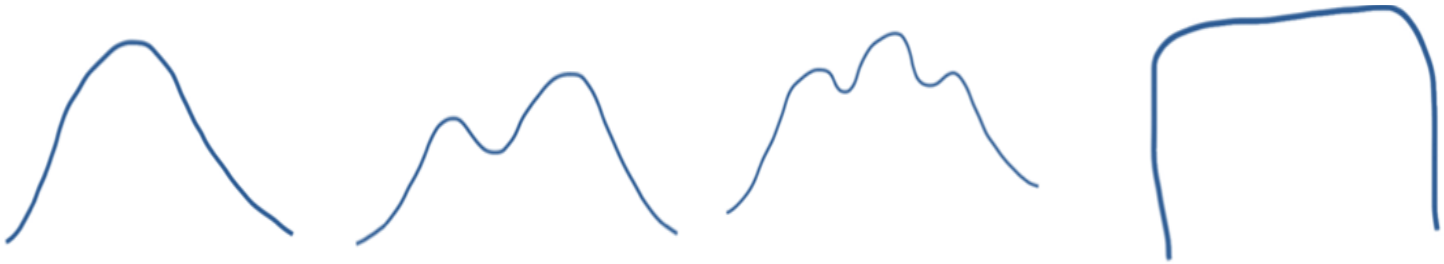
- The width of the column influences the interpretation of the chart.
- N.B.: Numeric-discrete data can be better represented with bar charts.

Common shapes of distributions

Histograms contain information about the…

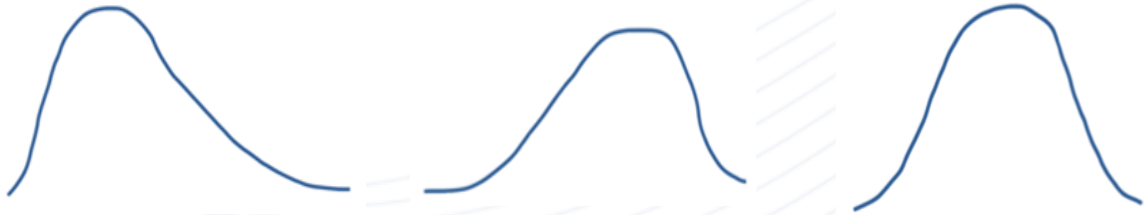
…modality of the distribution, and the

unimodal bimodal multimodal uniformly distributed



…skewness or symmetry of the distribution.

right skewed left skewed symmetrical



Mean, variance and standard deviation

- The (arithmetic) **mean** $\bar{x}$ of a sample $x_1, x_2, \ldots, x_n$, also denoted as average, is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

- The **variance** s^2 is roughly the average squared deviation from the mean value. It has the disadvantage that it cannot be interpreted in the units of the original data. Is is defined as:

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

- The **standard deviation** s is calculated as square root of the variance. It can bei interpreted in the units of the original data:

$$s = \sqrt{s^2}$$

Percentiles (Quantiles)

The interpretation of the p%-percentile (p%-quantile) is: p% of all values in a sample are at most this large, and (1-p)% of all values in a sample are at least this large.

Special percentiles

- 0%-percentile: Minimum.
- 25%-percentile: Lower quartile (first quartile Q1).
- 50%-percentile: Median (second quartile Q2).
- 75%-percentile: Upper quartile (third quartile Q3).
- 100%-percentile: Maximum.

The interval between Q1 and Q3 contains the middle 50% of all values, called **inter quartile range (IQR)**: IQR = Q3 - Q1

The interval between the minimum and the maximum represents the maximum possible distance between two arbitrary values, called **range**:
Range = Maximum – Minimum

Example for percentiles: Distribution of wealth (Vermögensverteilung)

Vermögensverteilung[1] in Deutschland

	Individuelle Nettovermögen											Inklusive dem Wert von Kraftfahrzeugen und nach Abzug von Studienkrediten		
	untere Grenze	2002	obere Grenze	untere Grenze	2007	obere Grenze	untere Grenze	2012	obere Grenze	untere Grenze	2017	obere Grenze	untere Grenze	2017	obere Grenze
Gini coefficient	0,768	0,776	0,784	0,790	0,799	0,809	0,769	0,779	0,790	0,769	0,779	0,789	0,749	0,759	0,769
Perzentilsverhältnisse															
p90/p50	13,4	14,1	14,6	13,3	14,5	15,9	11,4	12,8	14,2	12,0	13,2	14,0	9,8	10,5	11,2
p75/p50	6,4	6,6	6,9	5,8	6,4	6,9	5,2	5,9	6,5	5,8	6,1	6,5	4,8	5,0	5,2
Mean value (€)	77 721	80 469	83 233	78 417	82 189	85 948	81 126	84 530	87 933	98 745	102 868	107 026	104 246	108 449	112 620
Percentiles (€)															
p99	723 280	767 952	823 932	740 579	812 943	888 565	782 080	839 408	899 442	928 876	1 035 000	1 153 155	941 178	1 045 680	1 167 932
p95	310 922	323 941	335 968	308 572	324 148	340 145	315 676	331 800	349 719	389 606	406 365	427 219	400 576	419 766	437 215
p90	205 187	211 867	218 737	201 147	209 789	218 551	210 813	219 100	226 544	254 388	263 500	273 594	267 511	275 770	284 490
p75	95 346	99 568	102 000	88 285	92 482	96 647	97 616	100 190	103 558	118 957	122 792	126 609	125 396	130 040	134 269
Median	14 470	15 000	15 808	13 133	14 520	15 654	15 159	17 120	19 101	18 671	20 010	21 967	24 528	26 260	28 116
p25	0	0	0	0	0	0	0	0	0	0	0	0	1 259	1 590	2 131
p10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p5	−2 914	−1 920	−887	−4 656	−3 960	−3 119	−4 504	−3 718	−2 791	−3 665	−3 000	−2 239	−2 688	−2 044	−1 577
p1	−22 698	−20 255	−18 293	−33 594	−30 000	−24 925	−27 551	−24 374	−20 953	−26 661	−23 107	−19 441	−23 246	−20 360	−18 140
Anteil der Personen mit einem Nettovermögen unter 0 Euro in Prozent	5,3	5,7	6,2	7,1	7,7	8,3	7,1	7,6	8,0	6,4	6,9	7,3	5,9	6,4	6,9
Anteil der Personen mit einem Nettovermögen gleich 0 Euro in Prozent	21,2	21,9	22,6	19,5	20,3	21,1	18,9	19,7	20,4	21,4	22,1	22,9	14,0	14,5	15,1
Nettovermögen[2] insgesamt in Mrd. Euro		5 775			5 918			5 920			7 390			7 776	

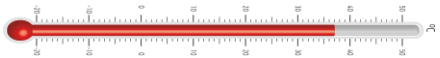
1 Individuelle Nettovermögen der Personen ab 17 Jahren in Privathaushalten, ohne Personen der Flüchtlingssamples M3 bis M5.
2 Ohne Top-Coding.
Anmerkungen: Statistisch signifikante Veränderungen gegenüber dem jeweiligen Erhebungsjahr zuvor sind grün markiert. Untere bzw. obere Grenze geben die Schwellenwerte eines 95-Prozent-Konfidenzintervalls an.

Quelle: SOEPv34, mit 0,1 Prozent Top-Coding; eigene Berechnungen.

https://www.diw.de/documents/publikationen/73/diw_01.c.679972.de/19-40-1.pdf, p. 4, accessed 30.01.2020.

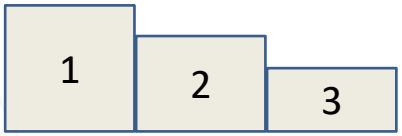
Feasible descriptive figures dependent on the measurement scales

Metric (numerical) scales



- Arithmetic mean, variance and standard deviation.
- Median and percentiles.
- IQR and range.
- Modal value (most frequent value).
- Geometric mean.

Ordinal scale



- Median and percentiles.
- Modal value.

Nominal scale

- Modal value.

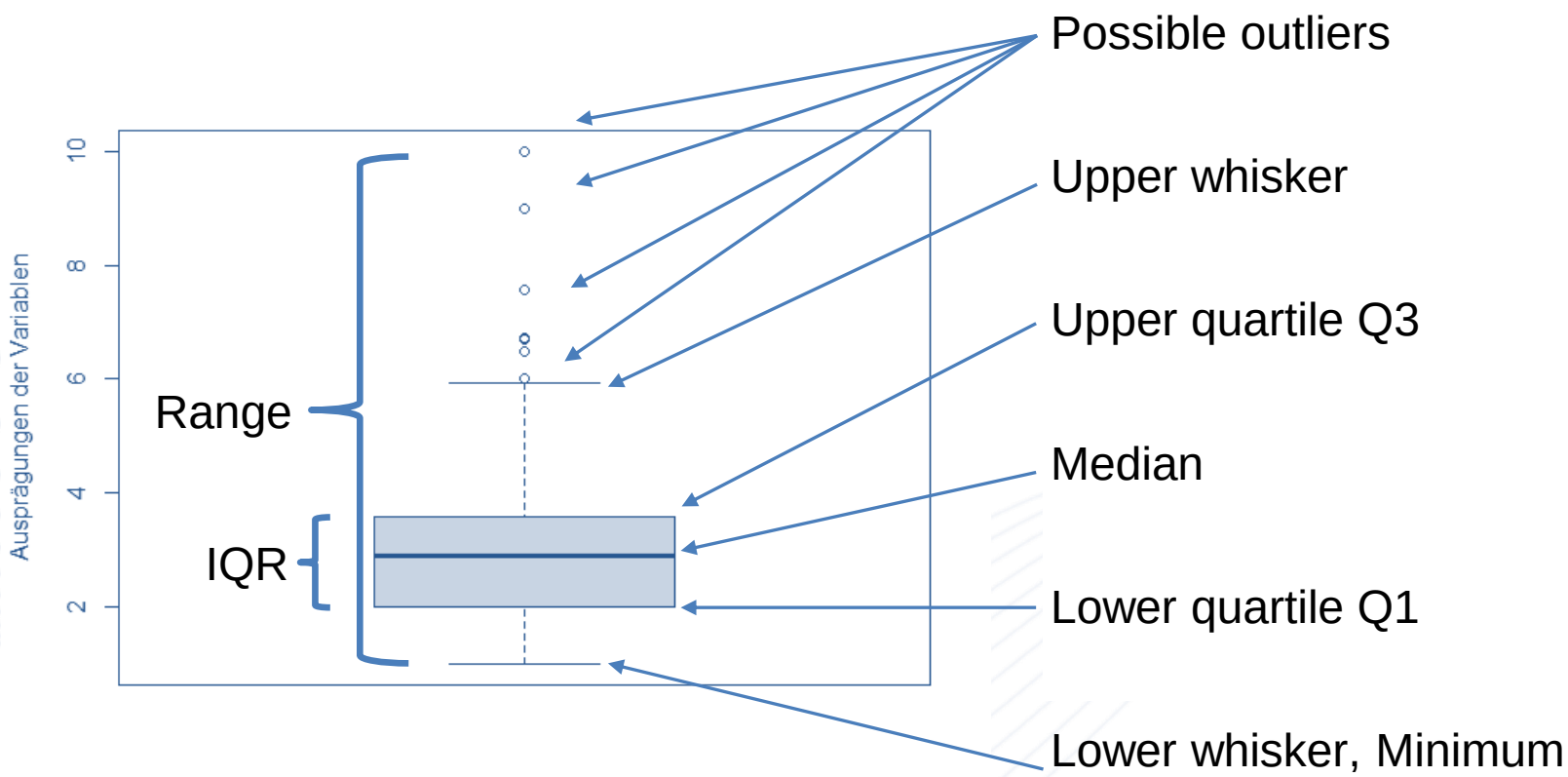
high

Information and interpretation quality

low

Boxplots (box-whiskers-plot)

- A **Boxplot** displays graphically the most important percentiles and thus enables a view on the **dispersion of the values**.
- Boxplots are an important tool for the detection of **outliers**.



Important R-commands for summarizing data

summary()

describe() from the package psych.

favstats() from the package mosaic.

hist()

boxplot()

Boxplot() (with a capital letter B) from the package car.

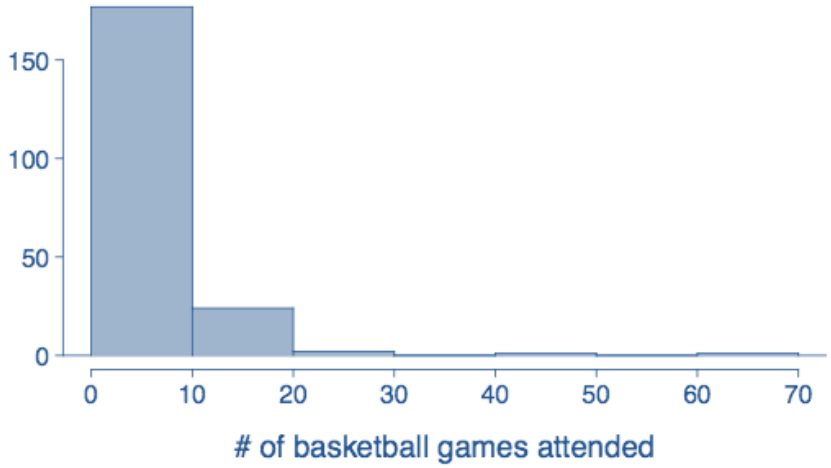
Exercise

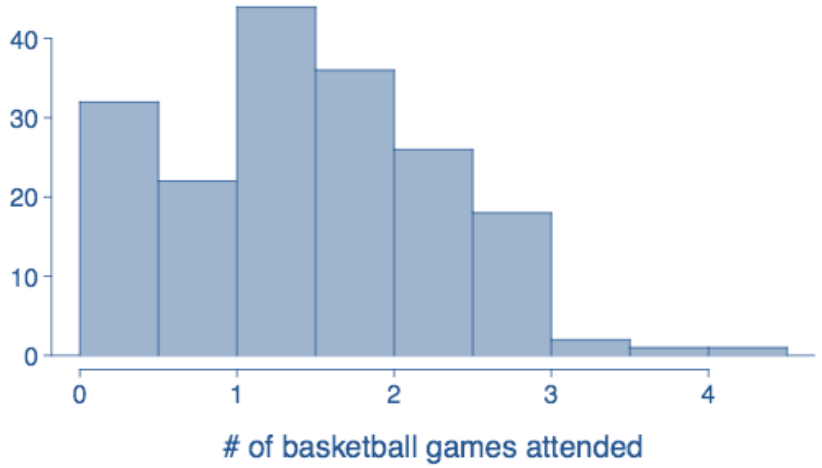
Practice the key figures and graphics for descriptive statistics with the tips data set.

Extremely skewed data

If data is extremely skewed, a transformation can make the analysis easier. A common transformation is the **logarithmic transformation**.

The histogram on the left shows the distribution of the number of basketball games attended by students. The histogram on the right shows the distribution of the logarithmic number of games attended.





Advantages and disadvantages of transformations

Skewed data is easier to analyze when it is transformed, as outliers become less significant after a transformation.

Number of games attended	70	50	25	…
ln(number of games attended)	4,25	3,91	3,22	…

However, the interpretation of the results is often more difficult if not impossible, as the logarithmized values cannot be interpreted directly (back-transformation required).

Exercise

Check whether the logarithms reduces the skewness of the tip.

R command: log()

Procedure:

1. Draw the histogram without transformation.
2. Draw the histogram with transformation.
3. Compare and evaluate both histograms.

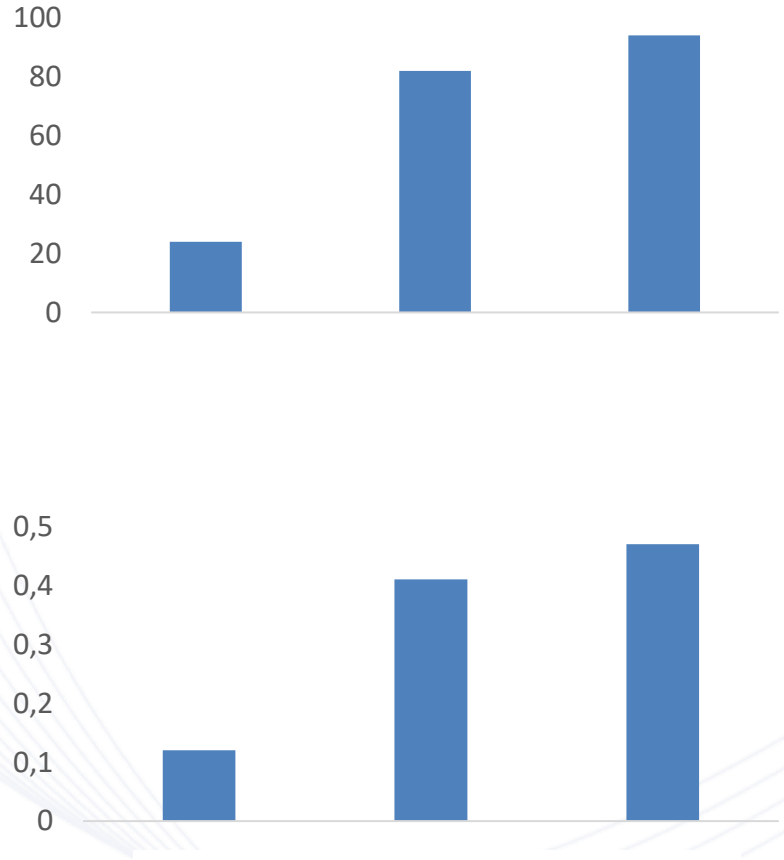
Exercise

In the company “PLT - PaidLeasureTime” the following salaries (in thousands of €) are paid to the employees: 87, 34, 98, 66, 51, 55, 45, 78, 104, 55, 49, 66, 47, 56, 356, 15, 45, 55. Please solve the following tasks.

1. Graphically check the shape of the distribution and whether there are any outliers.
2. Calculate the mean, median, standard deviation and inter quartile range.
3. Which key figure(s) would you use to describe the salaries and why?

2.3 Univariate descriptive statistic for categorical data

Frequency distribution and bar chart



- A **frequency distribution** is suitable for summarizing categorical data.
- The number of cases in which a certain characteristic occurs is called the **absolute frequency**.
- The proportion of the total number of cases in which a certain characteristic occurs is called the **relative frequency**.
- **Bar charts** are suitable for the graphical representation of a frequency distribution
- Bar charts for absolute and relative frequencies are identical except for the scaling of the Y-axis.

Exercise

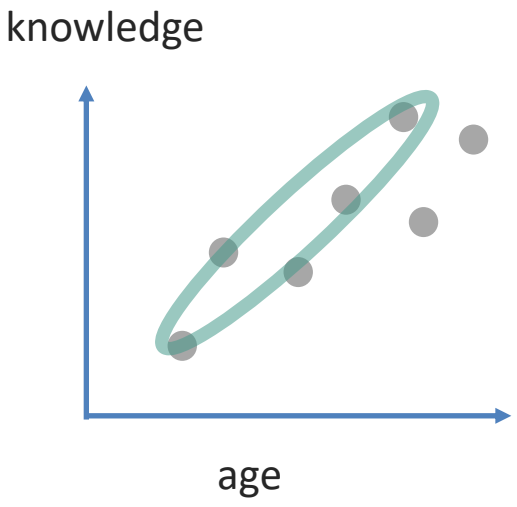
Create frequency distributions for the variables gender (sex) and group size (size) from the tips data set. How do the two variables differ?

R commands: table() and barplot()

2.4 Correlation analysis for numerical data

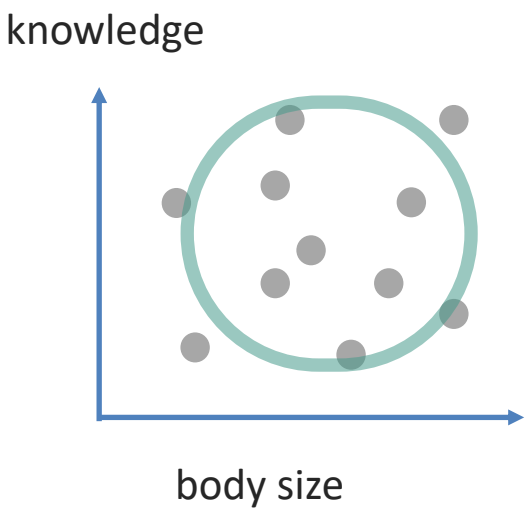
Strong and weak correlation

Strong correlation



- Strong correlation in this example means: The higher the age, the more extensive the knowledge.
- Strong correlation is graphically represented as a “cigar” (narrow ellipse).

Weak / no correlation

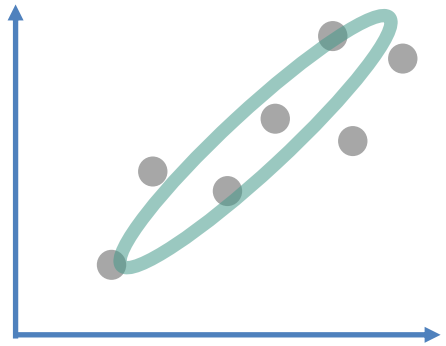


- Weak or no correlation in this example means: Small body sizes occurs with extensive knowledge as well as with little knowledge. The same is true for large body sizes.
- Weak or no correlation is graphically represented as a “cake” (sphere or rectangle).

Positive and negative correlation

Positive correlation

knowledge

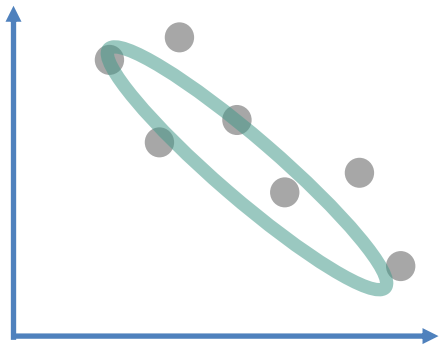


age

- High values in age correspond with high values in knowledge. Low values in age correspond with low values in knowledge.
- The ellipse is inclined upwards.

Negative correlation

leasure time



time spent studying

- High values in time spent studying correspond with low values in leasure time. Low values in time spend studying correspond with high values in leasure time.
- The ellipse is inclined downwards .

Pearson correlation coefficient

(Pearson) correlation coefficient r_{xy}:

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

Cov_{xy}	covariance of x and y
s_x	standard deviation of x
s_y	standard deviation of y
n	number of observations

- The (Pearson) correlation coefficient r_{xy} is insensitive to scale differences in the variables examined.
- Its value is always between -1 and 1.
- The sign (+ -) indicates the direction (positive or negative correlation), the specific value (from |0| to |1|) indicates the strength of the correlation.
- The (Pearson) correlation coefficient is only suitable for metrically scaled variables and is susceptible to outliers.
- The (Pearson) correlation coefficient is only suitable for linear relationships (and not, for example, curvilinear relationships).

Anscombes quartett – descriptive figures

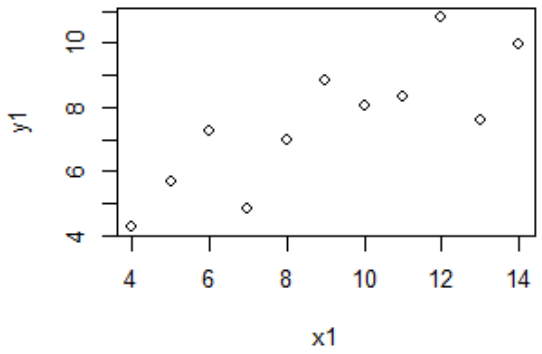
In 1973, the statistician Francis Anscombe published an example with four pairwise measurement series x and y whose mean values and standard deviations (separately for the x and y measurement series) as well as their pairwise correlation coefficients (of the four x-y pairs) are identical.

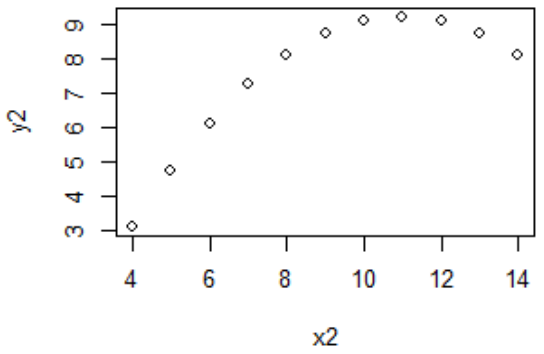
	x1	x2	x3	x4	y1	y2	y3	y4
	10	10	10	8	8,04	9,14	7,46	6,58
	8	8	8	8	6,95	8,14	6,77	5,76
	13	13	13	8	7,58	8,74	12,74	7,71
	9	9	9	8	8,81	8,77	7,11	8,84
	11	11	11	8	8,33	9,26	7,81	8,47
	14	14	14	8	9,96	8,1	8,84	7,04
	6	6	6	8	7,24	6,13	6,08	5,25
	4	4	4	19	4,26	3,1	5,39	12,5
	12	12	12	8	10,84	9,13	8,15	5,56
	7	7	7	8	4,82	7,26	6,42	7,91
	5	5	5	8	5,68	4,74	5,73	6,89
mean	9,00	9,00	9,00	9,00	7,50	7,50	7,50	7,50
standard deviation	3,16	3,16	3,16	3,16	1,94	1,94	1,94	1,94
correlation	0,82	0,82	0,82	0,82				

→ Based on the descriptive key figures, one would assume four identical pairs of measurement series.

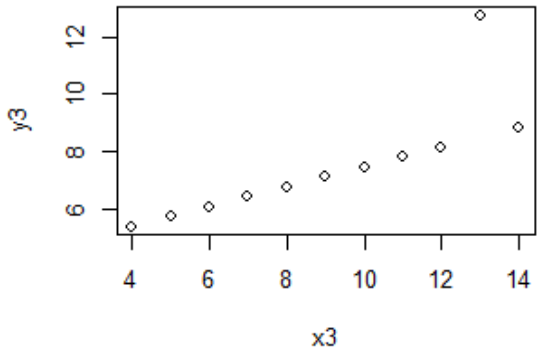
Anscombes quartett – scatterplots

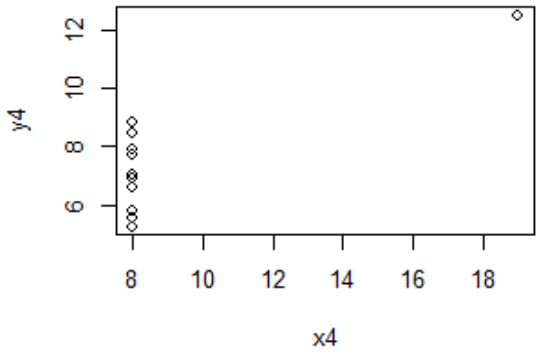
The differences only become apparent through the graphical representations in the scatterplots.





→ **If possible, always do both draw the scatterplot and calculate the correlation coefficient!**





The Spearman rank correlation coefficient

- The rank R_i indicates the position of a value in the list of all values sorted in ascending order. The observation number is placed in brackets as notation for ranks: $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$

Spearman rank correlation coefficient r_{RxRy}:

$$r_{RxRy} = \frac{\sum_{i=1}^{n}(R_{xi} - \overline{Rx}) \cdot (R_{yi} - \overline{Ry})}{n \cdot s_{Rx} \cdot s_{Ry}} = \frac{cov_{RxRy}}{s_{Rx} \cdot s_{Ry}}$$

- The interpretation of the Spearman rank correlation is identical to the Pearson correlation: The sign (+ -) indicates the direction (positive or negative correlation), the value (from |0| to |1|) indicates the strength of the correlation.
- The Spearman rank correlation is also suitable for ordinally scaled variables and is robust against outliers.
- The Spearman rank correlation is suitable for correlations, i.e. linear and some curvilinear correlations.

Exercise

Create a scatterplot for the relationship between the amount of the restaurant bill (total_bill) and the tip amount (tip).Compare the values for the Pearson correlation with the Spearman correlation for the strength of the correlation between the two variables. Are there any outliers, non-linear correlations or similar?

R commands: plot(), cor() and ?cor() for held with the cor()-function.

2.5 Correlation analysis for categorial data

Contingency tables

A table that combines data for two categorical variables is called a **contingency table**.

The contingency table below shows the distribution of gender among students and whether or not they are looking for a partner while attending college.

	Looking for a partner		
	No	Yes	Total
Gender	Female	10	90	100
	Malech	40	10	50
	Total	**50**	**100**	**150**

Is there a correlation between gender and the fact that students are looking for a partner?

Exercise

What should the table look like if there is no correlation between gender and the fact that students are looking for a partner?

	Looking for a partner		
	No	Yes	Total
Gender	Female	?	?	100
	Male	?	?	50
	Total	**50**	**100**	**150**

Contingency tables

This contingency table is a little more complex…

The contingency table below shows the distribution of gender among students and whether or not they are looking for a partner while attending college.

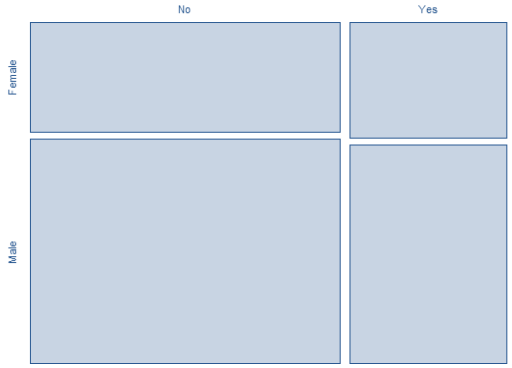
	Looking for a partner		
	No	Yes	Total
Gender			
Female	86	51	137
Male	52	18	70
Total	**138**	**69**	**207**

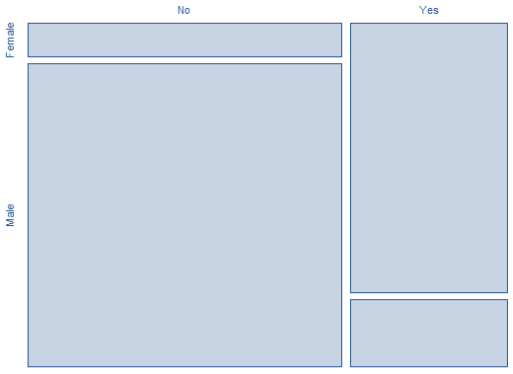
Is there a correlation between gender and the fact that students are looking for a partner?

Mosaicplots

A **mosaicplot** visualizes the correlation between two categorical variables.

If the ratios of the areas in the rows or columns are the same, then there is no correlation between the two variables.The mosaicplot shows a “Scandinavian cross”.





If the ratios of the areas in the rows or columns are different, then there is a correlation between the two variables.

The contingency coefficient

- A correlation exists if the numerical ratio in the rows or columns is different, and no correlation exists if the numerical ratio in the rows or columns is (approximately) the same.
- This approach is imprecise for several reasons:
  - At what point can we speak of a correlation?
  - How can the strength of a correlation be measured?
- The **contingency coefficient C** is a measure of the strength of the correlation between two categorical variables. It takes values between 0 and 1, whereby the extreme values are never reached. The greater C, the stronger the correlation between the two variables.
- C has the disadvantage that its value depends on the complexity of the contingency table. The more cells the table has, the closer gets C to 1. The **corrected contingency coefficient C_{corr}** is independent from the complexity of the table.
- Rule of thumb: C <= 0,2: no / weak correlation
  C >= 0,6: strong correlation

Exercise

Use the tips dataset for this exercise.

1. Calculate a contingency table for the correlation between smoking behavior and gender and assess whether there is a correlation.

2. Calculate the contingency coefficient C for this contingency table.

R commands: table(), mosaicplot() and ContCoef(table, correct = TRUE / FALSE) from the DescTools package.

3. Probability theory and the normal distribution

3.1 Case study: Discrimination

Gender discrimination

- In 1972, as part of a study, 48 male managers of a bank were asked on the basis of a personnel file whether the person described should be promoted to head a branch with routine tasks.
- The personnel files differed only in that one half described a male person and the other half an otherwise identical female person.
- A random decision was made as to which of those responsible received a “male” document and which a “female” document.
- Of the 48 application documents, 35 were recommended for promotion.
- The study investigated the extent to which women are disadvantaged.

Rosen and Jerdee (1974): Influence of sex role stereotypes on personnel decisions, Journal of Applied Psychology, 59 (1).

Data

Is there a correlation between promotion and gender?

	Promotion		
Gender	Yes	No	Total
Male	21	3	24
Female	14	10	24
Total	**35**	**13**	**48**

% of males to be promoted: 21 / 24 = 0,875

% of females to be promoted: 14 / 24 = 0,583

Exercise

There is a difference of almost 30 percentage points (29.2% to be exact) between the proportion of “male” and “female” personnel files that were proposed for promotion. Based on this information, which of the following statements is true?

a) If we repeat the experiment, more women will definitely be suggested. The result was a coincidence.

b) Promotion depends on gender. Men are more likely to be promoted, so there is a disadvantage for women in promotions.

c) The difference in proportions is random. This is not proof of discrimination against women in promotions.

d) Women are less qualified than men, which is why they are promoted less often.

Two competing statements

"There is nothing." **(null hypothesis)**

Promotion and gender are independent.

There is no discrimination based on gender.

The observed differences in the proportions are simply coincidence.

"There is something." **(alternative hypothesis)**

Promotion and gender are not independent.

There is discrimination based on gender.

The observed differences in the proportions are not random.

A court hearing is like a hypothesis test

Testing a hypothesis is a bit like a court hearing.

- H_0 : The accused is innocent
  H_A : The accused is guilty
- We present the evidence – collect data.

https://de.wikipedia.org/wiki/Gerichtsverhandlung, accessed 22.07.2020

- Then we evaluate the evidence - “Are this data plausibly random if the null hypothesis is true?”
- If the data is very unlikely, then there is more than reasonable doubt about the null hypothesis.
- We have to decide: How improbable is improbable?

A court hearing is like a hypothesis test (continued)

- If the evidence is not sufficient to rebut the presumption of innocence, then the court decides “not guilty”.
  - The court does not say that the defendant is innocent, but only that the evidence is insufficient.
  - The defendant may indeed be innocent, but the court cannot be sure.
- Statistically speaking, we cannot reject the null hypothesis.
  - We never say that the null hypothesis is true because we simply do not know.
  - Hence never: “Accepting the null hypothesis”.

A court hearing is like a hypothesis test (continued)

- In a court hearing, the burden of proof lies with the prosecution.
- In a hypothesis test, the burden of proof lies with the unusual statement.
- The null hypothesis is the usual statement (status quo), the alternative hypothesis is the unusual statement for which we need to collect evidence.

Summary: Structure of a hypothesis test

1. Formulate null hypothesis (and alternative hypothesis).

We start with a **null hypothesis (H_0)**, which reflects the current state of knowledge (status quo). This is contrasted with the **alternative hypothesis (H_A)**, which reflects our research question, i.e. what we want to test.

2. Define significance level

We define a **Significance level α**. The significance level represents the maximum probability up to which we want to be wrong if we reject H_0.

3. Determine p-value

We now carry out the hypothesis test under the assumption that the null hypothesis is true. To do this, we determine the probability p of the actually observed data under the assumption that H_0 is true. This probability is called the **p-value**.

4. Make a test decision

If the test results suggest that the data do not provide enough evidence for the alternative hypothesis (i.e. if $p > \alpha$), we retain the null hypothesis. Otherwise ($p \leq \alpha$), we reject the null hypothesis in favor of the alternative.

3.2 Probability theory

Random experiment and probability theory

Probability theory deals with the laws of **random experiments**, i.e. processes in which different results can occur under (apparently) identical conditions.

A **random experiment** is a process which,

- at least theoretically, can be repeated an infinite number of times,
- which has several possible outcomes, and
- in which the specific result is not known in advance.

In general, a **probability** can be defined as a measure to quantify the uncertainty of the occurrence of a certain event in the context of a random experiment.

Lübke/Vogt (2014): Angewandte Wirtschaftsstatistik: Daten und Zufall, Springer Gabler

Examples for random experiments

1. Toss a coin:
   What is the probability for a head?

2. Roll a dice:
   What is the probability for a 6?

3. Investment decision:
   What is the probability for a loss?

https://nerd-wiki.de/allgemein/einhorn-nierensteine-prismatische-metall-wuerfel-von-dndice/, accessed 12.2.2020.

Probability and elementary event

The possible, mutually exclusive results of a random experiment are referred to as **elementary events** or **atomic events.** The set of all elementary events is called the event space, symbol Ω. An event is a subset of Ω.

Dabei ist

- The event A and B is called intersection $A \cap B$
- The event A or B is called unification quantity $A \cup B$
- The complementary event A^c in Ω is the event that occurs if A does not occur

P is defined as probability, if:

- For any subset A of Ω: $0 \leq P(A) \leq 1$
- For the save event: $P(\Omega) = 1$
- For the impossible event: $P(\varnothing) = 0$

Quelle: Lübke/Vogt (2014): Angewandte Wirtschaftsstatistik: Daten und Zufall, Springer Gabler

Laplace probability

The probability P for an event A is the ratio of the number results where A occurs divided by the number of (equally) possible results.

$$\mathrm{P(A)} = \frac{\text{Number of results where A occurs}}{\text{Total number of equally possible results}}$$

Example: Toss a coin

Consider the event A = Head.

$$\mathrm{P(A)} = \frac{\text{Number of results where A occurs}}{\text{Total number of equally possible results}} = \frac{1}{2}$$

Bleymüller/Weißbach (2015), Statistik für Wirtschaftswissenschaftler, Vahlen

Exercise

Determine the following probabilities for a dice with six sides, numbered 1 to 6.

1. The probability that the dice shows a 1.
2. The probability that the dice shows an even number.
3. The probability that the dice shows a 7.
4. The probability that the dice shows a number between at least 1 and at most 6.
5. The probability that the dice does not show a 5.

Constraints of the Laplace probability

In reality, the possible results normally do not have the same probability or cannot be specified and counted. Sometimes even the number of results a specified event occurs cannot be counted.

Examples:

- Throw a dice: For a six-sided dice numbered from 1 to 6, you want to check whether it is fair, i.e. whether all the numbers really have the same probability. How can this be checked?

- Investment decision: What is the probability for a loss?

Empirical probability

The **empirical probability** equals the relative frequency.

$$P(A) = \frac{\text{Number of trials where A occurs}}{\text{Total number of trials}}$$

Whereby (at least mentally) more and more trials are carried out (limit value).

Example:

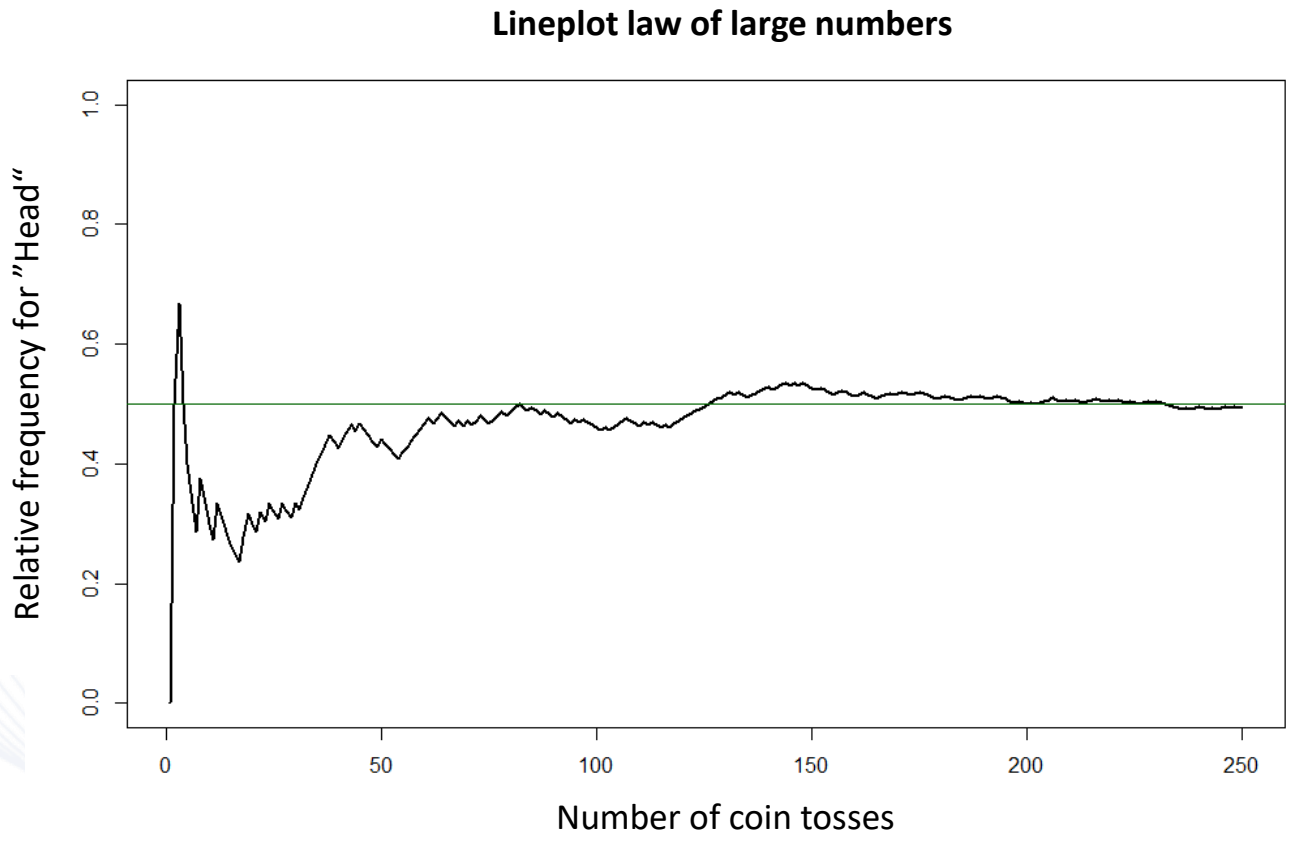
An urn contains an unknown number of M black and white (indistinguishable) balls in total. **What is the probability of drawing a white ball?**

- Draw K (<M) balls (with put back) and count the white balls W.
- The probability of a white ball is then P(white ball) = W / K.
- Repeat this “sufficiently often”.

Bleymüller, Weißbach (2015), Statistik für Wirtschaftswissenschaftler, Vahlen

Law of large numbers (LLN)

If a random experiment (e.g. coin toss) is carried out repeatedly under the same conditions, the empirical relative frequency converges to the theoretical probability.



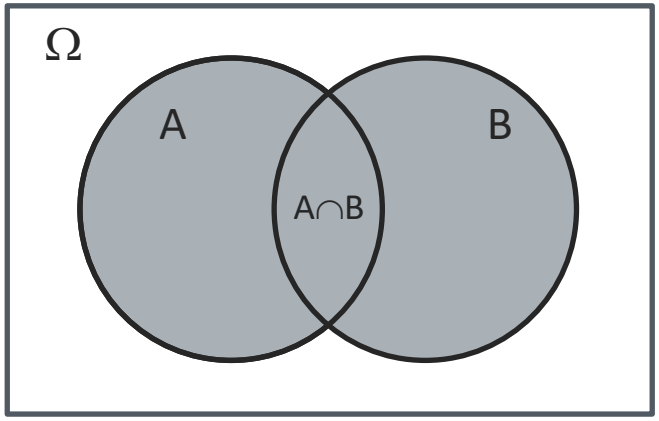
Exercise

1. How can it be checked, at least theoretically, whether a dice is fair or not?
2. Assume there are two hospitals in a town. In the larger hospital, an average of 45 babies is born every day, in the smaller hospital the average is 15 babies. It can be assumed that overall, 50% of all babies are boys, but the exact percentage varies from day to day. Sometimes it is more than 50%, sometimes less than 50%. For one year, each hospital records the days on which more than 60% of the newborns were boys. Which of the two hospitals do you think records more of these days?

For the second exercise: Randow, G.v. (2017): Das Ziegenproblem – Denken in Wahrscheinlichkeiten, 10. Auflage, Rowohlt, p. 79

Adding probabilites for events A and B

Not mutually exclusive events



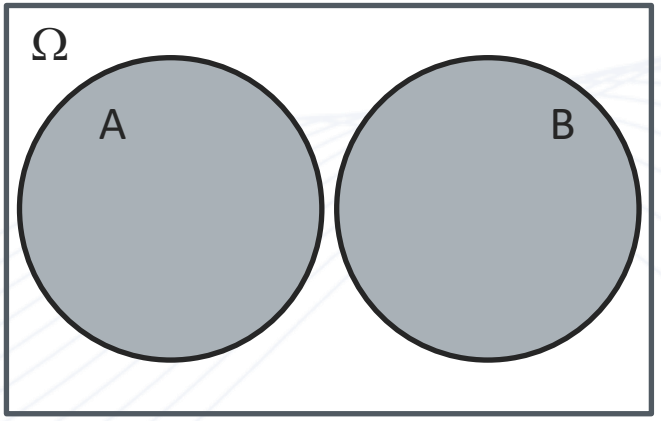
The probability of at least one of two events A or B occurring is the probability of A plus the probability of B minus the probability of the intersection (would otherwise be counted twice):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Mutually exclusive events

For mutually exclusive events, the probabilities for A and B can be added directly, the intersection is empty:

$$P(A \cup B) = P(A) + P(B)$$

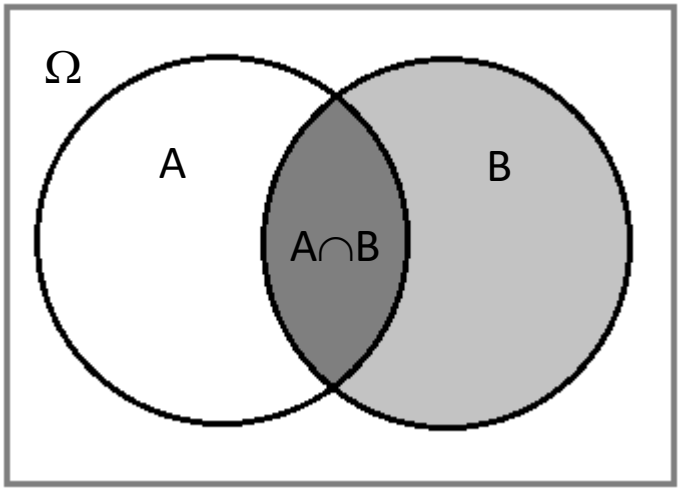


Lübke/Vogt (2014): Angewandte Wirtschaftsstatistik: Daten und Zufall, Springer Gabler

Conditional probability and multiplication of probabilities

The conditional probability for an event A given an event B P(A|B) is the probability for A given B has already occurred:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



From this, the following rule for the multiplication of probabilities can be derived:

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

Lübke/Vogt (2014): Angewandte Wirtschaftsstatistik: Daten und Zufall, Springer Gabler

Stochastic independence and multiplication of probabilities

Two events A and B with P(A)>0 and P(B)>0 are called **independent** if the following applies:

$$P(A \mid B) = P(A)$$

$$\Leftrightarrow P(B \mid A) = P(B)$$

$$\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

The occurrence of one event does not influence the probability of occurrence of the other.

The probability for two independent events A and B both occur together can be calculated as the product of the two individual probabilities.

Summary: Caculation rules für probabilies

Adding probabilites: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multiplying probabilities: $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$

Complementary probability: $P(A^c) = 1 - P(A)$

Monotony: $A \subset B \Rightarrow P(A) \leq P(B)$

3.3 Random variable and normal distribution

Random variable

- A **random variable** X is a characteristic (variable) whose value is the **result of a random experiment**. In mathematical terms, each element of the event space Ω is assigned a real number: X: Ω → R.
- The **probability distribution** indicates the probability with which the random variable takes on the values (probabilities of observations).
- Observations (real data) are **realizations of random variables**.

Examples for probability distribution: Coin and dice

A (probability) distribution specifies the associated probability for each possible result of a random experiment.

Example: Toss a coin:

Possible result	Probability (probability distribution)	Cumulated probability (distribution functionn)
Head	1/2	1/2
Tail	1/2	1

Example. Throw a dice:

Possible result	Probability (probability distribution)	Cumulated probability (distribution functionn)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	1

Exercise

What does the probability distribution for the number of “tails” tossed look like for two coin tosses?

Distribution function

The **distribution function** *F* at the point x indicates the probability with which a random variable X takes on values less than or equal to x:

$$F(x) = P(X \leq x)$$

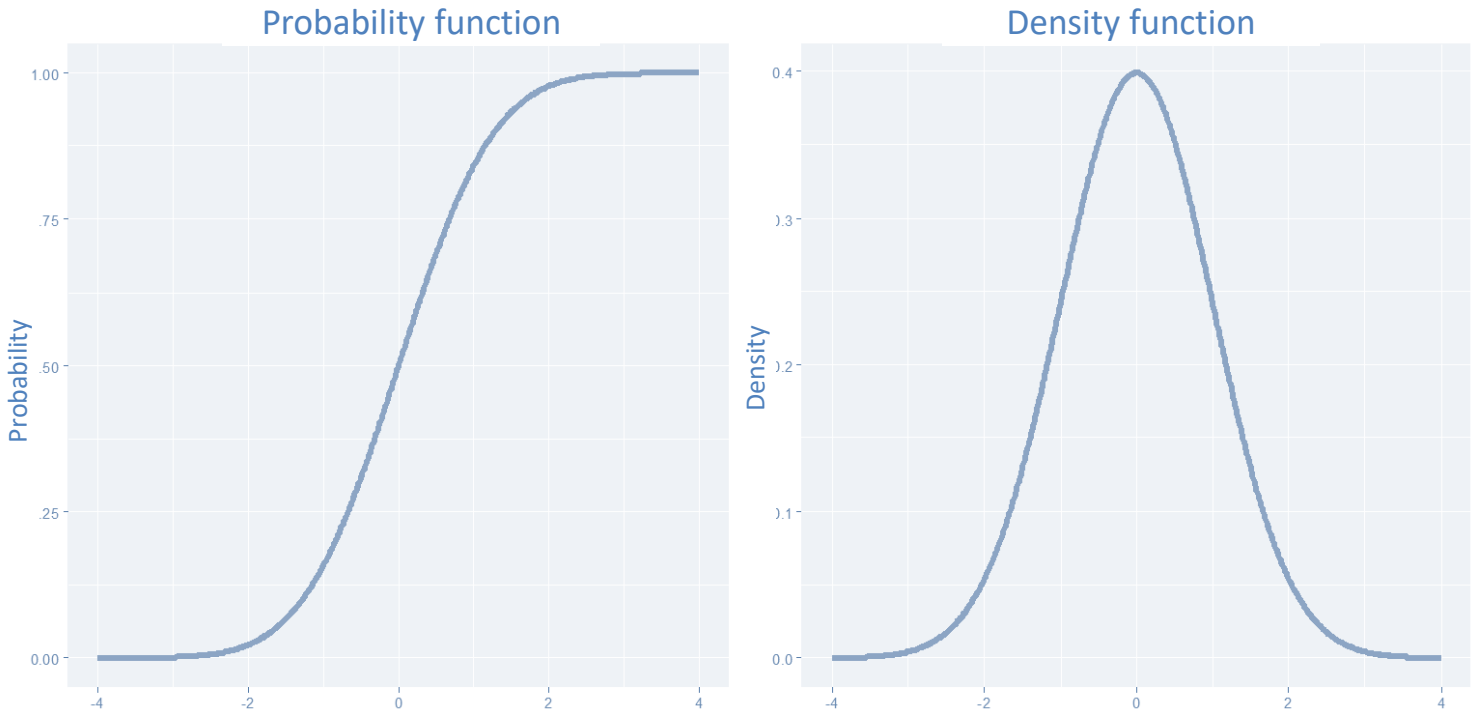
According to the calculation rules for probabilities:

- $P(X \leq b) = F(b)$
- $P(X > a) = 1 - F(a)$
- $P(a < X \leq b) = F(b) - F(a)$

The density function *f* (probability density) denotes the…

- Discrete: …probability for individual values x (probability distribution),
- Continuous: …derivative of the distribution function according to x.

Probability function and density function



- The probability can be read directly from the distribution function.
- For the density function, the area under the curve corresponds to the probability. The total area is 1 (= 100% probability).

Probability distribution and relative frequency distribution

If the number of possible results cannot be specified, the probability distribution can no longer be determined. This can be approximated empirically using a **relative frequency distribution**.

Example:

An urn contains an unknown number of M black and white (indistinguishable) balls in total.

What is the probability of drawing one, two, three, ..., six white ball(s)?

Draw K balls sufficiently often (e.g. 1000 times), count the white balls W:

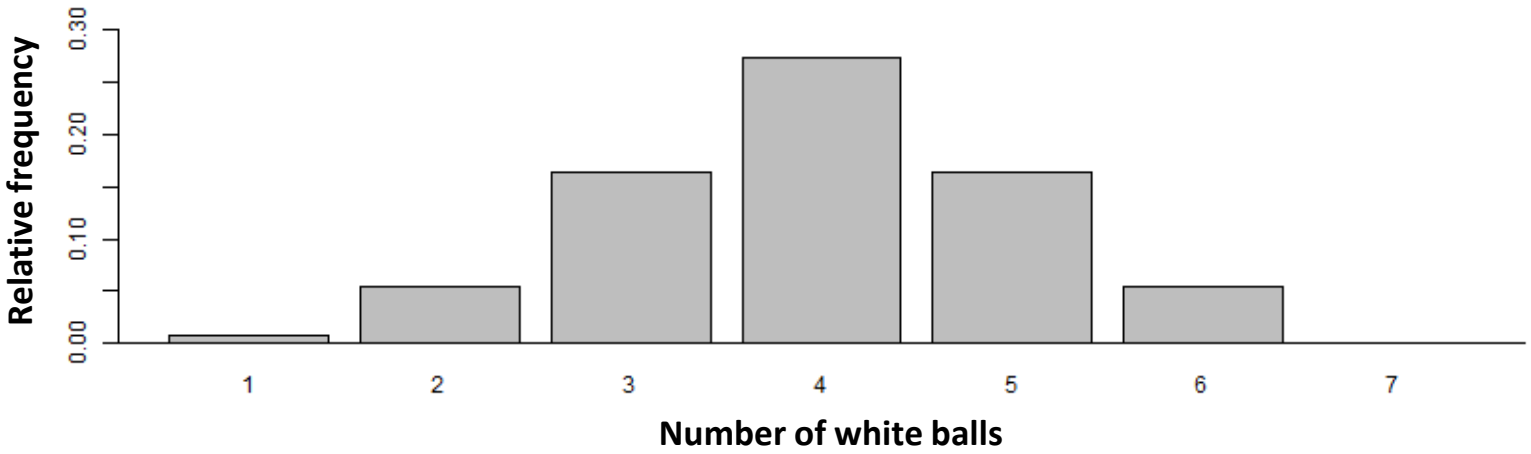
$$\mathrm{P}(1 \text{ white ball}) = \frac{\text{Number of draws with 1 white ball}}{1000}$$

$$\vdots$$

$$\mathrm{P}(6 \text{ white balls}) = \frac{\text{Number of draws with 6 white balls}}{1000}$$

Probability distribution and relative frequency distribution

The following relative frequency distribution is obtained:

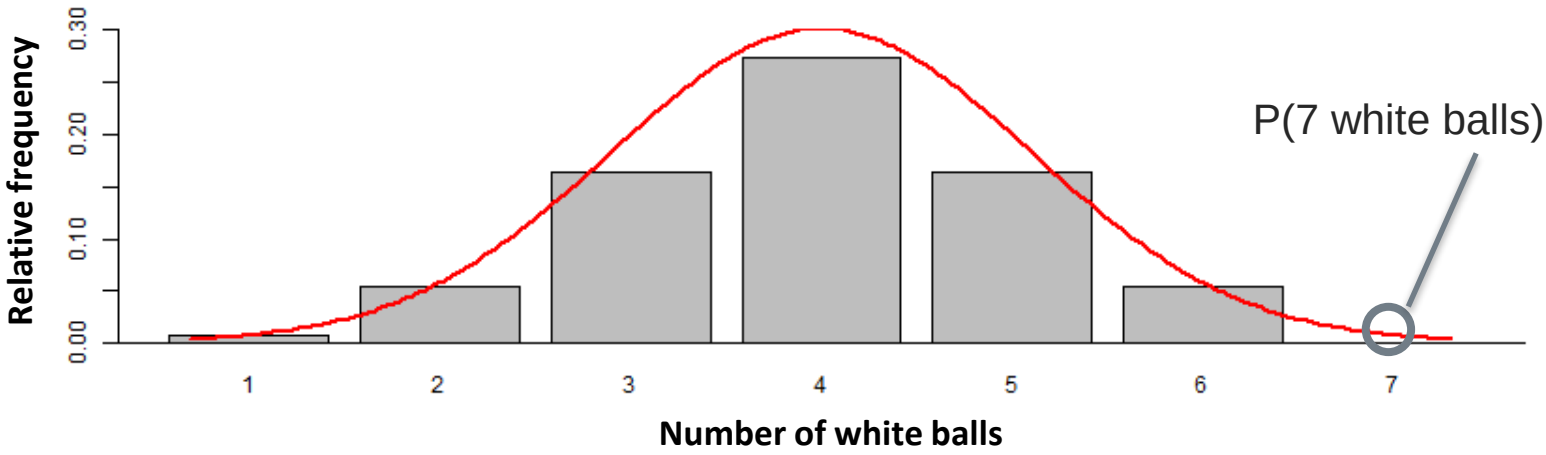


Is it possible to determine P(7 white balls) even if they have not been observed?

→ Yes, find a function that is a good approximation for the relative frequency distribution.

Probability distribution and relative frequency distribution

A function for approximating the relative frequency distribution is the **normal distribution**:

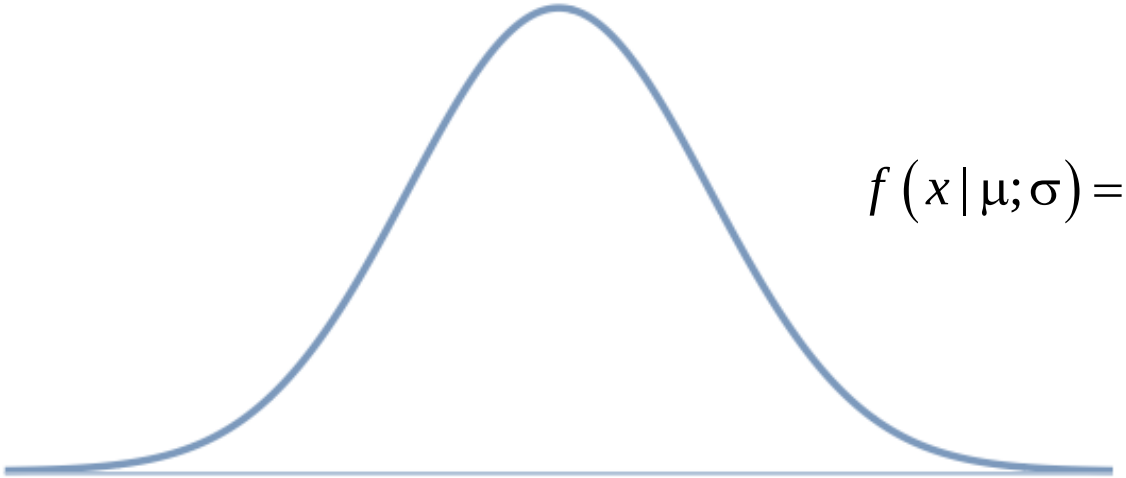


The normal distribution is particularly suitable for approximating symmetrical relative frequency distributions.

Other names are **Gaussian distribution** or **Gaussian bell-shaped curve**.

The normal distribution

- Unimodal and symmetrical, bell-shaped curve.
- Many variables are approximately normally distributed, but never exactly. This makes the normal distribution the most frequently used continuous distribution.
- Many other distributions converge to the normal distribution (central limit theorem).
- Designation N(μ, σ) → Normal distribution with mean value (expected value) μ and standard deviation σ.

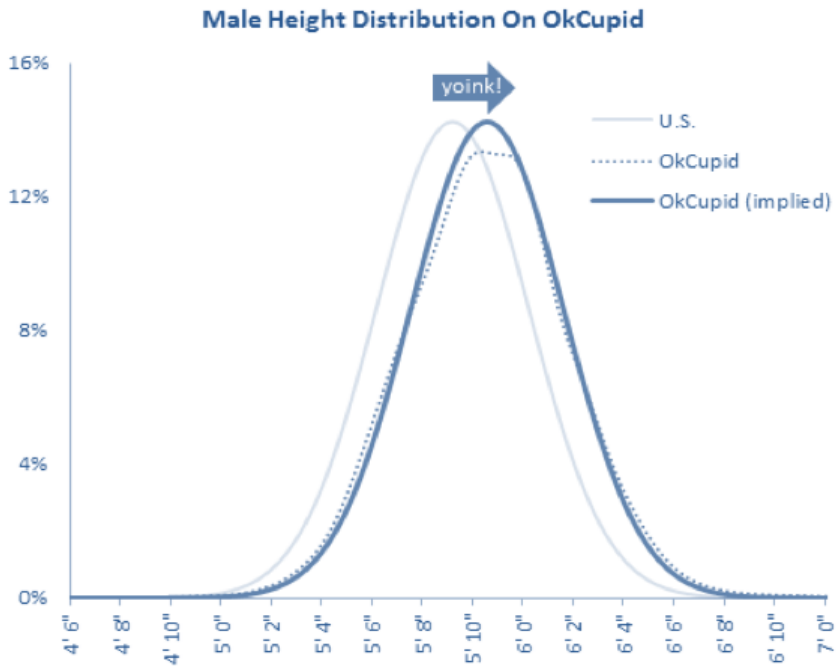


$$f\left(x \mid \mu;\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Example: Body size of men

"The size of men on OKCupid [dating site] is approximately normally distributed as expected - except that it's shifted a little to the right. Apparently, guys almost always make themselves a little taller."

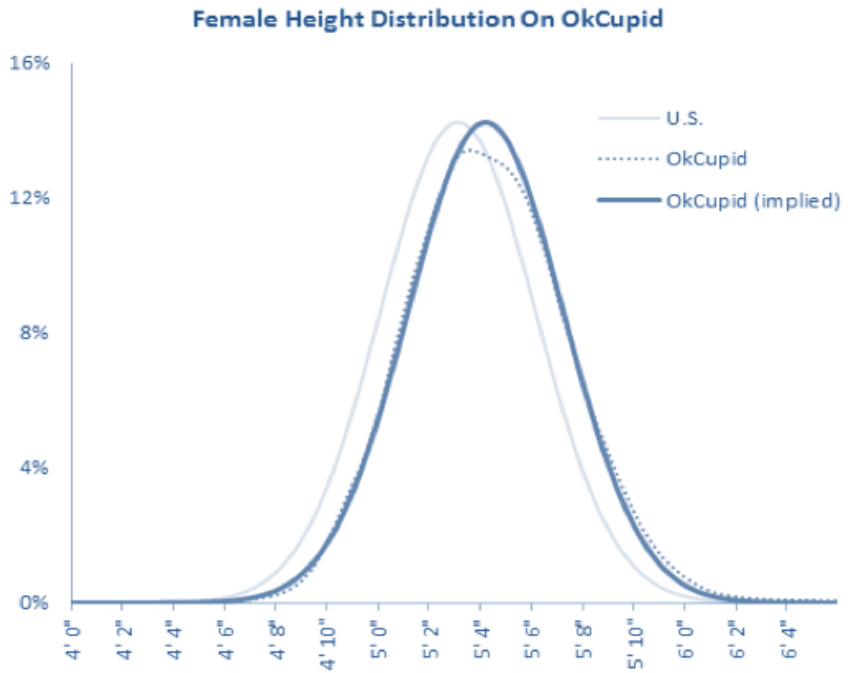
"Also, you can see a little vanity: from about 5' 8" (172.72 cm), the dotted curve goes even further to the right. This means that once guys get close to 6 feet (182.88 cm), they round up a bit more to reach that psychologically desirable benchmark."



https://theblog.okcupid.com/the-big-lies-people-tell-in-online-dating-a9e3990d6ae2, accessed 17.02.2020

Example: Body size of women

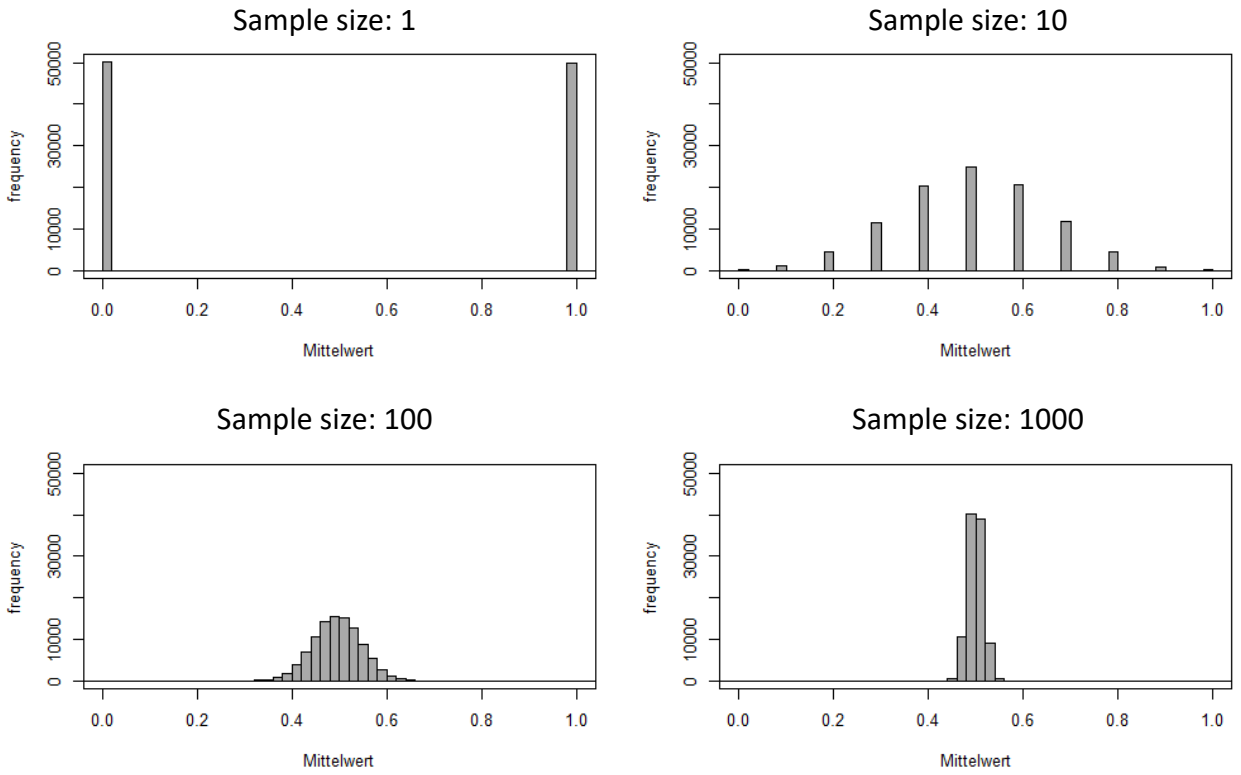
"When we looked at the women's data, we were surprised to find the same exaggeration, just without the jolt towards a benchmark".



https://theblog.okcupid.com/the-big-lies-people-tell-in-online-dating-a9e3990d6ae2, accessed 17.02.2020

Central limit theorem

Histograms: Central limit theorem for coin tosses and relative frequency of "Head"



If n independent, identically distributed (i.i.d.) random variables of an arbitrary distribution are added up, the distribution converges to a normal distribution as n increases.

The parameters of a normal distribution

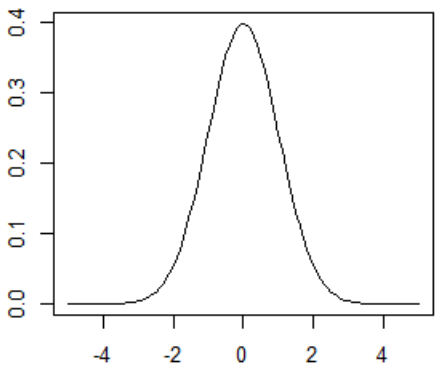
μ (mu) and σ (sigma) are theoretical parameters that must first be estimated for practical applications.

- **μ**: "Mean value" of the normally distributed random variable: **Expected value.**
- **σ**: Standard deviation of the normally distributed random variable (σ>0).

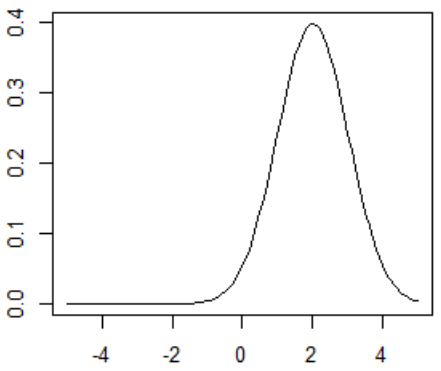
If μ and σ are known, the shape of the normal distribution curve can be determined exactly. μ can be estimated by the sample mean value $\bar{X}$ and σ can be estimated by the sample standard deviation s.

The parameters of a normal distribution

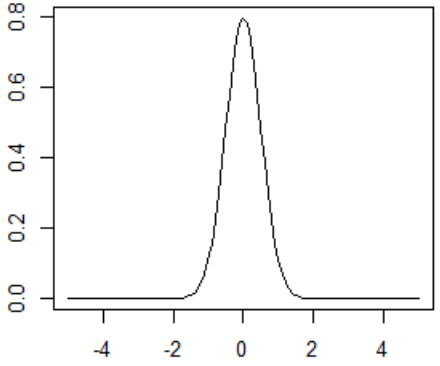
Normal distribution μ=0, σ=1



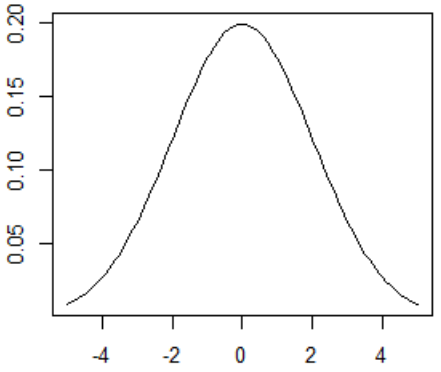
Normal distribution μ=2, σ=1



Normal distribution μ=0, σ=.5



Normal distribution μ=0, σ=2



μ shifts the bell curve on the axis, and σ stretches or compresses the curve.

The standard normal distribution

The transformation $z=(x-\mu)/\sigma$ can be used to transform a normally distributed variable X from any normal distribution into a standard normally distributed Z with $\mu=0$ and $\sigma=1$ (→ standardization, z-transformation).

The values of the distribution function of a standard normal distribution $\Phi(z)$ (Phi) are often printed in tables or stored in programs (e.g. R).

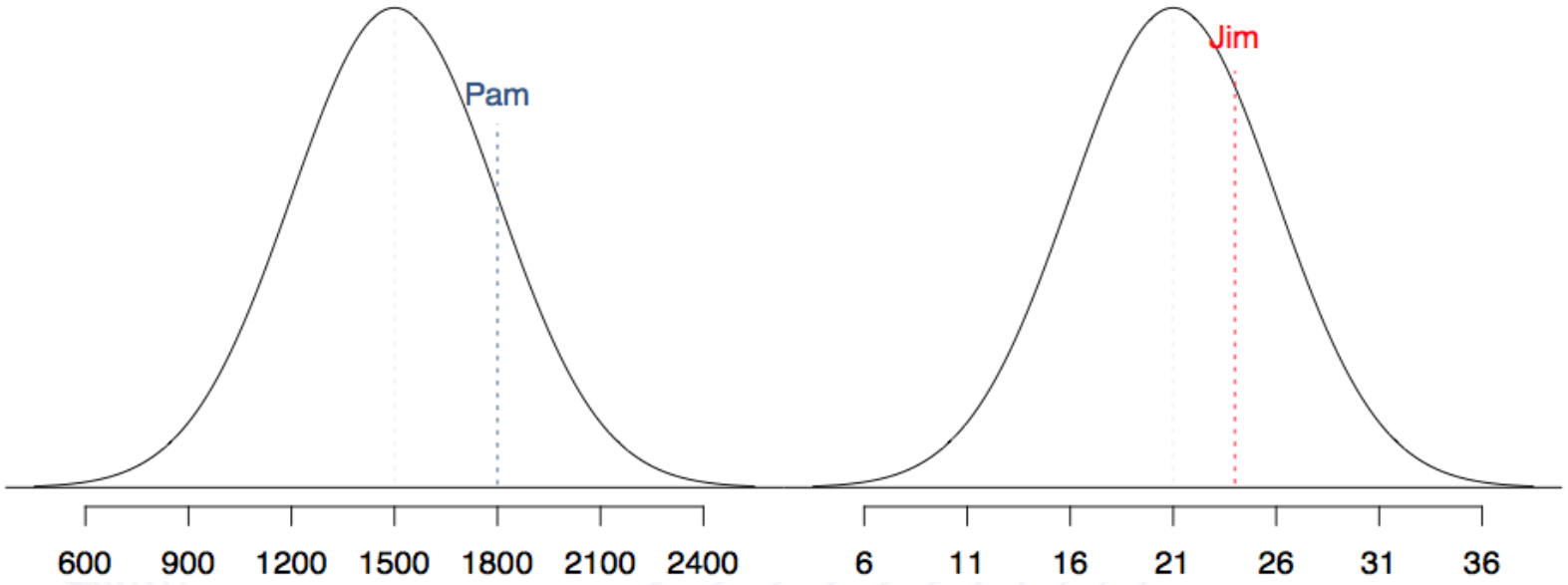
The standard normal distribution is symmetrical around 0. This means that $\Phi(-z) = 1 - \Phi(z)$.

$$\Phi(z) = F(z) = \int_{-\infty}^{z} \underbrace{\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}t^2)dt}_{\textbf{Density function}}$$

Example: Aptitude tests for university studies

SAT test scores are approximately normally distributed with a mean of 1500 and a standard deviation of 300. ACT test scores are approximately normally distributed with a mean of 21 and a standard deviation of 5.

Which of two college applicants scored higher on the standardized test than the other? Pam, SAT-score of 1800, or Jim, ACT-score of 24?

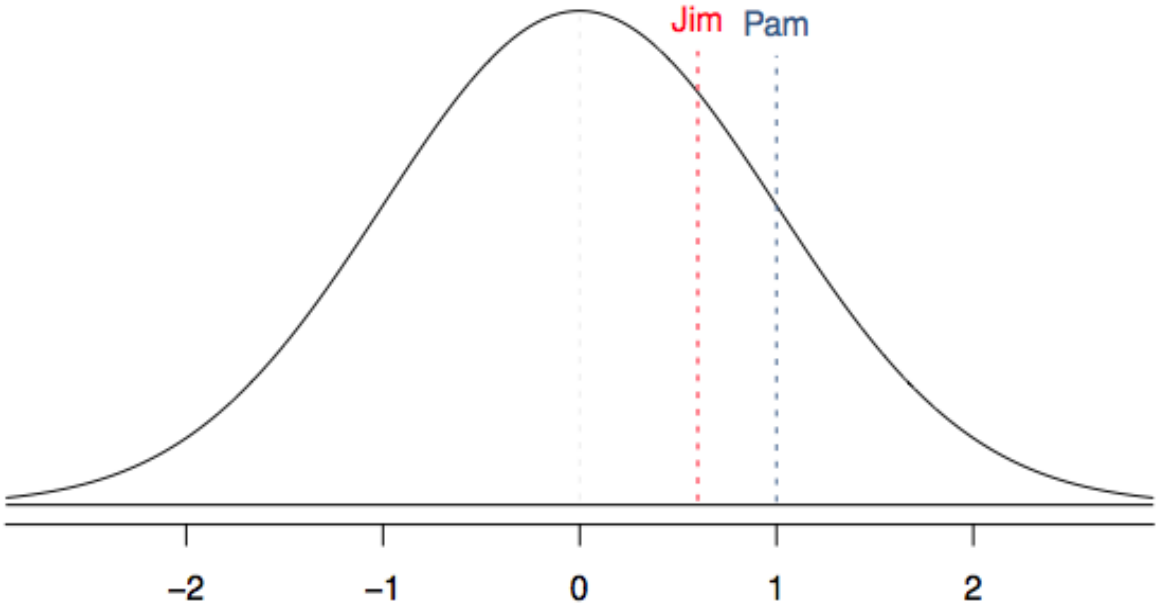


SAT: Scholastic Assessment Test; ACT: American College Test. These are two standardized American tests to assess the study ability of applicants for a Bachelor's degree.

Example: Standardization and z-values

It is impossible to compare the two raw scores directly, but it is possible to compare how many standard deviations the two scores are away from their mean value:

- Pam's score is z = (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is z = (24 - 21) / 5 = 0.6 standard deviation above the mean.

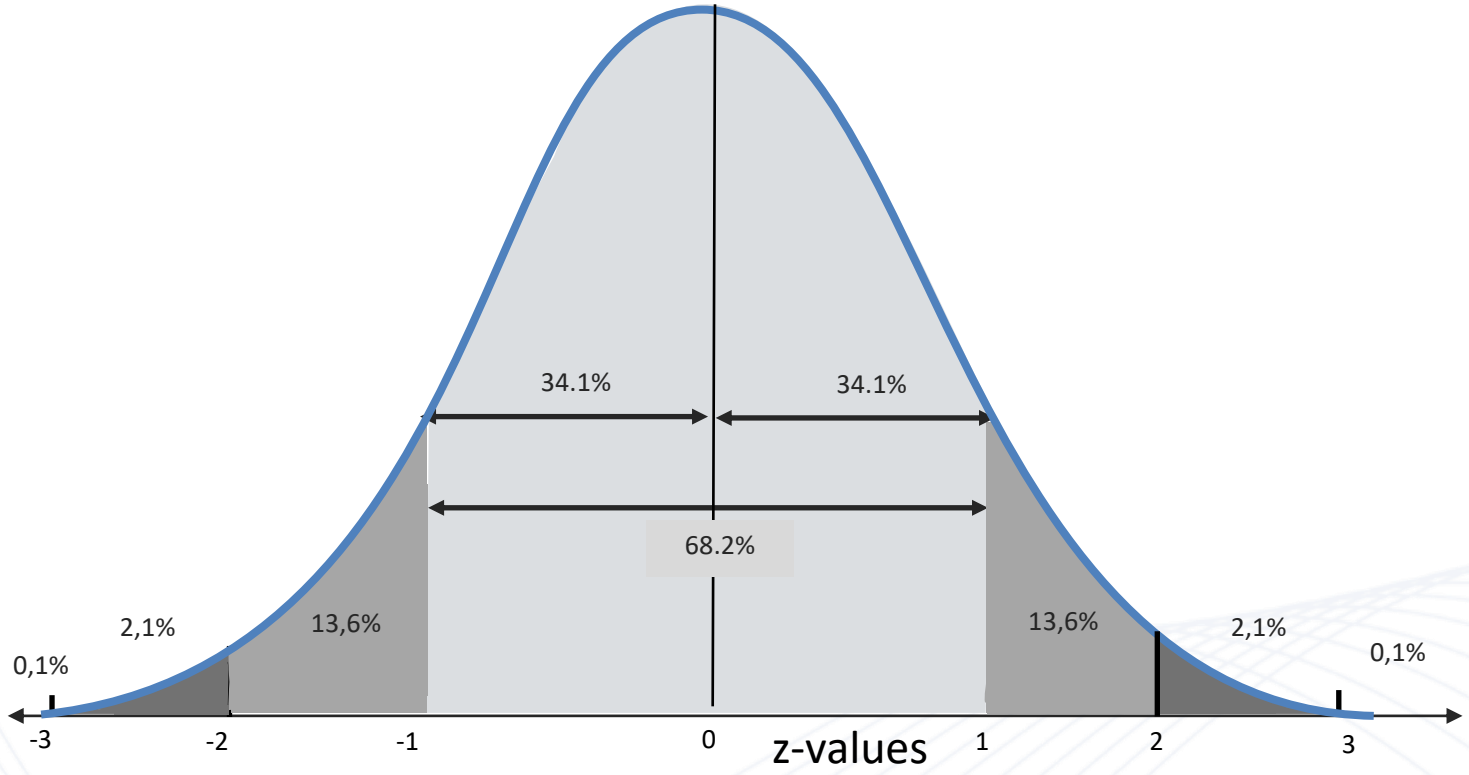


Example: Standardization and z-values (continued)

These are called **standardized** or **z-values**.

- The z-value of an observation is the number of standard deviations (SD) it is above or below the mean.
- z = (observation - mean) / standard deviation.
- z-values are defined for any distribution shape, but only for normal distributions it is possible to use them for the calculation of percentiles.
- Different normal distributions with different means and standard deviations can be compared using z-values.
- Observations that are more than 2 standard deviations away from the mean (|z| > 2) are usually considered unusual.

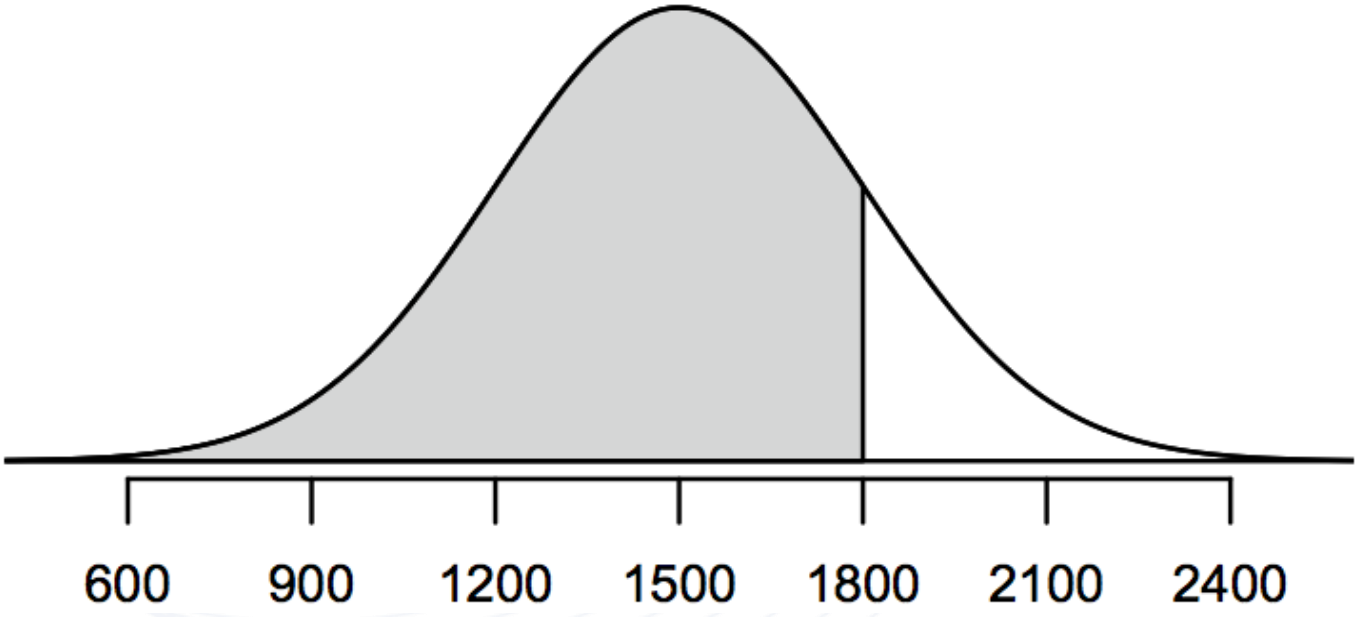
z-values and areas under the curve of the standard normal distribution



Here you can roughly see what proportion of the distribution lies within an interval of z-values or beyond a certain z-value.

Distribution function

- The **value of the distribution function** is the percentage of observations that are less than or equal to a given observed value.
- Graphically, the distribution function corresponds to the area under the probability density to the left of the observed value.
- Interpretation: The probability of achieving a score of 1800 or less corresponds to the grey area under the curve.



Exercise

Determine the requested probabilities both manually and with R.

Determine for the SAT (μ=1500, σ=300):

a) The probability of a test result of no more than 2100 points.
b) The probability of a test result of at most average.
c) The probability of a test result of at least 1800 points.

Determine for the ACT (μ=21, σ=5):

a) The probability of a test result of 31 points or less.
b) The probability of a test result of at most average.
c) The probability of a test result of at least 36 points.

Determining the values of the distribution function - with tables

z	0	1	2	3	4	5	6	7	8	9
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

$$\Phi(-z) = 1 - \Phi(z)$$

Quantiles of the standard normal distribution

z	Φ (z)
1,2816	0,900
1,6449	0,950
1,9600	0,975
2,3263	0,990
3,0902	0,999

Lübke/Vogt (2014): Angewandte Wirtschaftsstatistik: Daten und Zufall, Springer Gabler, p. 213

Determining the values of the distribution function - with R

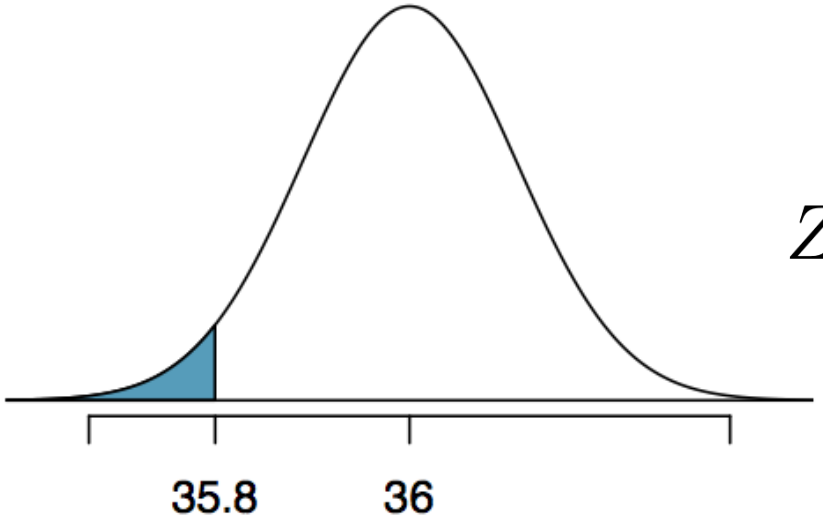
There are many ways in R to determine values of distribution functions/areas under the curve (and thus probabilities).

```
pnorm(1800, mean=1500, sd=300) # x

## [1] 0.8413447

pnorm((1800-1500)/300) # z

## [1] 0.8413447c
```

Quality control

For Heinz ketchup, the amount of ketchup filled into the bottle is normally distributed with a mean value of 36 oz. (1020.58 g) and a standard deviation of 0.11 oz. (3.12 g). Every 30 minutes, a bottle is removed from production and the contents measured. If the amount of ketchup in the bottle is less than 35.8 oz. (1014.91 g) or more than 36.2 oz. (1026.253 g), then the bottle has failed quality control. What percentage of bottles contain less than 35.8 oz. ketchup?

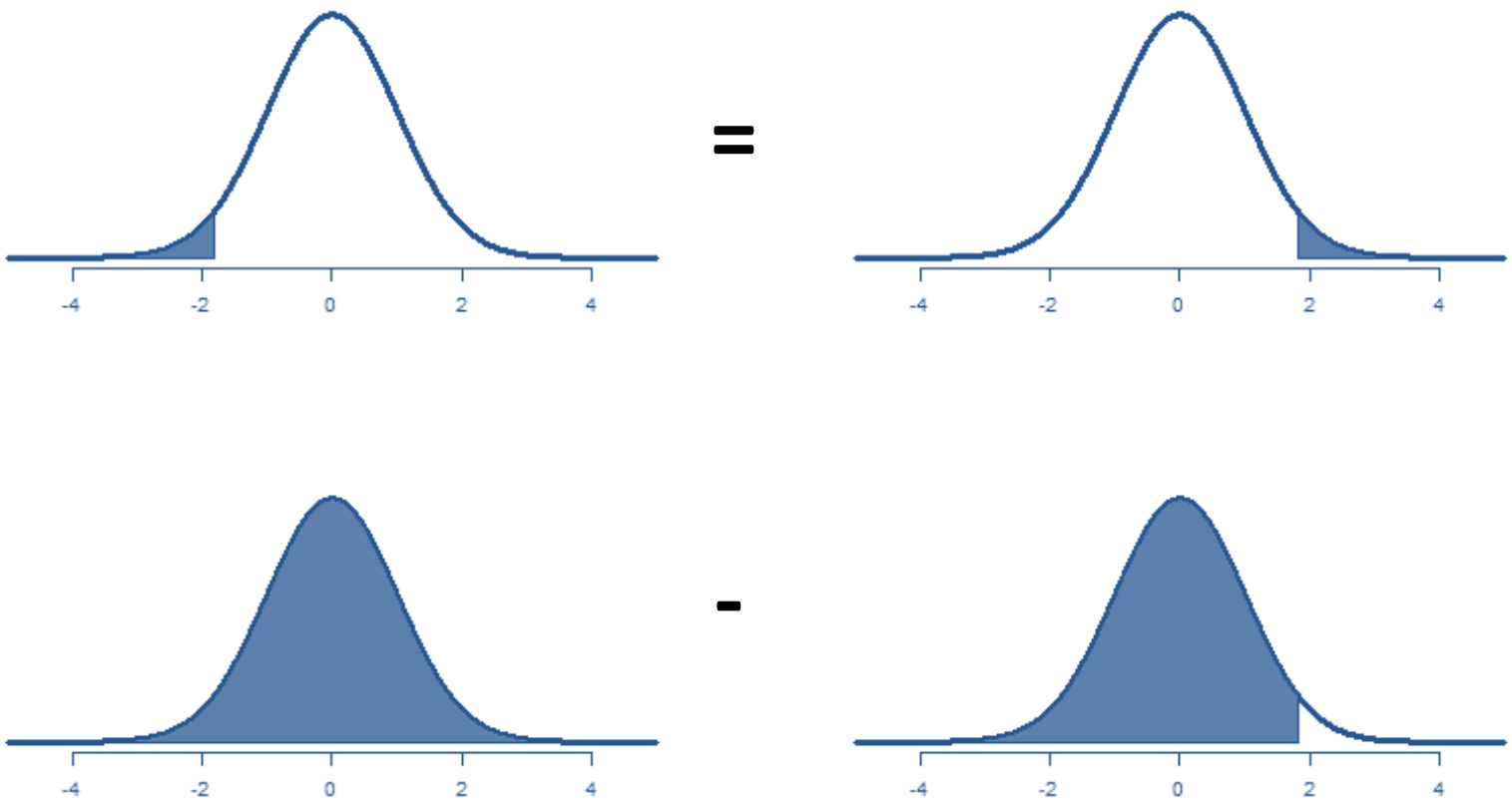
→ $Let\ X\ = Amount\ of\ ketchup\ in\ a\ bottle$: $X \sim N(\mu\ =\ 36, \sigma\ =\ 0{,}11)$



$$Z = \frac{35,8 - 36}{0,11} = -1,82$$

Application of the table for negative z-values

The normal distribution is symmetrical, i.e. $\Phi(-z) = 1 - \Phi(z)$



Determining the values of the distribution function - with tables

z	0	1	2	3	4	5	6	7	8	9
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

$$\Phi(-z) = 1 - \Phi(z) = 1 - \Phi(1,82) = 1 - 0,9656 = 0,0344$$

Exercise

What percentage of bottles successfully pass quality control?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

Exercise

Let the intelligence quotient IQ be normally distributed with $\mu = 100$ and $\sigma = 15$. Determine the following probabilities for the IQ of a random person. Calculate the probabilities both manually and using R:

a) $IQ \leq 115$

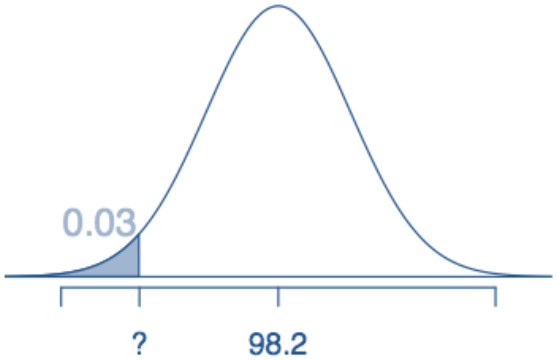
b) $IQ > 145$

c) $IQ \leq 70$

d) $70 < IQ \leq 130$

Finding the cut-off point – with tables

The body temperature of healthy people is approximately normally distributed with a mean of 98.2° F (36.77° C) and a standard deviation of 0.73° F (0.41° C). Where is the cut-off point for the lowest 3% of body temperature?



$$P(X < x) = 0.03\text{: } \Phi(z) = 0.03 \Leftrightarrow \Phi(-z) = 1 - 0.03 \Rightarrow -z = -1.88 \Leftrightarrow z = -1.88$$

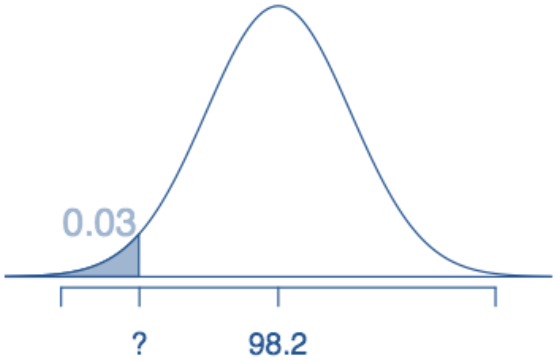
$$z = \frac{observed\ value\ - mean\ value}{standard\ deviation} \dashrightarrow z = \frac{x - 98.2}{0.73} \Leftrightarrow x = -1.88 \cdot 0.73 + 98.2 = 96.8$$

3% of the body temperature is below 96.8° F (36° C).

Mackowiak, Wasserman, and Levine (1992): A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick, JAMA The Journal of the American Medical Association, 268 (12).

Finding the cut-off point – with R

The body temperature of healthy people is approximately normally distributed with a mean of 98.2° F (36.77° C) and a standard deviation of 0.73° F (0.41° C). Where is the cut-off point for the lowest 3% of body temperature?




```
qnorm(0.03, mean=98.2, sd=0.73)

## [1] 96.82702
```

3% of the body temperature is below 96.8° F (36° C).

Mackowiak, Wasserman, and Levine (1992): A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick, JAMA The Journal of the American Medical Association, 268 (12).

Exercise

The body temperature of healthy people is approximately normally distributed with a mean of 98.2° F (36.77° C) and a standard deviation of 0.73° F (0.41° C). Where is the cut-off point for the highest 10% values of body temperature?

(a) 97.3° F

(b) 99.1° F

(c) 99.4° F

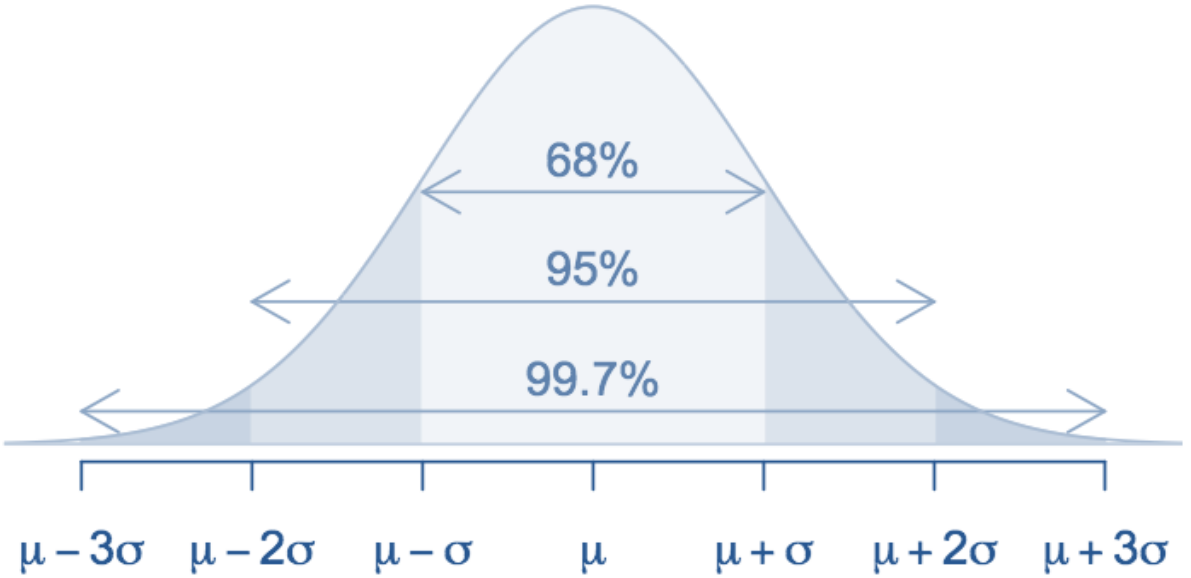
(d) 99.6° F

68-95-99.7 rule

For approximately normally distributed data….

- approx. 68% are within 1 SD around the mean value,
- approx. 95% are within 2 SD around the mean value,
- approx. 99,7% are within 3 SD around the mean value.

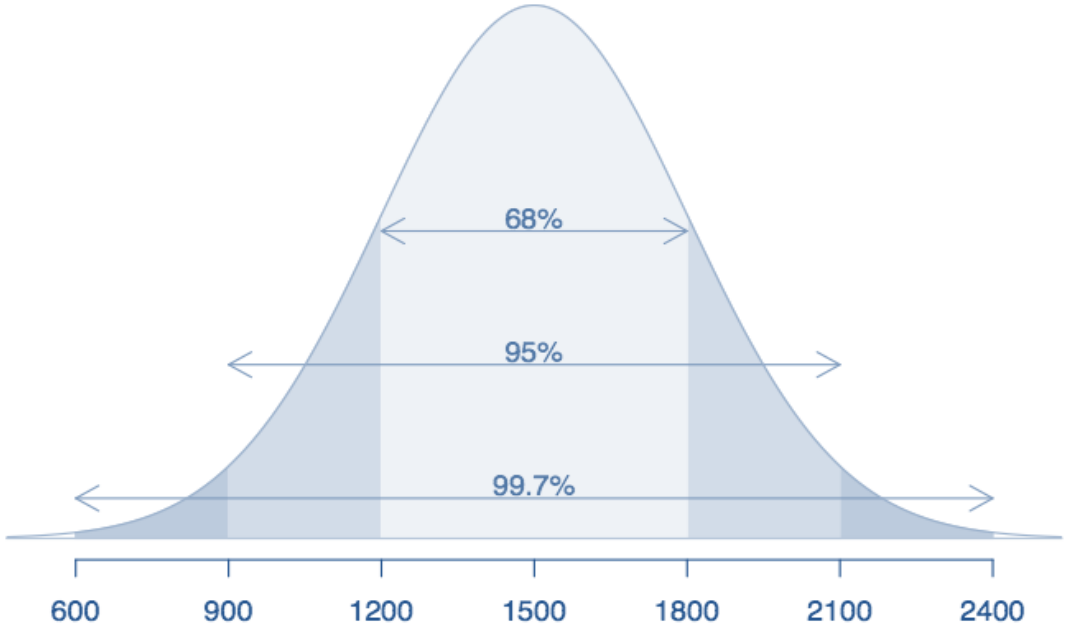
It is possible for observations to be 4, 5 or more standard deviations away from the mean, but this is very rare for normally distributed data.



Description of the variability with the 68-95-99.7 rule

SAT scores are approximately normally distributed with a mean of 1500 and a standard deviation of 300.

- ~68% of SAT test scores are between 1200 and 1800.
- ~95% of SAT test scores are between 900 and 2100.
- ~99.7% of SAT test scores are between 600 and 2400.



Exercise

Which of the following statements is wrong?

a) The majority of z-values of a right-skewed distribution are negative.

b) With a skewed distribution, the z-value of the mean can be unequal to 0.

c) For a normal distribution, the IQR is less than μ +/- σ.

d) z-values are useful in recognizing how unusual a single observation is compared to the rest of the data within the distribution.

Exercise

Use a suitable website/app to find out the weather forecast for tomorrow and discuss the question of whether you will need an umbrella tomorrow.

3.4 Important probability distributions

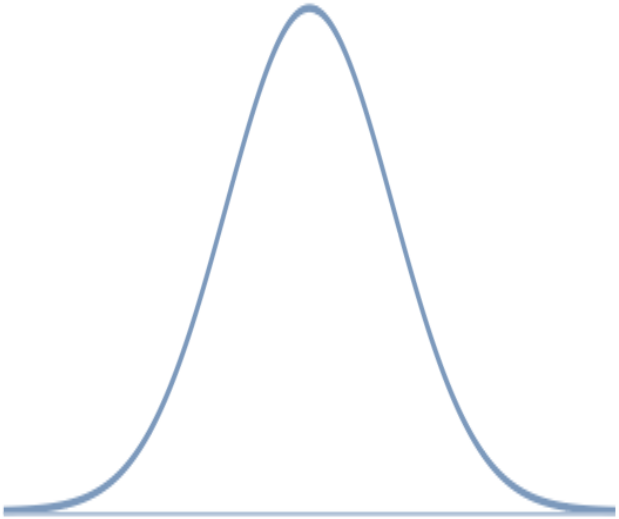
Overview

- The methods of inferential statistics are based on probability theory. If the probability distribution of a sample effect is known, then a statement can be made about the effect in the population.
- The most important probability distribution is the normal distribution.
- In addition, there are other important probability distributions that are used in many statistical hypothesis tests: The t-distribution, the chi-square distribution and the F-distribution.

Normal distribution and standard normal distribution

- The **normal distribution** is unimodal and symmetrical around μ, bell-shaped curve.
- Designation N(μ, σ) → Normal distribution with mean value (expected value) μ and standard deviation σ.
- The parameters μ and σ determine the exact shape of the normal distribution: μ shifts the bell curve on the x-axis, and σ stretches or compresses the curve.
- Each normal distribution can be transformed into the standard normal distribution by the z-transformation $z=(x-\mu)/\sigma$.
- The **standard normal distribution** has μ = 0 and σ = 1.

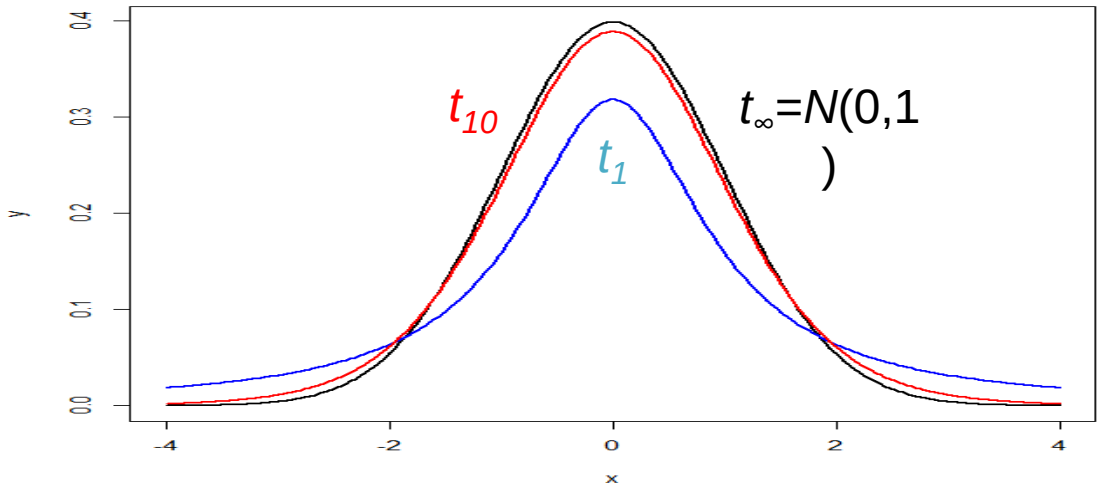
$$f(x|\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$$f(x|\mu;\sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The t distribution

- Unimodal and symmetrical, bell-shaped curve as the normal distribution.
- A normally distributed random variable with unknown population variance (and thus unknown standard error, see next section) is t-distributed with df degrees of freedom (df).
- Designation: t_{df}.



- The greater the number of degrees of freedom, the closer the t distribution is to the normal distribution.

Degrees of freedom

- The number of degrees of freedom (df) is the number of independent observations that are used to estimate a key figure, minus any estimated intermediate or auxiliary key figures.
- Example: Mean value for $x_1, \ldots, x_n$ independent observations:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

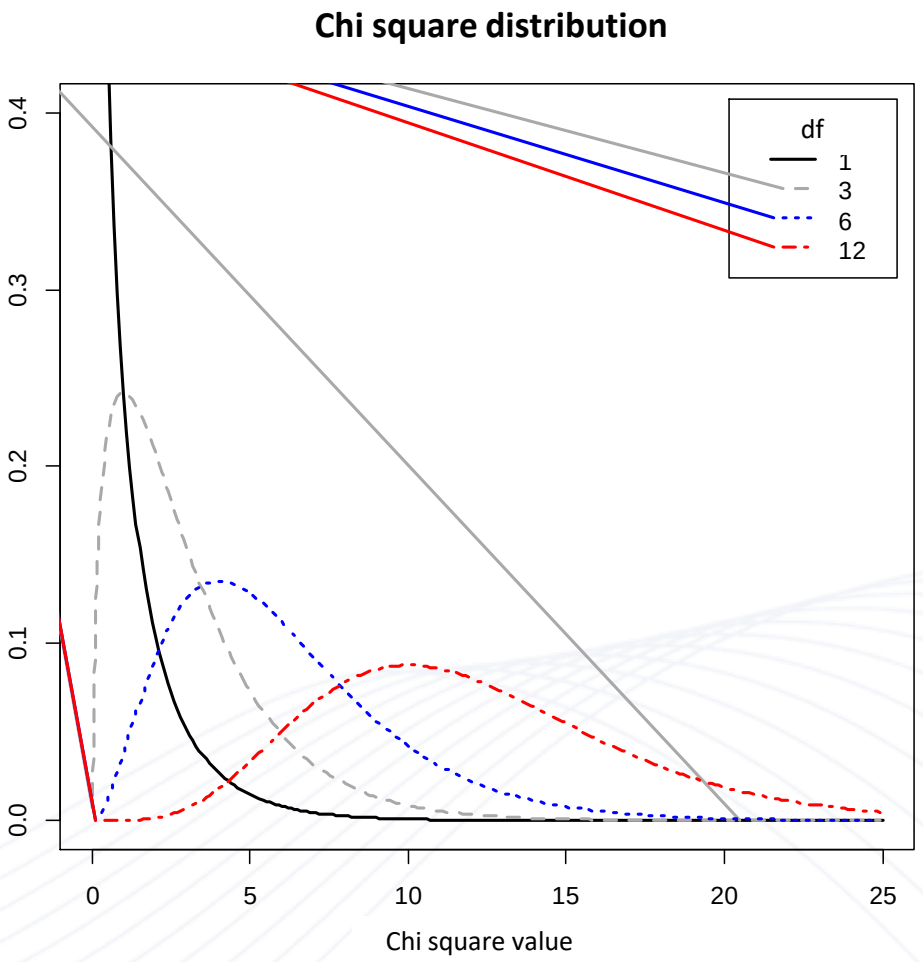
  - No further key figures are required to calculate the mean value; the number of degrees of freedom corresponds to the sample size: df = n.
- Example: Standard deviation for $x_1, \ldots, x_n$ independent observations:

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

  - To calculate the standard deviation, an additional key figure (mean value) is required, the number of degrees of freedom corresponds to the sample size less 1: df = n-1.

The chi-square distribution

- Mostly unimodal and right-skewed.
- For standard-normally distributed random variables $X_1,\ldots,X_n$, the sum $X_1^2 + \cdots + X_n^2$ follows a chi-square distribution with n degrees of freedom.
- This makes the chi-square distribution the typical probability distribution for variance estimates.
- Designation: χ_{df}.



The F-distribution

- Mostly unimodal and right-skewed.
- The quotient of two chi-squared distributed random variables (each divided by their degrees of freedom) is F-distributed with df1 and df2 degrees of freedom.
- This makes the F-distribution the typical probability distribution for quotients from variance estimates.
- Designation: $F_{df1,df2}$