Datenanalyse: Eye-tracking 2020/2021

Katja Suckow

2021-01-07

Contents

Chapter 1

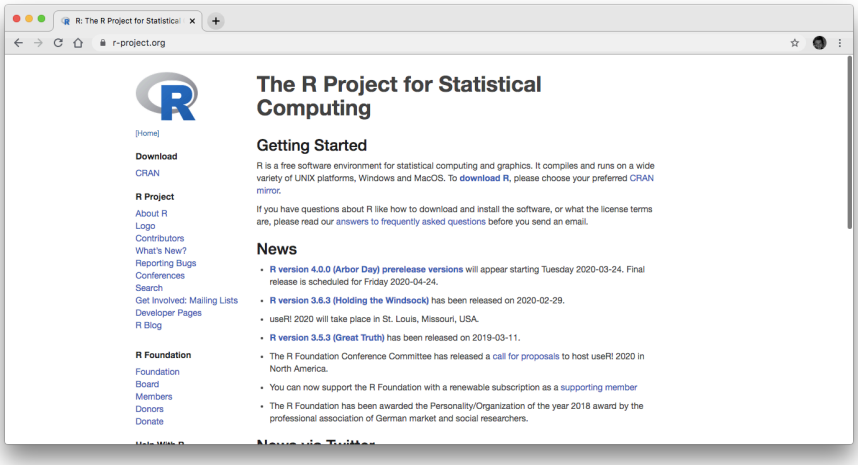
Willkommen

Da wir uns leider in diesem Jahr erstmal nicht persönlich treffen können, habe ich hier eine kleine Übersicht zu den verwendeten R-Befehlen erstellt.

1.1 Was braucht man dafür?

Dafür muss man auf seinem Rechner **R** und **Rstudio** installieren (beide sollten auf Windows, Mac und Linux laufen).

1. **R** frei verfügbare Software https://www.r-project.org/



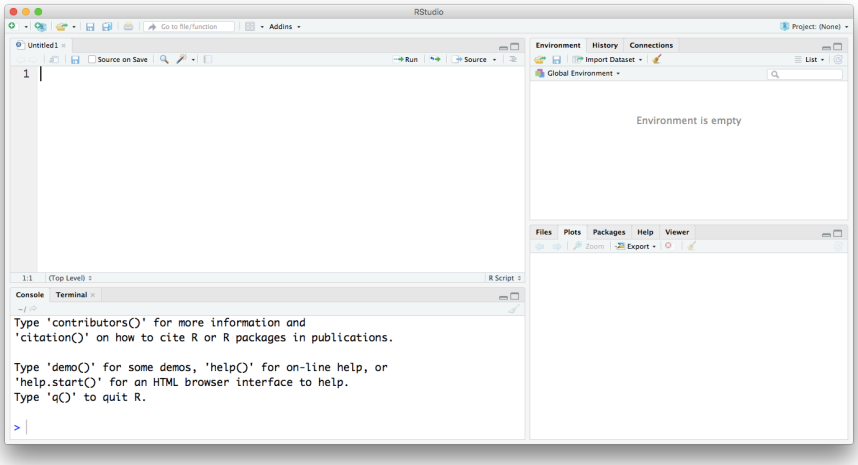
R ist eine freie Programmiersprache zur statistischen Datenanalyse und

Erstellung von Grafiken. Sie kann auch als Skriptsprache benutzt werden, um einfache Skripte und Programme zu schreiben. **R** ist vor allem in der Wissenschaft weit verbreitet und löst hier zunehmend SPSS ab. **R** bietet viele Methoden und Pakete zu statistischen Auswertung und Datendarstellung, die ständig in open-source weiterentwickelt werden, es steht kostenlos zur Verfügung und kann auf jeder Plattform (Windows, Mac, Linux) laufen. R besteht aus 3 Hauptfenstern: (1) **Konsole**, um direkt Befehle einzugeben, (2) **Editor**, um eine Abfolge an Befehlen zu speichern und auszuführen, (3) **Grafikfenster**

2. **RStudio** https://www.rstudio.com/products/rstudio/download Es gibt verschiedene Möglichkeiten mit R zu arbeiten. Wir werden die grafische Oberfläche, die **RStudio** bietet, nutzen. **RStudio** bietet eine gut handhabe Oberfläche, viel Unterstützung und viele integrierte Apps (Shiny, Markdown, Bookdown …), die auf R zurückgreifen. Unterschiedliche Editoren zur Bearbeitung von R Dateien sind:

- RStudio (für die gemeinsame Datenauswertung im Praktikum empfohlen)
- Notepad++ (Windows)
- Textwrangler
- . . .

Figure 1.1: RStudio Logo



So sieht RStudio ohne Inhalt oder Daten aus.

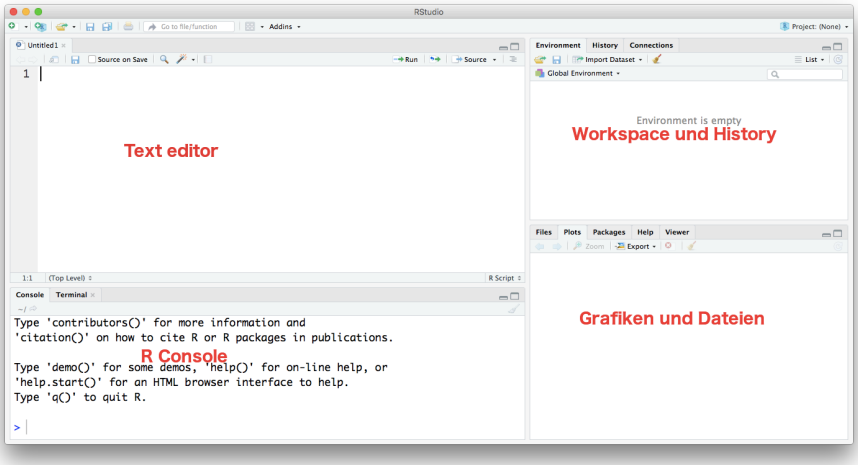


Figure 1.2: Die 4 Hauptfenster von RStudio

In der Standarddarstellung befindet sich oben links der **Code Editor**, oben rechts können sie **Workspace und History** ansehen, unten links befindet sich die **R Console**, in der Sie den Code direkt eingeben können und unten rechts lassen sich unter anderem **Plots und Files** anzeigen. Schauen Sie sich die verschiedenen Möglichkeiten an, die die verschiedenen Tabs in den Ecken bieten.

Skript

Das Skript enthält eine Sammlung von Befehlen. Diese können auch direkt über Konsole ausgeführt werden. Es empfiehlt sich jedoch sehr, diese im **Code**

Editor zu speichern, um später noch einmal darauf zugreifen zu können.

Workspace

Der Workspace enthält Sammlung von Objekten, die in einer Session erstellt wurden. Diese kann explizit gespeichert werden, um die einzelnen Objekte noch einmal neu zu laden.

So sieht Rstudio mit ein bischen mehr Inhalt aus.

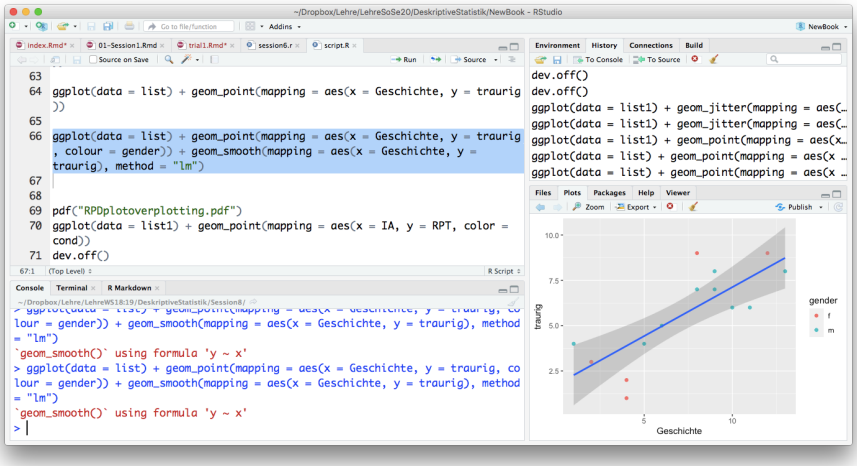


Figure 1.3: RStudio mit Datensatz und grafischer Darstellung eines Datensatzes

- **oben links** steht der Code, der gespeichert und auch aus dem Editor ausgeführt werden kann. (Mac: Markierung des auszuführenden Codes und dann **CMD + Enter**)
- **unten links** man kannn den auszuführenden Code auch direkt unten in das Konsolenfenster eingeben. Dann wird dieser allerdings nicht gespeichert. Es wird empfohlen, Code der erstmal nur ausprobiert wird, direkt in das Konsolenfenster einzutragen. Wenn der Code das korrekte Ergebnis liefert, sollte er im Editor in einem Skript gespeichert werden.
- **oben rechts** lassen sich History der eingegebenen Befehle und auch die Inhalte des Workspace anzeigen
- **unten rechts** kann man sich die geladenen Pakete für **R**, sowie die verschiedenen erstellten Grafiken anzeigen lassen

Die Anordnung der Fenster lassen sich über **RStudio** > **Preferences** und dann **Pane Layout** ändern. Über **Preferences** können Sie sich generell anzeigen lassen, welche Änderungen Sie u.a. bei der Anzeige vornehmen können bzw.~wollen.

3. **Shiny App**

Das ist eine App, die in RStudio integriert ist und die deskriptive Daten mit **R** bildlich und interaktiv veranschaulichen kann und auch veröffentlicht. Wenn man sich einmal ein bisschen mit R auseinandergesetzt hat, kann man sich mit Shiny weiter "austoben".

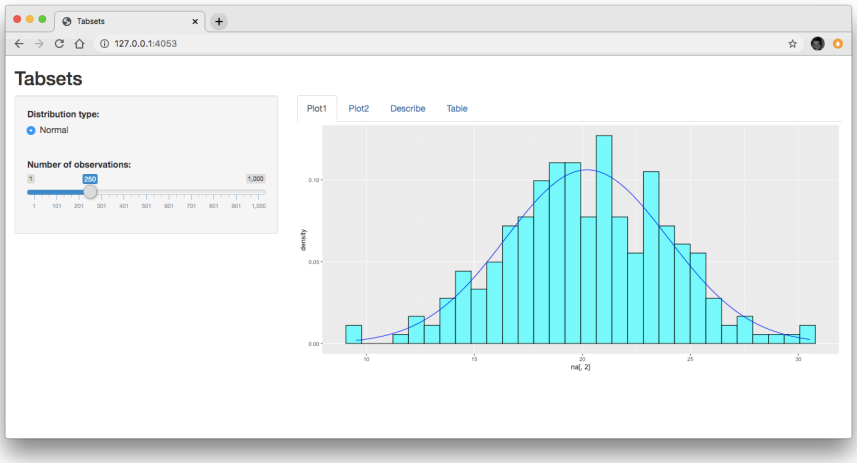


Figure 1.4: Beispiel des Screenshots einer Shiny App, die interaktiv die Normalverteilung einer übergebenen Anzahl beobachteter Werte anzeigt.

1.2 Einleitung: Datentypen in der Linguistik

Qualitative Daten Nicht-numerische, oft verschriftlichte oder in audiovisueller Form vorliegende Daten. Verwendung häufig explorativ und hypothesengenerierend. (e.g., Interviews)

Quantitative Daten Numerische (bzw. in numerischer Form überführte) Daten, die mit dem Ziel der Überprüfung von Hypothesen und Theorien erhoben werden

1.3 Population und Stichprobe

In der psycholinguistischen Forschung werden Daten in einer kontrollierten Umgebung (Experiment) erhoben, die Aussagen für die Gesamtpopulation zu einer bestimmten linguistischen Fragestellung machen zu können.

Dabei wird eine kleine Gruppe der Population (Stichprobe) in einem Experiment getestet. Bestimmte Daten dieser Stichprobe werden erhoben und anschliessend die Ergebnisse dieser Daten auf die Gesamtpopulation projiziert. Die erhobenen Daten der Stichprobe müssen erst in deskriptive statistische Kennwerte umgewandelt werden und anschliessend mittels statistischer Verfahren Hypothesen geprüft werden, um Aussagen über die Gesamtpopulation machen zu können.

1.4 Deskriptive Statistik und Inferenzstatistik

Deskriptive Statistik übersichtliche Darstellung der erhobenen Daten einer Stichprobe in Form statistischer Kennwerte (Mittelwerte, Streuung, Verteilung, Grafiken)

Inferenzstatistik Statistische Verfahren, die es ermöglichen Rückschlüsse aus den Daten auf die Gesamtpopulation zu ziehen, um Schätzungen für Populationswerte zu berechnen oder experimentelle Hypothesen zu testen.

1.5 R als Taschenrechner: Beispiele

Um sich erstmal mit **R** vertraut zu machen, kann man R erstmal als Taschenrechner einsetzen. Dabei ist die Anwendung recht assoziativ und kann wie folgt eingesetzt werden.

Probieren Sie diese Beispiele erstmal selbst zuerst in Ihrem Terminal und dann aus dem Editor heraus aus!

Addition +

```
5 + 7
```

```
## [1] 12
```

Subtraktion -

```
325 - 18
```

```
## [1] 307
```

Multiplikation *

```
43 * 21
```

```
## [1] 903
```

Division /

```
442 / 13
```

```
## [1] 34
```

Exponent ^

```
32^2
```

```
## [1] 1024
```

1.6 Einige vordefinierte Beispielfunktionen

Funktionen können auch sehr einfach in **R** selbst geschrieben werden. Dazu kommenw wir später auch noch. Man muss allerdings nicht alles neu erfinden, besonders weil **R** auch viele Funktionen vordefiniert mitbringt.

Die Syntax ist relativ simpel. Die Funktion hat einen eigenen Namen und die übergebenen Argumente, mit denen die Funktion Berechnungen durchführt werden in der Klammer nach dem Funktionsnamen übergeben.

Unten sind ein paar Beispiele für frequente leicht nachvollziehbare Funktionen aufgelistet.

Summe sum() - Bildung der Summe, der übergebenen Argumente:

```
sum(3,45,12,34)
```

```
## [1] 94
```

Wurzel sqrt() - Bildung der Wurzel von einem übergebenen Argument:

```
sqrt(1024)
```

```
## [1] 32
```

Minimum min() - Ermittlung der kleinsten Zahl aus einer Liste von übergebenen Werten:

```
min(43,24,11,23,76,14,56,99,12)
```

```
## [1] 11
```

Maximum max() - Ermittlung der höchsten Zahl aus einer Liste von übergebenen Werten:

```
max(43,24,11,23,76,14,56,99,12)
```

```
## [1] 99
```

Absoluter Wert (Betrag) abs() - Ermittlung des absoluten Betrags aus einem Wert:

```
abs(-25)
```

```
## [1] 25
```

Aneinanderreihung concatenate c() verbindet in Klammern stehende Zahlen zu einer Liste

```
c(2,6,7)
```

List ls()

- listet alle Elemente im aktuellen Workspace auf

```
ls()
```

Gib Pfad des aktuellen working directory zurück getwd()

- Gibt Verzeichnis des aktuellen Workspace aus

```
getwd()
```

Setze working directory auf folgenden Pfad setwd()

- Setzt den aktuellen Workspace in angegebenes Verzeichnis

```
setwd(<Pfad>)
```

Lies Datensatz ein read.table()

- liest einen Datensatz in aktuellen Workspace ein

```
read.table(<file>, header = T)
```

write.table()

- exportiert eine Variable aus aktuellem working directory in die angegebene Dateinamen

```
write.table(<Variablenname>,file = "")
```

1.7 Vergleichsoperatoren

Hier sind die Beispiele logischer Vergleichsoperatoren. Deren Interpretation gibt entweder **TRUE** oder **FALSE** zurück.

gleich ==

- Vergleich ob die Werte beider Ausdrücke identisch sind
- ja: **TRUE**
- nein: **FALSE**

```
4 == 6
```

```
## [1] FALSE
```

ungleich: **!=**

- Vergleich ob die Werte beider Ausdrücke ungleich sind
- ja sie sind ungleich: **TRUE**
- nein sie sind gleich: **FALSE**

```
4 != 6
```

```
## [1] TRUE
```

größer: >

- ist der Wert links größer als der Wert rechts?

```
4 > 6
```

```
## [1] FALSE
```

kleiner: $<$

- ist der Wert links kleiner als der Wert rechts?

```
4 < 6
```

```
## [1] TRUE
```

größer gleich: $>=$

- ist der Wert links größer oder gleich gleich der Wert rechts?

```
4 >= 6
```

```
## [1] FALSE
```

```
4 >= 4
```

```
## [1] TRUE
```

kleiner gleich: $<=$

- ist der Wert links kleiner als der Wert rechts?

```
4 <= 6
```

```
## [1] TRUE
```

```
4 <= 4
```

```
## [1] TRUE
```

1.8 Logische Vergleichsoperatoren

Konjunktion, logisches "Und": **&&** oder **&**

- beim logischen **Und** werden die Werte der beiden Ausdrücke verglichen - nur wenn beide **TRUE** ergeben, wird der Gesamtausdruck **TRUE** ergeben
- im unteren Beispiele sind beide Ausdrück **TRUE**

```
4 <= 4 & 4 >= 4
```

```
## [1] TRUE
```

Disjunktion, logisches "oder": || oder |

- beim logischen **Oder** reicht es, wenn ein zu vergleichender Ausdruck **TRUE** ist, um den gesamten Ausdruck **TRUE** zu machen.

```
4 < 6 || 5 > 8
```

```
## [1] TRUE
```

Negation: **!**

- bei der Negation wird der Wahrheitsgehalt eines Ausdruckes umgekehrt.

```
!(4 < 6 || 5 > 8)
```

```
## [1] FALSE
```

1.9 Variablen in R

Verschiedene Werte oder Ergebnisse einer Berechnung können leicht in Variablen (Platzhalter) gespeichert werden. Diese Variablen müssen in **R** mit einem Buchstaben beginnen. Die Zuweisung von Werten in eine Variable erfolgt entweder mit einem Pfeil **<-** oder mit einem simplen = (beides ohne Unterschied möglich):

```
a <- 15
b = 23
ab = c(a,b)
ab
```

```
## [1] 15 23
```

1.10 Hilfefunktionen

Mit der Hilfefunktion **?** lässt sich zu jeder R-Funktion eine hilfreiche Beschreibung anzeigen. Schreiben Sie einfach das **?** direkt vor die Funktion zu der Sie mehr erfahren wollen und **R** öffnet ein Hilfsfenster. Probieren Sie es einmal aus und schreiben:

?sum

Mit **?sum** können Sie sich die Informationen zum **sum** Befehl anzeigen lassen. Meist werden am Ende der Beschreibung auch Beispiele zur Benutzung angezeigt.

Alternativ können Sie auch den **help()** benutzen:

```
?sum
help(sum)
```

1.11 Pakete Laden

Pakete sind eine Sammlung verschiedener Funktionen zu einem Thema.

```
# installiert das Paket
install.packages("tidyverse")

# lädt das Paket
library(tidyverse)
```

Chapter 2

Datenbeschreibungen in R

Kurze Zusammenfassung einiger zentraler Funktionen für das Praktikum.

```
# Daten einlesen
getwd()
# Der Pfad zu Eurem Workspace ist auf jedem Rechner anders
# Deshalb läßt sich das nicht vorgeben und Ihr könnt diesen nicht einfach
# von jemand kopieren.
setwd("<Setzt bitte hier den Pfad zum Workspace auf Eurem Rechner ein>")
# Der Datensatz sollte 10VP.dat sollte im Pfad Eures Workspace gespeichert sein
list = read.table("gugVP11.txt",header = T)
```

```
# Inhalt des Workspace
# welche Variablen wurden eingelesen oder selbst erstellt
ls()

# listet die ersten 6 Zeilen auf
head(list)

# Berechne die Summe aller Lesezeiten
sum(list$time)
```

2.1 R-Vektoren

Vektoren sind eine Datenstruktur, die Elemente desselben Datentyps in Form einer Liste enhält. Diese Datentypen können *logisch*, *integer*, *double*, *character* oder *komplex* sein.

```
x = c(1:30)
#typeof() gibt den Datentyp einer Variable wieder
typeof(x)
```

```
## [1] "integer"
```

```
#str() listet den Datentyp und die verschiedenen Ausprägungen einer Variable auf
str(x)
```

```
##  int [1:30] 1 2 3 4 5 6 7 8 9 10 ...
```

```
#length() gibt die Länge einer Liste wieder
length(x)
```

```
## [1] 30
```

2.2 R-Datentypen

R weist die Datentypen automatisch zu - Variablentyp muss nicht explizit festgelegt werden. Manchmal kann man den Variablentyp ändern, wenn bestimmte Funktionen durchgeführt werden sollen.

integer sind ganzzahlige Werte im endlichen Bereich \ - standardisierter Bereich ist normalerweise von -32768 bis +32768 - wenn der Wert größer als 32768 ist, dann gibt es eine Art Überlauf

```
x = c(1:30)
typeof(x)
```

```
## [1] "integer"
```

```
str(x)
```

```
##  int [1:30] 1 2 3 4 5 6 7 8 9 10 ...
```

double bezeichnet Fliesskommazahlen, wird in R auch manchmal als **numeric** wiedergegeben.

```
x = 3.232
str(x)
```

```
##  num 3.23
```

```
typeof(x)
```

```
## [1] "double"
```

Character enthält Buchstabenfolgen. Diese müssen in ” ” stehen, um sie von den Variablen zu unterscheiden.

```
x = "Langeweile"
x
```

```
## [1] "Langeweile"
```

```
str(x)
```

```
##  chr "Langeweile"
```

```
typeof(x)
```

```
## [1] "character"
```

Factor sind Variablen, die bestimmte endliche Werte annehmen können. Sie bezeichnen kategorische Werte. **Factor** können als Zahlen oder Buchstaben dargestellt werden.

```
x = as.factor(c("apfel","birne","erdbeere","erdbeere"))
y = as.factor(c(1,2,2,1,1,1,8))
str(x)
```

```
##  Factor w/ 3 levels "apfel","birne",..: 1 2 3 3
```

```
str(y)
```

```
##  Factor w/ 3 levels "1","2","8": 1 2 2 1 1 1 3
```

logical Evaluation einer logischen Frage: Darstellung als **TRUE** oder **FALSE**; kann nur zwei Werte annehmen.

- Siehe hierzu die Inhalte von letzter Woche und die Folien und Übungen im

complex Daten, die nicht in traditioneller Weise dargestellt werden.

Die **str()** Funktion stellt die Datentypen der Werte in einer Tabelle dar:

```
# listet Datentypen aller Spalten auf
list = read.table("gugVP11.txt",header = T)
str(list)
```

str() listet die Tabelle nach den Spalten auf:

1. zuerst erscheint der Spaltennahme
2. dann der Datentyp: **int** = integer, ganzzahliger Datentyp in einem bestimmten Bereich **Factor**, Daten können nur einen bestimmten Wert annehmen
3. danach erscheinen die verschiedenen Beobachtungen der jeweiligen Spalte

2.3 R-Data frames

Data Frames ist die Datenstruktur in Form einer Tabelle (zweidimensional) mit **Zeilen** und **Spalten**. Deren Werte kann eine Mischung aus Zahlen und Buchstaben enthalten.

```
# ein Data Frame hat folgende Form
# diese Information wird später noch einmal wichtig
name[<zeile>, <spalte>]
```

```
# so kann man eine Liste generieren
# c() steht für "concatenate"", verbinden oder aneinanderreihen
y = c("Unsinn","Quatsch","rot")
y
```

```
## [1] "Unsinn"  "Quatsch" "rot"
```

```
#so kann man einen Data Frame generieren
a = data.frame(5:7,y,8:10, check.names = FALSE)
a
```

```
##   5:7       y 8:10
## 1   5  Unsinn    8
## 2   6 Quatsch    9
## 3   7     rot   10
```

So sieht der Inhalt des erstellten Data Frame schlußendlich aus; wie eine Tabelle ohne Überschrift und ohne Spalten und Zeilennamen

2.4 R-Matritzen

Matritzen in **R** haben ähnlich wie Dataframes die Form einer Tabelle, nur bestehen Matritzen nur aus Zahlen und enthalten keine Buchstaben.

```
# so kann man eine Matrix von 1 bis 30 aufsteigend erstellen
#und diese auf 10 Spalten und 3 Zeilen verteilen
a = matrix(data = 1:30, ncol = 10, nrow = 3, byrow = TRUE)

# so sieht der Inhalt dann aus
a
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    2    3    4    5    6    7    8    9    10
## [2,]   11   12   13   14   15   16   17   18   19    20
## [3,]   21   22   23   24   25   26   27   28   29    30
```

Mit der Funktion **dim()** lässt sich die Größe eines Dataframes oder einer Matritze anzeigen. Nehmen wir mal die selbsterstellte Matrix mit dem Namen **a**: **dim(a)** gibt die Anzahl der Zeilen und die Anzahl der Spalten der jeweiligen Tabelle wieder.

```
dim(a)
```

```
## [1]  3 10
```

Hier zeigt sich die schon angedeutete Anordnung in den eckigen Klammern: zuerst die links die **Zeile** dann rechts die **Spalte**.

```
# folgender Code zeigt, wie man auf einzelne Spalten zugreifen kann
# so kann man sich die 1. Spalte der Tabelle anzeigen lassen
a[,1]
```

```
## [1]  1 11 21
```

```
# hier zeigt sich, wie man sich die 1. Zeile der Tabelle anzeigen lassen kann
a[1,]
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Die **mean()** Funktion gibt den Mittelwert von einer Reihe von Zahlen wieder

```
# hier wird der Mittelwert der 1. Zeile der Tabelle a ermittelt und ausgegeben
mean(a[1,])
```

```
## [1] 5.5
```

Man kann in der Tabelle den einzelnen Spalten und Zeilen Namen geben. Falls dies im originalen eingelesenen Datensatz nicht schon geschehen ist, gibt es hierfür die Funktionen **colnames()** und **rownames()**

```
# die einzelnen Spalten werden alphabetisch angeordnet
colnames(a) = c("a","b","c","d","e","f","g","h","i","j")

# die einzelnen Zeilen bekommen so einen eigenen Farbnamen
# wenn die Zuweisungen Buchstaben enthalten, müssen diese in Hochkommata geschrieben w
# bei der Zuweisung von Zahlen, müssen diese nicht in Hochkommata stehen
rownames(a) = c("rot","blau","weiss")

# gib den Inhalt der Tabelle a aus
a
```

```
##        a  b  c  d  e  f  g  h  i  j
## rot    1  2  3  4  5  6  7  8  9 10
## blau  11 12 13 14 15 16 17 18 19 20
## weiss 21 22 23 24 25 26 27 28 29 30
```

2.5 R - nützliche Funktionen

as.factor() Mit dieser Funktion lässt sich ein Vector als Faktor kodieren. Ein als Faktor kodierter Vektor kann nur Werte aus bestimmten vordefiniertem Bereich enthalten.

Kopieren Sie jeweils den Code und lassen Sie sich die Ergebnisse anzeigen.

```
x = c(1:30)
str(x)
```

```
##  int [1:30] 1 2 3 4 5 6 7 8 9 10 ...
```

```
x = as.factor(x)
str(x)
```

```
##  Factor w/ 30 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

levels() Zeigt die Levels eines Factors an.

```
x = c(1:30)
x = as.factor(x)
levels(x)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
## [16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
```

table() erstellt eine Übersicht über die Werte einer Spalte.

```
x = c(1:30)
x = as.factor(x)
table(x)
```

```
## x
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## 27 28 29 30
##  1  1  1  1
```

```
# Hilfefunktionen in R
help("paketname")
?Paketname
```

```
install.packages("dplyr")
library(dplyr)

# zeige nur Zeilen, die bestimmte Bedingung erfüllen
filter()
filter(list,time < 150)

# ordne Zeilen nach einer bestimmten Spalte
arrange()
arrange(list,time)
```

```
# zusammenfassen der Daten
# min, max, mean, median, var, sd - anwendbar
summarise()
summarise(list,avg = mean(time))
```

```
# wähle Spalten nach Namen aus
select()
select(list, VP, time)

# mache Berechnungen und hänge Spalten an
mutate()
mutate(list, time2 = time/100)
```

Eine **Funktion** ist die Aneinanderreihung auszuführender Befehle. Man kann sie selbst definieren, schnell wiederholen und schnell anwenden.

```
<Name> = function(<Liste von übergebenen Argumenten>){
  <Befehle, die ausgeführt werden sollen>
}
```

Beispiel einer Funktion. Probiert diese doch einmal aus!

```
# "percent" berechnet Prozent a von b in Variable d
percent = function(a,b){
    d = 100/b * a
    d
}
percent(34,78)
```

```
## [1] 43.58974
```