

Prof. Dr. Hendrik Godbersen

Kurs

Angewandte Statistik & R

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

RStudio

Project: (None)

Environment History Connections

Global Environment

Data

- frame_Gesc... 2 obs. of 2 variables
- radio 275 obs. of 26 variables

Values

table_Gesc... 'table' num [1:2(1d)] 46.9 5...

```

41 radio$SenderCluster2[radio$SenderCluster=="Privat jung"] <- "Privat"
42 radio$SenderCluster2[radio$SenderCluster=="Privat alt"] <- "Privat"
43 radio$SenderCluster2[radio$SenderCluster=="SWR"] <- "SWR"
44
45 radio$CoGe <- (radio$Comedy + radio$Gewinnspiele) / 2
46
47 radio_weiblich <- radio[radio$Geschlecht == "weiblich",]
48 View(radio_weiblich)
49
50
51 # Häufigkeiten, Lageparameter & Streuungsmasse
52 # Häufigkeiten
53 tally(~Geschlecht, data = radio)
54 tally(~Geschlecht, format = "proportion", data = radio)
55 tally(~Geschlecht, format = "percent", data = radio)
56
57 tally(~SenderCluster, format = "percent", data = radio)
58
59 tally(SenderCluster~Geschlecht, format = "percent", data = radio)
60
61 # Lageparameter & Streuungsmasse
157:31 (Top Level)
  
```

Console

```

>
>
> # Setup
> require(mosaic)
> require(psy)
> require(psych)
> require(openxlsx)
> load("~/Desktop/Rworkshop/radio.RData")
> View(radio)
> # Häufigkeiten, Lageparameter & Streuungsmasse
> # Häufigkeiten
> tally(~Geschlecht, data = radio)
Geschlecht
maennlich weiblich
129      146
> # Graphiken - Häufigkeiten
  
```

Files Plots Packages Help Viewer

Zoom Export

Geschlecht	Freq
maennlich	46.91
weiblich	53.09

Einleitung

- **Dieser Kurs vermittelt das Wissen & die Fähigkeiten,...**
 - ...die gängigen Verfahren der deskriptiven & inferenzstatistischen Auswertung zu verstehen und anzuwenden
 - ...mit R zu arbeiten & die für statistische Auswertungen notwendigen Befehle anzuwenden
 - ...ein empirisches Projekt mit den „gängigen“ statistischen Verfahren in R auszuwerten
- **Sie erhalten auf www.godbersen.online...**
 - ...das Buch zu diesem Kurs zum strukturierten Lernen & Üben
 - das Befehlssammlung mit allen notwendigen R-Befehlen, um ein empirisches Projekt mit den „gängigen“ Verfahren auszuwerten
 - ...einen „echten“ Datensatz, mit dem die Übungsaufgaben zu diesem Kurs realisiert werden

Inhalt

- 1 Einleitung
- 2 Installation & Grundstruktur von R Studio
- 3 R-Befehle & Übungsdaten
- 4 Datenmanagement
- 5 Deskriptive Statistik**
- 6 Inferenzstatistische Verfahren**
- 7 Graphiken & Export von Ergebnissen nach Excel

Einleitung

- **Kern des Kurses sind die statistischen Verfahren (Kap. 5 & 6):**

- 1) Erklärung der statistischen Verfahren
- 2) Übungsaufgabe mit einem „echten Datensatz“
- 3) Umsetzung der statischen Analyse in R

- **Nutzung des Kurses für den größten Lernerfolg:**

- Nutzung von Kurs, Buch & Videos:
 - Am Anfang: Kapitel-für-Kapitel & Schritt-für-Schritt
 - Später: Nachschlagewerk
- Herangehensweise:
 - Wann wende ich welches Verfahren an & wie interpretieren ich die Ergebnisse? (& nicht: Wie programmiere ich mit R?)
- Lernziel:
 - Expertise in anwendungs- & umsetzungsorientierter empirischer Forschung (& nicht in „abgedrehtem“ Statistikwissen)

Inhalt

- 1 Einleitung
- 2 Installation & Grundstruktur von R Studio
- 3 R-Befehle & Übungsdaten
- 4 Datenmanagement
- 5 Deskriptive Statistik**
- 6 Inferenzstatistische Verfahren**
- 7 Graphiken & Export von Ergebnissen nach Excel

RStudio

Project: (None)

Environment History Connections

Global Environment

Data

- frame_Gesc... 2 obs. of 2 variables
- radio 275 obs. of 26 variables

Values

table_Gesc... 'table' num [1:2(1d)] 46.9 5...

```

41 radio$SenderCluster2[radio$SenderCluster=="Privat jung"] <- "Privat"
42 radio$SenderCluster2[radio$SenderCluster=="Privat alt"] <- "Privat"
43 radio$SenderCluster2[radio$SenderCluster=="SWR"] <- "SWR"
44
45 radio$CoGe <- (radio$Comedy + radio$Gewinnspiele) / 2
46
47 radio_weiblich <- radio[radio$Geschlecht == "weiblich",]
48 View(radio_weiblich)
49
50
51 # Häufigkeiten, Lageparameter & Streuungsmasse
52 # Häufigkeiten
53 tally(~Geschlecht, data = radio)
54 tally(~Geschlecht, format = "proportion", data = radio)
55 tally(~Geschlecht, format = "percent", data = radio)
56
57 tally(~SenderCluster, format = "percent", data = radio)
58
59 tally(SenderCluster~Geschlecht, format = "percent", data = radio)
60
61 # Lageparameter & Streuungsmasse
157:31 (Top Level)
  
```

Console

```

>
>
> # Setup
> require(mosaic)
> require(psy)
> require(psych)
> require(openxlsx)
> load("~/Desktop/Rworkshop/radio.RData")
> View(radio)
> # Häufigkeiten, Lageparameter & Streuungsmasse
> # Häufigkeiten
> tally(~Geschlecht, data = radio)
Geschlecht
maennlich weiblich
129 146
> # Graphiken - Häufigkeiten
  
```

Files Plots Packages Help Viewer

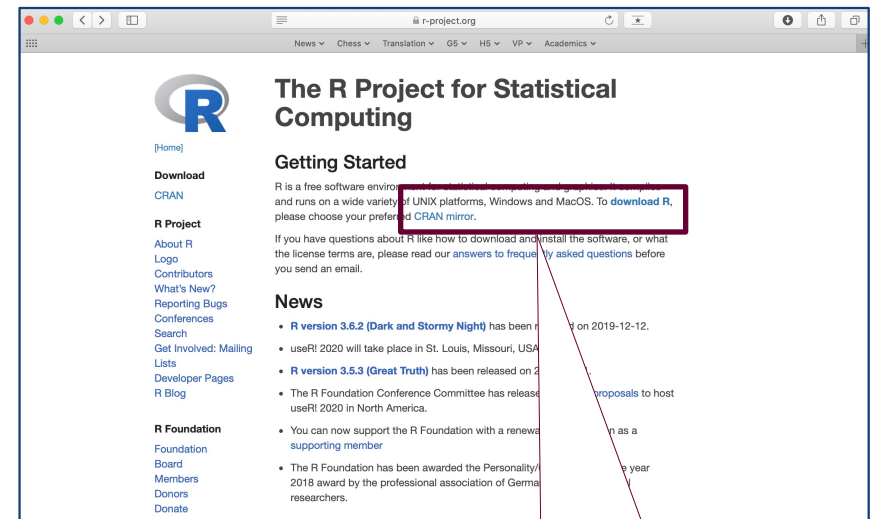
Zoom Export

Geschlecht	Freq
maennlich	46.91
weiblich	53.09

1 Einleitung	6 Inferenzstatistische Verfahren
2 Installation & Grundstruktur von R Studio	6.1 Überblick & Systematisierung der Verfahren
2.1 Installation von R, R Studio & relevanten Paketen	6.2 Chi-Quadrat-Test
2.2 Grundstruktur von R Studio	6.3 t-Test
3 R-Befehle & Übungsdaten	6.4 ANOVA
3.1 R-Befehle	6.5 Shapiro-Wilk-Test
3.2 Übungsdaten	6.6 Wilcoxon-Test
4 Datenmanagement	6.7 Korrelation
4.1 Datenvorbereitung & Laden von Daten	6.8 Regression
4.2 Daten bearbeiten in R	6.9 Hauptkomponentenanalyse
5 Deskriptive Statistik	7 Graphiken & Export von Ergebnissen nach Excel
5.1 Skalenniveaus	7.1 Graphiken in R
5.2 Häufigkeiten	7.2 Export von Ergebnissen nach Excel
5.3 Lageparameter & Streuungsmaße	

1) Download & Installation von R (R ist das Basisprogramm)

- 1) Gehen sie zu: <https://www.r-project.org/>
- 2) Wählen Sie einen Mirror zum Download aus
 - Achten Sie darauf, dass Sie R passend zu Ihrem Betriebssystem wählen
- 3) Laden Sie R herunter
- 4) Installieren Sie R auf Ihrem Computer



Zum Download von R auf
www.r-project.org

2) Download & Installation von R Studio (R Studio ist eine Entwicklungsumgebung für R & Erleichtert die Bedienung)

- 1) Gehen Sie zu:
<https://rstudio.com/products/rstudio/download/>
- 2) Laden Sie R Studio herunter
- 3) Installieren Sie R auf Ihrem Computer

3) Installieren Sie die Zusatzpakete in R Studio

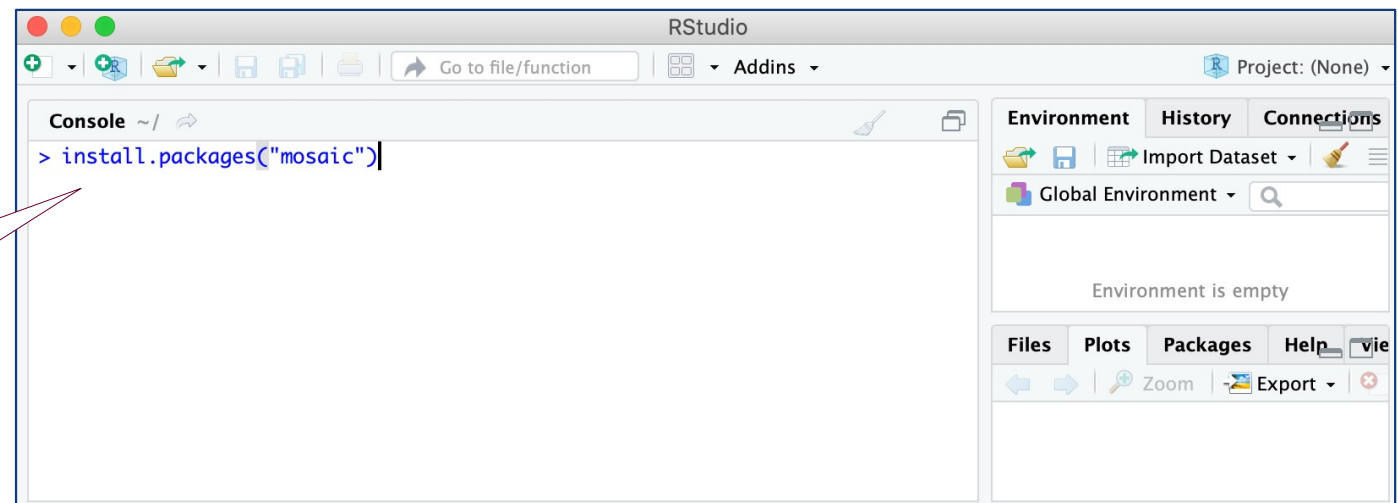
Installation von R, R Studio & relevanten Zusatzpaketen

- 1) Download & Installation von R ✓
- 2) Download & Installation von R Studio ✓
- 3) **Installieren Sie die Zusatzpakete in R Studio**
 - 1) Stellen Sie sicher, dass Ihr Computer mit dem Internet verbunden ist
 - 2) Geben Sie in die Console folgenden Befehl ein: `install.packages("mosaic")`
 - 3) Betätigen Sie die Return-Taste
 - 4) Wiederholen Sie dieses Vorgehen für alle relevanten Pakete

Benötigte R-Pakete:

- mosaic
- psych
- psy
- ggplot2
- openxlsx
- readxl

Befehl `install.packages("mosaic")`
eingeben & Return-Taste betätigen



1 Einleitung	6 Inferenzstatistische Verfahren
2 Installation & Grundstruktur von R Studio	6.1 Überblick & Systematisierung der Verfahren
2.1 Installation von R, R Studio & relevanten Paketen	6.2 Chi-Quadrat-Test
2.2 Grundstruktur von R Studio	6.3 t-Test
3 R-Befehle & Übungsdaten	6.4 ANOVA
3.1 R-Befehle	6.5 Shapiro-Wilk-Test
3.2 Übungsdaten	6.6 Wilcoxon-Test
4 Datenmanagement	6.7 Korrelation
4.1 Datenvorbereitung & Laden von Daten	6.8 Regression
4.2 Daten bearbeiten in R	6.9 Hauptkomponentenanalyse
5 Deskriptive Statistik	7 Graphiken & Export von Ergebnissen nach Excel
5.1 Skalenniveaus	7.1 Graphiken in R
5.2 Häufigkeiten	7.2 Export von Ergebnissen nach Excel
5.3 Lageparameter & Streuungsmaße	

Grundstruktur von R Studio

The screenshot shows the R Studio interface with several callouts explaining its components:

- Script mit R-Befehlen:** Points to the script editor on the left, which contains R code for loading data and performing calculations.
- Öffnen, Speichern etc. von Dateien erfolgt wie bei anderen Programmen über das Menü:** Points to the top menu bar.
- Button, um Befehle auszuführen (Zeilen im Skript müssen markiert sein):** Points to the 'Run' button in the top toolbar.
- Geladener Datensatz & generierte Objekte:** Points to the 'Environment' pane on the right, which lists loaded data objects like 'frame_Gesc...' and 'radio'.
- Anzeige des Datensatzes (Achtung: Anzeige ≠ geladener Datensatz):** Points to the 'Data' section in the 'Environment' pane.
- Generierte Graphiken:** Points to the 'Plots' pane on the right, which displays a bar chart showing the frequency of 'Geschlecht' (maennlich and weiblich).
- Console mit Ergebnisse (Befehlseingabe auch hier möglich; Befehle werden aber nicht gespeichert):** Points to the 'Console' pane at the bottom, which shows the output of the R commands.

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

- [illegible]

Installieren & Laden
der (Zusatz-)Pakete

Bearbeiten von
Variablen &
Datensätzen

Häufigkeiten

Lageparameter &
Streuungsmaße

Inferenzstatistische
Verfahren

Hauptkomponenten-
analyse & Cronbachs
Alpha

R (mosaic) – Key Commands

Get going

```
install.packages("xyz")
require(xyz) ← starting mosaic, psych, psy, ggplot2, openxlsx, plspm (required for each session)
```

Recoding, computing, formatting variables & creating new datasets:

Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:

```
dataset$new_variable <- dataset$variable1 + dataset$variable2
```

Changing the variable formats:

```
dataset$new_variable <- as.numeric(dataset$variable)
→ factor to numeric
(1): dataset$variable <- as.factor(dataset$variable)
(2): levels(dataset$variable) <- c("attribute1", "attribute1")
→ numeric to factor
```

Creating new datasets (is equal, is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value",]
new_dataset <- dataset[dataset$variable != "value",]
new_dataset <- dataset[dataset$variable > value,]
```

General command structure

```
command(y~x, z)
```

Frequencies

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

1

Prof. Dr.
Godbersen

Chi² test

```
xchisq.test(variable1~variable2, data = dataset)
```

t-test

```
t.test(variable~grouping_variable, data = dataset)
```

ANOVA

```
summary(aov(variable~grouping_variable, data = dataset))
→ alternative way:
modelxy <- aov(variable~grouping_variable, data = dataset)
summary(modelxy)
```

Shapiro-Wilk test

```
shapiro.test(dataset$variable)
```

Mann-Whitney & Wilcoxon test

```
wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

Correlation

```
cor.test(variable1~variable2, data = dataset)
```

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")
```

Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

R-Befehle: Befehlsstruktur

R (mosaic) – Key Commands

Get going

install.packages("xyz")
require(xyz) ← starting mosaic, psych, psy, ggplot2, openxlsx, plspm (required for each)

Recoding, computing, formatting variables & creating new datasets:

Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values

Calculating new variables:

dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the variable formats:

dataset\$new_variable <- as.numeric(dataset\$variable)

→ factor to numeric

(1): dataset\$variable <- as.factor(dataset\$variable)

(2): levels(dataset\$variable) <- c("attribute1", "attribute1")

→ numeric to factor

Creating new datasets (is equal, is not equal etc.):

new_dataset <- dataset[dataset\$variable == "value",]

new_dataset <- dataset[dataset\$variable != "value",]

new_dataset <- dataset[dataset\$variable > value,]

General command structure

command(y~x, z)

Frequencies

tally(~variable1, data = dataset)

tally(~variable1, format = "percent", data = dataset)

tally(variable1~variable2, data = dataset)

tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

max(~variable, data = dataset) / min(~variable, data = dataset)

favstats(~variable, data = dataset)

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values)

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

General command structure

command(y~x, z)

Frequencies

tally(~variable1, data = dataset)

tally(~variable1, format = "percent", data = dataset)

tally(variable1~variable2, data = dataset)

tally(variable1~variable2, format = "percent", data = dataset)

Hinweis zu den Befehlen

rot R-Befehl (darf nicht verändert werden)

blau Bezeichnungen von Variablen oder des Datensatzes, die individuell angepasst werden muss

y Abhängige Variable

x Unabhängige Variable

z Datensatz

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

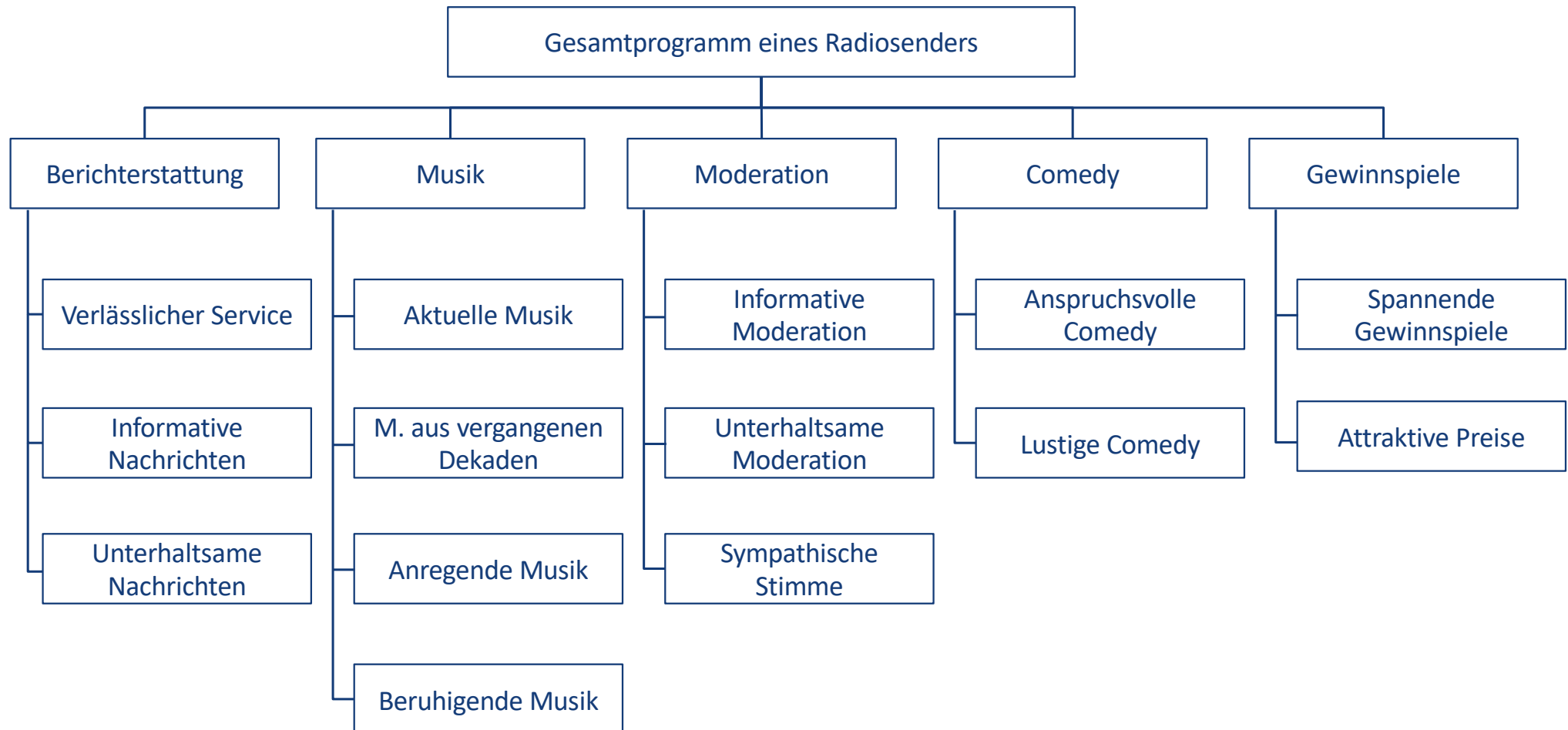
6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel



Übungsdaten: Fragebogen & Variablen

- **Stammsender**
 - Messung: in den letzten 2 Wochen am häufigster gehörter Sender
- **SenderCluster**
 - Messung: neu gebildete Variable mit drei Ausprägungen
 - Merkmalsausprägungen: SWR / Privat jung / Privat alt
- **Programmeigenschaften**
 - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
 - 14 Items
 - Verlässlicher_Service
 - Informative_Nachrichten
 - Unterhaltsame_Nachrichten
 - Aktuelle_Musik
 - Alte_Musik
 - Anregende_Musik
 - Beruhigende_Musik
 - Informative_Moderation
 - Unterhaltsame_Moderation
 - Sympathische_Stimme
 - Anspruchsvolle_Comedy
 - Lustige_Comedy
 - Spannende_Gewinnspiele
 - Attraktive_Preise
- **Programmelemente**
 - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
 - 5 Items
 - **Berichterstattung**
 - **Musik**
 - **Moderation**
 - **Comedy**
 - **Gewinnspiele**
- **Gesamtprogramm**
 - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
- **Geschlecht**
 - Messung: Auswahlfrage
- **Alter**
 - Messung: offene Altersangabe
- **Altersgruppen** (alt vs. jung, 2 Gruppen)
- **Aktuelle_Musik_A** (Affinität, 2 Gruppen)

Hinweis: Variablennamen in rot

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

Datenvorbereitung & Laden von Daten: Erstellen des Datensatzes

- **1. Schritt:**
 - Datentabelle im Excel-Format erstellen (auch andere Formate sind möglich)
 - Aus **Online-Fragebogen** im Excel-Format herunterladen
 - Von **schriftlicher Fragebogen** in Excel eintragen
- **Datenstruktur:**
 - 1. Zeile = Variablenbezeichnungen
 - 1 Spalte = 1 Variable
 - 1 Zeile = 1 Befragter
- **Empfehlung:**
 - Nur vollständige Datensätze in die Auswertung nehmen
- **Voraussetzungen für die Auswertung mit R:**
 - **Variablenbezeichnungen:** keine Leerzeichen & keine Umlaute
 - Empfehlung: kurze & bedeutungsvolle Bezeichnungen
 - **Merkmalsausprägungen:**
 - Metrische Variablen (z.B. Alter) als numerischen Wert
 - Kategoriale Variablen (z.B. Geschlecht) als Text

Variablenbezeichnungen

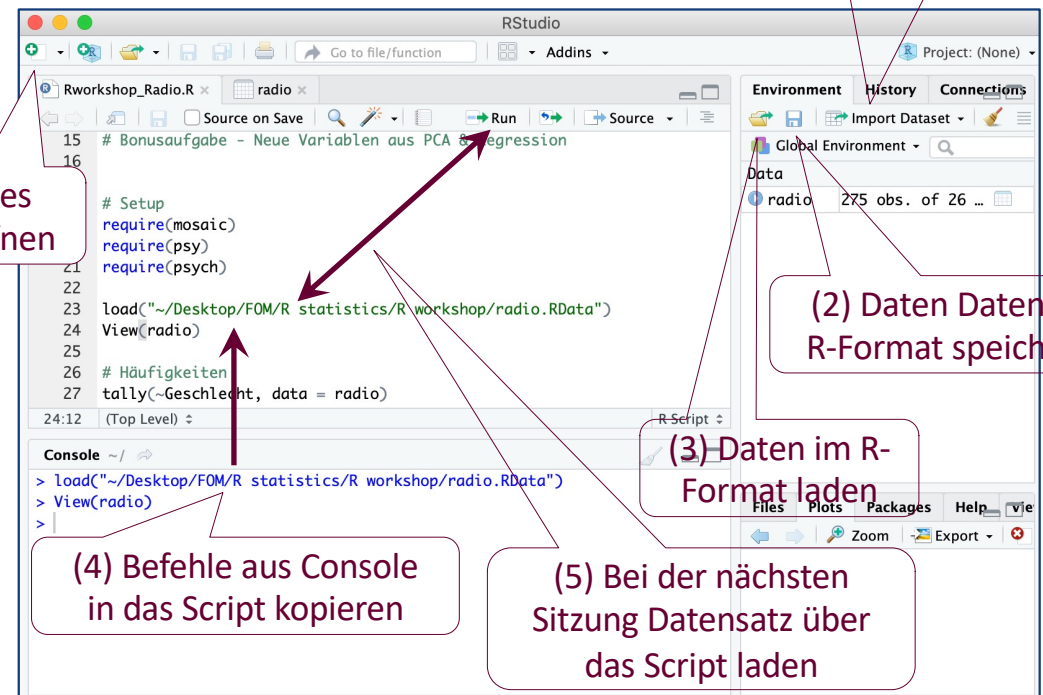
Befragter

Variable

1	Stammsender	SenderCluster	Verlaesslicher_Service	Informative_Nachrichten	Unterhalts_Akte_M	Alte_M
2	Energy	Privat jung	3	48	3	99
3	SWR3	SWR	6	7	3	20
4	Energy	Privat jung	6	86	86	100
5	Energy	Privat jung	6	54	41	100
6	BigFM	Privat jung	9	21	71	100
7	Energy	Privat jung	9	28	51	91
8	Antenne1	Privat alt	12	24	25	60
9	BigFM	Privat jung	14	27	22	86
10	BigFM	Privat jung	15	18	8	100
11	egoFM	Privat jung	16	27	21	63
12	BigFM	Privat jung	16	17	31	94
13	BigFM	Privat jung	18	38	7	100
14	BigFM	Privat jung	18	24	65	89
15	egoFM	Privat jung	19	37	43	96
16	BigFM	Privat jung	19	3	4	100
17	Das Ding	Privat jung	20	80	40	20
18	Antenne1	Privat alt	22	17	16	79
19	Energy	Privat jung	28	26	6	71
20	Das Ding	Privat jung	28	77	62	100
21	SWR3	SWR	29	37	33	29
22	SWR3	SWR	30	70	10	70
23	Ohr	Privat alt	33	47	33	78
24	SWR3	SWR	35	23	28	64
25	BigFM	Privat jung	35	78	6	86
26	BigFM	Privat jung	36	47	15	93
27	Energy	Privat jung	37	0	0	27
28	BigFM	Privat jung	38	62	82	95

Datenvorbereitung & Laden von Daten: Daten in R laden

- **Vorbereitung (wichtig):**
 - Speichern Sie die Daten in einem Ordner auf Ihrem Computer, der extra für das Projekt angelegt ist
- **Daten aus Excel importieren (1):**
 - Unter „Environment“ Daten über „Import Dataset“ Excel-Daten auswählen
- **Daten im R-Format speichern (2):**
 - Unter Environment über Speichersymbol Daten im Projektordner speichern
- **Daten im R-Format laden (3):**
 - Unter „Environment“ Daten über Öffnungssymbol R-Daten auswählen
 - Unter „Environment“ auf den Datensatz klicken
 - In der Console generierten Befehle in das Script kopieren (4)
 - Bei der nächsten Anwendung können die Daten über das Script gestartet werden (5)



1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

Daten bearbeiten in R: Beispiele

- Recodieren von Variablen:
 - Inverse Skalen, z.B. 1 zu 6, 2 zu 5, 3 zu 4...
 - Zusammenfassen von Werten, z.B. Einwohner von Bayern und Baden-Württemberg (erhobene Variable) zu Süddeutschland (neu zu bildende Variable)
 - Ändern von metrischen zu kategorialen Werte (siehe Datenvorbereitung für R)
 - ...
- Berechnen neuer Variablen:
 - Bildung eines neuen („übergeordneten“) Konstrukts aus mehreren Variablen, z.B. Berechnung einer Gesamtbewertung aus mehreren Einzelbewertungen über das arithmetische Mittel
 - Berechnung einer neuen Variable, die sich aus dem Verhältnis mehrerer Variablen bestimmt, z.B. Berechnung der Rentabilität aus den Variablen Gewinn und Umsatz
 - ...
- Erstellen neuer Datensätze:
 - Ausschluss von Befragten mit bestimmten Merkmalsausprägungen, z.B. zu alte oder zu junge Befragte
 - Analyse nur eines Teils der Befragten, z.B. ausschließliche Analyse der Antworten von Frauen
 - ...

Daten bearbeiten in R: Übungsaufgaben & R-Befehle

Aufgaben:

- Generieren Sie aus SenderCluster eine neue Variable (SenderCluster2): „Privat jung“ & „Privat alt“ zu „Privat“ und „SWR“ zu „SWR“.
- Berechnen Sie aus den Variablen „Comedy“ und „Gewinnspiele“ in einer neuen Variable (CoGe) den Mittelwert.
- Generieren Sie einen neuen Datensatz, der nur aus Frauen besteht.

- Stammsender
- **SenderCluster**
- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten

Variablen

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) <- starting mosaic, psych

Recoding, computing, formatting variables
Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable == "XXX"] <- "new_value"
→ Please note: categorial values require quotes (""), numerical values not

Calculating new variables:
dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the the variable formats:
dataset\$new_variable <- as.factor(dataset\$variable1)
→ factor to numeric:
(1) dataset\$variable <- as.numeric(dataset\$variable)
(2) levels(dataset\$variable) <- as.factor(levels(dataset\$variable))
→ numeric to factor

Creating new datasets (is equal, is not equal etc.):
new_dataset <- dataset[dataset\$variable == "value",]
new_dataset <- dataset[dataset\$variable != "value",]
new_dataset <- dataset[dataset\$variable > value,]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation
median(~variable, data = dataset)
var(~variable, data = dataset)
max(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, mean, max

Recoding (new) variables:

dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"

→ Please note: categorial values require quotes (""), numerical values not

Calculating new variables:

dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the the variable formats:

...

Creating new datasets (is equal, is not equal etc.):

new_dataset <- dataset[dataset\$variable == "value",]

new_dataset <- dataset[dataset\$variable != "value",]

new_dataset <- dataset[dataset\$variable > value,]

- Berichterstattung
- Musik
- Moderation
- **Comedy**
- **Gewinnspiele**
- Gesamtprogramm
- **Geschlecht**
- Alter
- Altersgruppen
- Aktuelle_Musik_A

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

Skalenniveau		Math. Eigenschaften	Beschreibung der Messwertcharakteristika	Beispiel	Lageparameter
Kategorial	Nominal	$\neq$	Die Messwerte sind identisch oder nicht identisch.	Geschlecht	• Modus
	Ordinal	$\neq ; </>$	Die Messwerte sind größer, kleiner oder identisch.	Olympische Ränge	• Modus • Median
Metrisch	Intervall	$\neq ; </> ; - ; +$	Die Distanz zwischen den Messwerten kann angegeben werden.	Temperatur	• Modus • Median • Arithm. Mittel
	Verhältnis	$\neq ; </> ; +/- ; * / \div$	Die Distanz zwischen den Messwerten und deren Verhältnis kann angegeben werden.	Körpergröße	

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

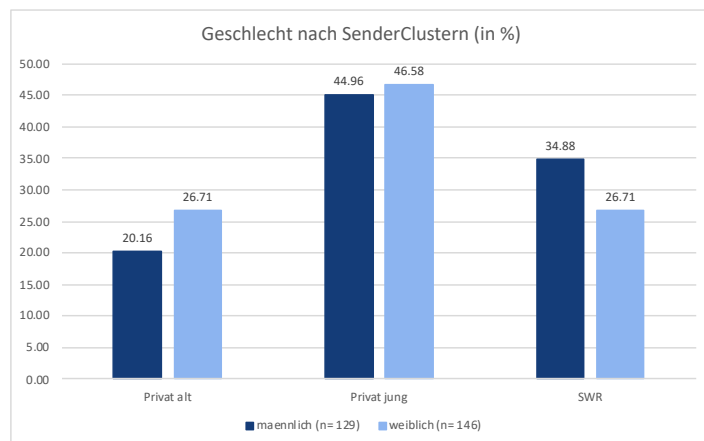
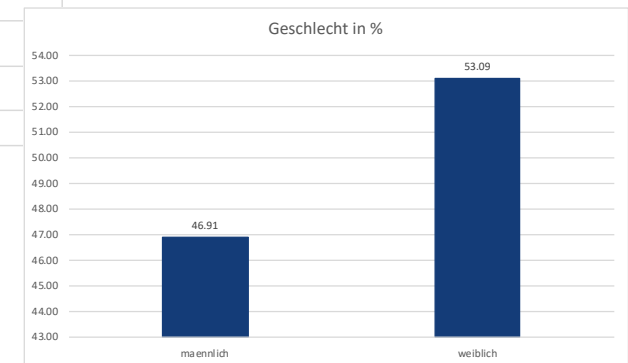
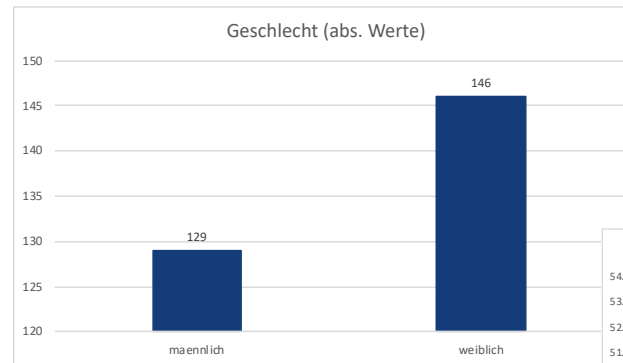
7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

Häufigkeiten: Steckbrief

- Absolute Häufigkeit:
 - gibt an wie häufig ein Wert auftritt
- Relative Häufigkeit:
 - gibt das Verhältnis eines untersuchten Merkmalsträgers zu einem Wert oder einer Klasse von Werten an



- Kreuztabelle:
 - Darstellung der gemeinsamen Verteilung zweier Variablen

Häufigkeiten: Übungsaufgaben & R-Befehle (1)

Aufgaben:

- Wie viele Männer und Frauen gibt es in der Stichprobe?
- Wie hoch ist der Anteil der Männer und Frauen in %?
- Wie hoch ist der Anteil der SenderCluster in %?

Frequencies

`tally(~variable1, data = dataset)`

`tally(~variable1, format = "percent", data = dataset)`

`tally(variable1~variable2, data = dataset)`

`tally(variable1~variable2, format = "percent", data = dataset)`

→ numeric to factor

```
Creating new datasets (a equal, is not equal etc.):
new_dataset <- dataset(dataset$variable == "value")
new_dataset <- dataset(dataset$variable != "value")
new_dataset <- dataset(dataset$variable > value.)
```

General command structure

```
command(y~x, z)
```

Frequencies

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 1

Wilcoxon test

```
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

Correlation

```
cor.test(variable1~variable2, data = dataset)
```

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 2

Variablen

- Stammsender
- **SenderCluster**
- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise
- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele
- Gesamtprogramm
- **Geschlecht**
- Alter
- Altersgruppen
- Aktuelle_Musik_A

Häufigkeiten: Übungsaufgaben & R-Befehle (2)

Aufgaben:

- Wie verteilen sich Männer und Frauen auf die SenderCluster (%)?

Frequencies

```
tally(~variable1, data = dataset)
```

```
tally(~variable1, format = "percent", data = dataset)
```

```
tally(variable1~variable2, data = dataset)
```

```
tally(variable1~variable2, format = "percent", data = dataset)
```

→ numeric to factor

```
Creating new datasets (a equal, is not equal etc.):
new_dataset <- dataset(dataset$variable == "value")
new_dataset <- dataset(dataset$variable != "value")
new_dataset <- dataset(dataset$variable > value)
```

General command structure

```
command(y~x, z)
```

Frequencies

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Wilcoxon test

```
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

Correlation

```
cor.test(variable1~variable2, data = dataset)
```

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

- | | |
|--|---|
| <ul style="list-style-type: none"> Stammsender SenderCluster Verlaesslicher_Service Informative_Nachrichten Unterhaltsame_Nachrichten Aktuelle_Musik Alte_Musik Anregende_Musik Beruhigende_Musik Informative_Moderation Unterhaltsame_Moderation Sympathische_Stimme Anspruchsvolle_Comedy Lustige_Comedy Spannende_Gewinnspiele Attraktive_Preise | <p>Variablen</p> <ul style="list-style-type: none"> Berichterstattung Musik Moderation Comedy Gewinnspiele Gesamtprogramm Geschlecht Alter Altersgruppen Aktuelle_Musik_A |
|--|---|

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

- **Lageparameter:**
 - Arithmetisches Mittel:
 - Division der Summe aller Werte durch die Anzahl der Werte; metrisches Skalenniveau notwendig
 - Median:
 - Wert, der die Grenze zwischen oberer und unterer Hälfte markiert; mindestens ordinales Skalenniveau
 - Modalwert:
 - Häufigste Wert einer Häufigkeitsverteilung; mindestens nominales Skalenniveau
- **Streuungsparameter**
 - Varianz
 - Durchschnittliche quadrierte Abweichung der gemessenen Werte vom arithmetischen Mittel (beachte: empirische Varianz $\rightarrow /n$; Stichprobenvarianz $\rightarrow /n-1$)
 - Standardabweichung
 - Durchschnittliche Abweichung der gemessenen Werte vom arithmetischen Mittel

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

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

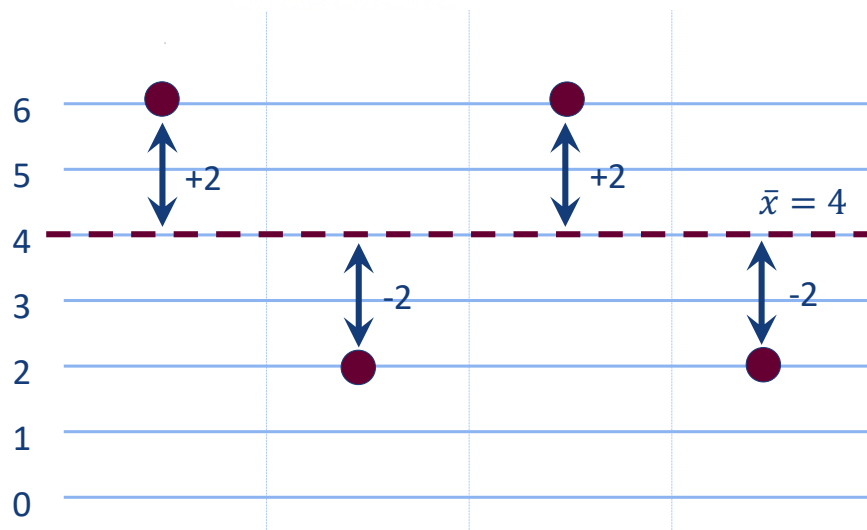
• Streuungsparameter

- Varianz

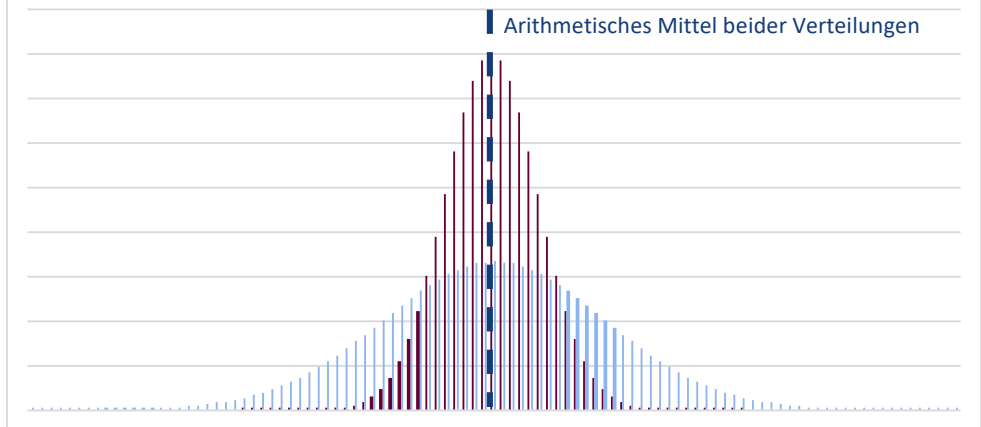
$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

- Standardabweichung

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



Normalverteilungen mit unterschiedlichen Varianzen



$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}) &= \{ (+2) + (-2) + (+2) + (-2) = 0 \} \\ \sum_{i=1}^n (x_i - \bar{x})^2 &= \{ (+2)^2 + (-2)^2 + (+2)^2 + (-2)^2 = 4 + 4 + 4 + 4 = 16 \} \\ \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} &= \{ 16 / 4 = 4 \} \\ \sigma = \sqrt{\sigma^2} &= \{ \sqrt{4} = 2 \} \end{aligned}$$

Lageparameter & Streuungsmaße: Übungsaufgaben & R-Befehle (1)

Aufgaben:

- Was ist das durchschnittliche Alter der Befragten (arith. Mittel)?
- Welchen Wert hat der Median bei der Variable Alter?
- Wie stark streuen die gemessenen Werte der Variable Alter um das arithmetische Mittel (Varianz & Standardabweichung)?

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
```

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

```
Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

Variablen

- Stammsender
- SenderCluster
- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise
- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele
- Gesamtprogramm
- Geschlecht
- **Alter**
- Altersgruppen
- Aktuelle_Musik_A

Lageparameter & Streuungsmaße: Übungsaufgaben & R-Befehle (2)

Aufgaben:

- Wie hoch ist die durchschnittliche Bewertung des Gesamtprogramms (arith. Mittel)?
- Welchen Wert hat der Median bei der Variable Gesamtprogramm?
- Wie stark streuen die gemessenen Werte der Variable Gesamtprogramm um das arithmetische Mittel (Varianz & Standardabweichung)?

Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

Variablen

- Stammsender
- SenderCluster

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele

• Gesamtprogramm

- Geschlecht
- Alter
- Altersgruppen
- Aktuelle_Musik_A

Aufgaben:

- Bestimmen Sie für die Variable Alter das arithmetische Mittel, den Median, die Standardabweichung etc. mit nur einem R-Befehl.
- Bestimmen Sie für die Variable Gesamtprogramm das arithmetische Mittel, den Median, die Standardabweichung etc. mit nur einem R-Befehl.

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
```

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

- Stammsender
- SenderCluster

Variablen

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele

Gesamtprogramm

- Geschlecht
- Alter
- Altersgruppen
- Aktuelle_Musik_A

Lageparameter & Streuungsmaße: Übungsaufgaben & R-Befehle (4)

Aufgaben:

- Stellen Sie das durchschnittliche Alter von Männern und Frauen gegenüber (arith. Mittelwerte).
- Stellen Sie die durchschnittliche Bewertung des Gesamtprogramms von Männern und Frauen gegenüber (arith. Mittelwerte).

Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`
`var(~variable, data = dataset) / sd(~variable, data = dataset)`
`max(~variable, data = dataset) / min(~variable, data = dataset)`
`favstats(~variable, data = dataset)`

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
  -> result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

- Stammsender
- SenderCluster

Variablen

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele

Gesamtprogramm

- Geschlecht
- Alter

- Altersgruppen
- Aktuelle_Musik_A

1 Einleitung	6 Inferenzstatistische Verfahren
2 Installation & Grundstruktur von R Studio	6.1 Überblick & Systematisierung der Verfahren
2.1 Installation von R, R Studio & relevanten Paketen	6.2 Chi-Quadrat-Test
2.2 Grundstruktur von R Studio	6.3 t-Test
3 R-Befehle & Übungsdaten	6.4 ANOVA
3.1 R-Befehle	6.5 Shapiro-Wilk-Test
3.2 Übungsdaten	6.6 Wilcoxon-Test
4 Datenmanagement	6.7 Korrelation
4.1 Datenvorbereitung & Laden von Daten	6.8 Regression
4.2 Daten bearbeiten in R	6.9 Hauptkomponentenanalyse
5 Deskriptive Statistik	7 Graphiken & Export von Ergebnissen nach Excel
5.1 Skalenniveaus	7.1 Graphiken in R
5.2 Häufigkeiten	7.2 Export von Ergebnissen nach Excel
5.3 Lageparameter & Streuungsmaße	

Überblick & Systematisierung der Verfahren

Prof. Dr.
Godbersen

			Unabhängige Variable			
Struktur	Skalenniveau		Math. Eigenschaften	Beschreibung der Messwertcharakteristika	Beispiel	Lageparameter
Abhängige Variable	Kategorial	Nominal	$\neq$	Die Messwerte sind identisch oder nicht identisch.	Geschlecht	• Modus
		Ordinal	$\neq ; </>$	Die Messwerte sind größer, kleiner oder identisch.	Olympische Ränge	• Modus • Median
*** „Ergänzung“ • Struktur (p) • Maß (p)	Metrisch	Intervall	$\neq ; </> ; - ; +$	Die Distanz zwischen den Messwerten kann angegeben werden.	Temperatur	• Modus • Median • Arithm. Mittel
		Verhältnis	$\neq ; </> ; +/- ; */\div$	Die Distanz zwischen den Messwerten und deren Verhältnis kann angegeben werden.	Körpergröße	
Strukturentdeckende Verfahren			• Hauptkomponentenanalyse (Faktoranalyse) • Clusteranalyse			

Strukturprüfende Verfahren		Unabhängige Variable	
		nicht-metrisch	metrisch
Abhängige Variable	nicht-metrisch	Chi ² -Test	Diskriminanzanalyse
	metrisch	t-Test*** (bei 2 Gruppen) & Varianzanalyse (bei ≥ 3 Gruppen)	Regressionsanalyse (Dependenz) & Korrelationsanalyse (Interdependenz)

*** „Ergänzende“ Tests:

- Shapiro-Wilk-Test
(prüft Stichproben auf Normalverteilung, die für den t-Test notwendig ist)
- Mann-Whitney-Test/Wilcoxon-Test
(prüft mind. ordinal skalierte Daten auf Unterschiede ohne Berücksichtigung der Verteilungsform)

Strukturentdeckende Verfahren	• Hauptkomponentenanalyse (Faktoranalyse) • Clusteranalyse
-------------------------------	---

1 Einleitung	6 Inferenzstatistische Verfahren
2 Installation & Grundstruktur von R Studio	6.1 Überblick & Systematisierung der Verfahren
2.1 Installation von R, R Studio & relevanten Paketen	6.2 Chi-Quadrat-Test
2.2 Grundstruktur von R Studio	6.3 t-Test
3 R-Befehle & Übungsdaten	6.4 ANOVA
3.1 R-Befehle	6.5 Shapiro-Wilk-Test
3.2 Übungsdaten	6.6 Wilcoxon-Test
4 Datenmanagement	6.7 Korrelation
4.1 Datenvorbereitung & Laden von Daten	6.8 Regression
4.2 Daten bearbeiten in R	6.9 Hauptkomponentenanalyse
5 Deskriptive Statistik	7 Graphiken & Export von Ergebnissen nach Excel
5.1 Skalenniveaus	7.1 Graphiken in R
5.2 Häufigkeiten	7.2 Export von Ergebnissen nach Excel
5.3 Lageparameter & Streuungsmaße	

- **Anwendung:**
 - Testen auf (Un-)Abhängigkeit bei nominalen Daten (Weichen die gemessenen Werte nicht-zufällig von den erwarteten Werten ab?)
- **Fragestellung:**
 - Gibt es einen „nicht-zufälligen“ statistischen Zusammenhang zwischen zwei nominalskalierten Merkmalen? Sind zwei kategoriale Variablen unabhängig?
- **Vorgehen bei der Analyse:**
 - 1) Ist der p-value < 0.05 (< 0.01 , < 0.001)? → Signifikanz?
 - 2) Wie unterscheiden sich die Häufigkeiten?
- **„Statistischen“ Analyseschritte („im Hintergrund“):**
 - 1) Berechnung der erwarteten Häufigkeiten (e), wenn kein Unterschied zwischen den Variablen bestehen würde (Spaltensumme * Zeilensumme / Merkmalsträger)
 - 2) Bestimmung, ob die gemessenen Werte (h) signifikant von den erwarteten Werten abweichen?

	männlich	weiblich	Summe
blau	h = 35 e = 25	h = 15 e = 25	50
rot	h = 15 e = 25	h = 35 e = 25	50
Summe	50	50	100

- „Statistischen“ Analyseschritte („im Hintergrund“):
 - 1) Berechnung der erwarteten Häufigkeiten (e), wenn kein Unterschied zwischen den Variablen bestehen würde (Spaltensumme * Zeilensumme / Merkmalsträger)
 - 2) Bestimmung, ob die gemessenen Werte (h) signifikant von den erwarteten Werten abweichen?

V1	männlich	weiblich	Summe
blau	h = 15 e = 25	h = 35 e = 25	50
rot	h = 35 e = 25	h = 15 e = 25	50
Summe	50	50	100

p-value = 0.0001447

V2	männlich	weiblich	Summe
blau	h = 24 e = 25	h = 26 e = 25	50
rot	h = 26 e = 25	h = 24 e = 25	50
Summe	50	50	100

p-value = 0.8415

Chi-Quadrat-Test: Übungsaufgaben & R-Befehle

Aufgaben:

- Besteht ein signifikanter Unterschied zwischen Männern und Frauen bei der Präferenz für die SenderCluster?
- Unterscheidet sich das junge und alte Alterssegment bei der Affinität zu aktueller Musik? Und wenn ja, wie fällt dieser Unterschied aus?

Prof. Dr. Godbersen

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) <- starting mosaic, psych, psy, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:

Chi² test

xchisq.test(variable1~variable2, data = dataset)

(2): levels(dataset\$variable) <- c("attribute1", "attribute1")
→ numeric to factor

Creating new datasets (is equal, is not equal etc.):
new_dataset <- dataset[dataset\$variable == "value",]
new_dataset <- dataset[dataset\$variable != "value",]
new_dataset <- dataset[dataset\$variable > value,]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 1

Prof. Dr. Godbersen

Chi² test
xchisq.test(variable1~variable2, data = dataset)

t-test
t.test(variable~grouping_variable, data = dataset)

ANOVA
summary(aov(variable~grouping_variable, data = dataset))

wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))

Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
cronbach(factorX)

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 2

- | | |
|--|--|
| <ul style="list-style-type: none"> • Stammsender • SenderCluster • Verlaesslicher_Service • Informative_Nachrichten • Unterhaltsame_Nachrichten • Aktuelle_Musik • Alte_Musik • Anregende_Musik • Beruhigende_Musik • Informative_Moderation • Unterhaltsame_Moderation • Sympathische_Stimme • Anspruchsvolle_Comedy • Lustige_Comedy • Spannende_Gewinnspiele • Attraktive_Preise | <p>Variablen</p> <ul style="list-style-type: none"> • Berichterstattung • Musik • Moderation • Comedy • Gewinnspiele • Gesamtprogramm • Geschlecht • Alter • Altersgruppen • Aktuelle_Musik_A |
|--|--|

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

- **Anwendung:**
 - Der t-Test ist ein parametrisches Verfahren. Beim Zweistichproben-t-Test wird überprüft, ob sich eine metrische Variable zwischen zwei Stichproben signifikant (nicht zufällig) unterscheidet.
- **Fragestellung:**
 - Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen zwei Stichproben in Bezug auf ein metrisches Merkmal?
 - Gibt es einen „nicht-zufälligen“ Einfluss einer nominalen Variable (2 Gruppen) auf eine metrische Variable?
- **Differenzierung zwischen:**
 - t-Test bei unabhängigen Stichproben – Messungen bei zwei Gruppen
 - t-Test bei abhängigen Stichproben – Messungen bei derselben Gruppe (z.B. vorher – nachher)
- **Vorgehen bei der Analyse:**
 - 1) Ist der p-value <0.05 (<0.01 , <0.001)? → Signifikanz?
 - 2) Wie unterscheiden sich die Mittelwerte?

t-Test: Übungsaufgaben & R-Befehle

Aufgaben:

- Beurteilen Männer und Frauen das Gesamtprogramm von Radiosendern (signifikant) unterschiedlich?
- Beurteilen Männer und Frauen die Musik von Radiosendern (signifikant) unterschiedlich?

- Stammsender
- SenderCluster

Variablen

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- Berichterstattung
- **Musik**
- Moderation
- Comedy
- Gewinnspiele
- **Gesamtprogramm**
- **Geschlecht**
- Alter
- Altersgruppen
- Aktuelle_Musik_A

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) <- starting mosaic, psych, psy, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:
Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:
dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the the variable format:
dataset\$new_variable <- as.numeric(dataset\$variable)

Chi² test
xchisq.test(variable1~variable2, data = dataset)

t-test
t.test(variable~grouping_variable, data = dataset)

ANOVA
summary(aov(variable~grouping_variable, data = dataset))
→ alternative way:
modelxy <- aov(variable~grouping_variable, data = dataset)
summary(modelxy)

Shapiro-Wilk test
shapiro.test(dataset\$variable)

t-test

t.test(variable~grouping_variable, data = dataset)

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
fvstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- prcomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
cronbach(factorX)

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

- **Verwendung:**
 - Die Varianzanalyse (*Analysis of Variance, ANOVA*) ist ein parametrisches Verfahren und untersucht, ob sich die Mittelwerte mehrerer (mehr als zwei) unabhängiger Stichproben systematisch unterscheiden.
- **Fragestellung:**
 - Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen mehr als zwei Stichproben in Bezug auf ein metrisches Merkmal?
- **Abgrenzung zum t-Test:**
 - Während der t-Test metrische Merkmale von maximal zwei Stichproben auf signifikante Unterschiede prüft, werden bei der Varianzanalyse mehrere Mittelwerte verglichen
- **Beispiel:**



ANOVA: Übungsaufgaben & R-Befehle

Aufgaben:

- Beurteilen die Hörer der verschiedenen SenderCluster das Gesamtprogramm von Radiosendern (signifikant) unterschiedlich?
- Beurteilen die Hörer der verschiedenen SenderCluster die Moderation von Radiosendern (signifikant) unterschiedlich?

- | | |
|---|--|
| <ul style="list-style-type: none"> • Stammsender • SenderCluster | Variablen |
| <ul style="list-style-type: none"> • Verlaesslicher_Service • Informative_Nachrichten • Unterhaltsame_Nachrichten • Aktuelle_Musik • Alte_Musik • Anregende_Musik • Beruhigende_Musik • Informative_Moderation • Unterhaltsame_Moderation • Sympathische_Stimme • Anspruchsvolle_Comedy • Lustige_Comedy • Spannende_Gewinnspiele • Attraktive_Preise | <ul style="list-style-type: none"> • Berichterstattung • Musik • Moderation • Comedy • Gewinnspiele • Gesamtprogramm • Geschlecht • Alter • Altersgruppen • Aktuelle_Musik_A |

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) ← starting mosaic, psych, psy, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:

Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:

dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the variable format:

dataset\$new_variable <- as.numeric(dataset\$variable)
→ factor to numeric

Prof. Dr.
Godbersen



Prof. Dr.
Godbersen

Chi² test

xchisq.test(variable1~variable2, data = dataset)

t-test

t.test(variable~grouping_variable, data = dataset)

ANOVA

summary(aov(variable~grouping_variable, data = dataset))

→ alternative way:

modelxy <- aov(variable~grouping_variable, data = dataset)

summary(modelxy)

Shapiro-Wilk test

shapiro.test(dataset\$variable)

Mann-Whitney & Wilcoxon test

ANOVA

summary(aov(variable~grouping_variable, data = dataset))

command1(~x, z)

Frequencies

tally(~variable1, data = dataset)

tally(~variable1, format = "percent", data = dataset)

tally(variable1~variable2, data = dataset)

tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

max(~variable, data = dataset) / min(~variable, data = dataset)

favstats(~variable, data = dataset)

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

1

Linear regression

summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression

summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)

dataset_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)

colnames(dataset_new) <- c("item_1", "item_2")

KMO(dataset_new)

pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)

summary(pcaX)

plot(pcaX)

principal(dataset_new, nfactors = x, rotate = "varimax")

Cronbach's Alpha (package: psy)

factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)

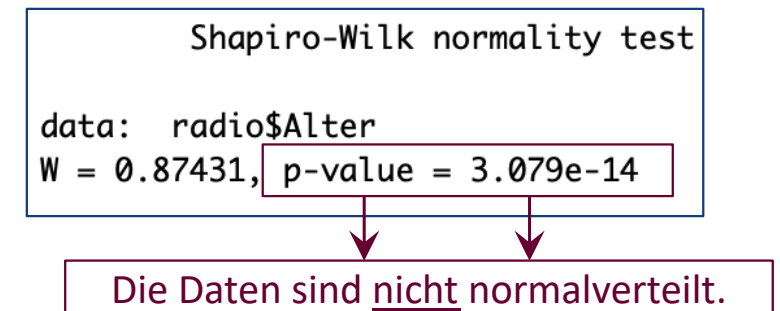
cronbach(factorX)

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

2

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

- **Verwendung:**
 - Der Shapiro-Wilk-Test vergleicht die Werte einer Stichprobe mit normalverteilten Werten, um zu überprüfen, ob eine Normalverteilung angenommen werden kann. **Bei kleinen Stichproben (ca. $n < 50$) ist die Normalverteilung Voraussetzung für die Anwendung des t-Tests.**
- **Fragestellung & Ergebnisinterpretation:**
 - Sind die Werte einer mindestens intervallskalierten Variable normalverteilt?
 - Ist der Test nicht signifikant ($p > .10$), unterscheidet sich die Verteilung der vorliegenden Stichprobe nicht von einer normalverteilten Stichprobe → **Normalverteilung**
 - Ist der Test signifikant ($p \leq .10$), unterscheidet sich die Verteilung der vorliegenden Stichprobe von einer normalverteilten Stichprobe → **keine Normalverteilung**
- **Konsequenz bei fehlender Normalverteilung:**
 - U.a. werden statt des t-Tests Rangverfahren wie Wilcoxon-/Mann-Whitney-Test angewendet
 - Beachte: I.d.R. ist es sinnvoll, nicht nur den Shapiro-Wilk-Test durchzuführen, sondern auch einen Blick auf die (graphische) Verteilung zu werfen.



Shapiro-Wilk-Test: Übungsaufgaben & R-Befehle

Aufgaben:

- Ist die Variabel Alter normalverteilt?
- Ist die Variabel Gesamtprogramm normalverteilt?

- Stammsender
- SenderCluster

Variablen

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele
- Gesamtprogramm**
- Geschlecht
- Alter**
- Altersgruppen
- Aktuelle_Musik_A

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) <- starting mosaic, psych, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:

Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:

dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the variable format:

dataset\$new_variable <- as.numeric(dataset\$variable)
→ factor to numeric
(1): dataset\$variable <- as.factor(dataset\$variable)
(2): levels(dataset\$variable) <- c("attribute1", "attribute1")
→ numeric to factor

Creating new datasets (e.g. repeat, in, not repeat, etc.)

Chi² test

xchisq.test(variable1~variable2, data = dataset)

t-test

t.test(variable~grouping_variable, data = dataset)

ANOVA

summary(aov(variable~grouping_variable, data = dataset))
→ alternative way:
modelxy <- aov(variable~grouping_variable, data = dataset)
summary(modelxy)

Shapiro-Wilk test

shapiro.test(dataset\$variable)

Mann-Whitney & Wilcoxon test

wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)

Shapiro-Wilk test

shapiro.test(dataset\$variable)

tally(~variable1, data = dataset)

tally(~variable1, format = "percent", data = dataset)

tally(variable1~variable2, data = dataset)

tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

max(~variable, data = dataset) / min(~variable, data = dataset)

fvstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

ANOVA - ANCOVA

summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)

dataset_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)

colnames(dataset_new) <- c("item_1", "item_2")

KMO(dataset_new)

pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)

summary(pcaX) plot(pcaX)

principal(dataset_new, nfactors = 4, rotate = "varimax")

Cronbach's Alpha (package: psy)

factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)

cronbach(factorX)

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

Wilcoxon-Test: Steckbrief

- **Verwendung:**
 - Der Rangsummen-Test ist ein non-parametrisches Verfahren und vergleicht die *mittleren Rangzahlen* zweier Gruppen. (Es liegt keine Verteilungsannahme vor, u.a. Normalverteilung wurde auf Grund von Shapiro-Wilk-Test abgelehnt)
- **Fragestellung:**
 - Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen zwei Gruppen in Bezug auf ein mindestens ordinalskaliertes Merkmal?
- **Beispiel:**

```
Wilcoxon rank sum test with continuity correction  
  
data: Gewinnspiele by Geschlecht  
W = 7577.5, p-value = 0.005188  
alternative hypothesis: true location shift is not equal to 0
```

Gruppen unterscheiden sich

Wilcoxon-Test: Übungsaufgaben & R-Befehle

Aufgaben:

- Liegt nach dem Wilcoxon-Test ein signifikanter Unterschied zwischen Männern und Frauen bei der Beurteilung des Gesamtprogramms vor?
- Liegt nach dem Wilcoxon-Test ein signifikanter Unterschied zwischen Männern und Frauen bei der Beurteilung der Musik vor?

- | | |
|---|---|
| <ul style="list-style-type: none"> Stammsender SenderCluster Verlaesslicher_Service Informative_Nachrichten Unterhaltsame_Nachrichten Aktuelle_Musik Alte_Musik Anregende_Musik Beruhigende_Musik Informative_Moderation Unterhaltsame_Moderation Sympathische_Stimme Anspruchsvolle_Comedy Lustige_Comedy Spannende_Gewinnspiele Attraktive_Preise | <p>Variablen</p> <ul style="list-style-type: none"> Berichterstattung Musik Moderation Comedy Gewinnspiele Gesamtprogramm Geschlecht Alter Altersgruppen Aktuelle_Musik_A |
|---|---|

Prof. Dr.
Godbersen

Prof. Dr.
Godbersen

R (mosaic) – Key Commands

Get going
install.packages("xyz")
require(xyz) <- starting mosaic, psych, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:
Recoding (new) variables:
dataset\$new_variable[dataset\$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:
dataset\$new_variable <- dataset\$variable1 + dataset\$variable2

Changing the variable format:
dataset\$new_variable <- as.numeric(dataset\$variable)
→ factor to numeric
(1): dataset\$variable <- as.factor(dataset\$variable)
(2): levels(dataset\$variable) <- c("attribute1", "attribute1")
→ numeric to factor

Creating new datasets (is equal, is not equal etc.):
new_dataset <- dataset[dataset\$variable == "value",]
new_dataset <- dataset[dataset\$variable != "value",]
new_dataset <- dataset[dataset\$variable > value,]

Chi² test
xchisq.test(variable1~variable2, data = dataset)

t-test
t.test(variable~grouping_variable, data = dataset)

ANOVA
summary(aov(variable~grouping_variable, data = dataset))
→ alternative way:
model <- aov(variable~grouping_variable, data = dataset)
summary(model)

Shapiro-Wilk test
shapiro.test(dataset\$variable)

Mann-Whitney & Wilcoxon test
wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))

Correlation
cor.test(variable1~variable2, data = dataset)

Mann-Whitney & Wilcoxon test

wilcox.test(variable~grouping_variable, data = dataset)

(→ Mann-Whitney test)

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 1

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands 2

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

Korrelation: Steckbrief

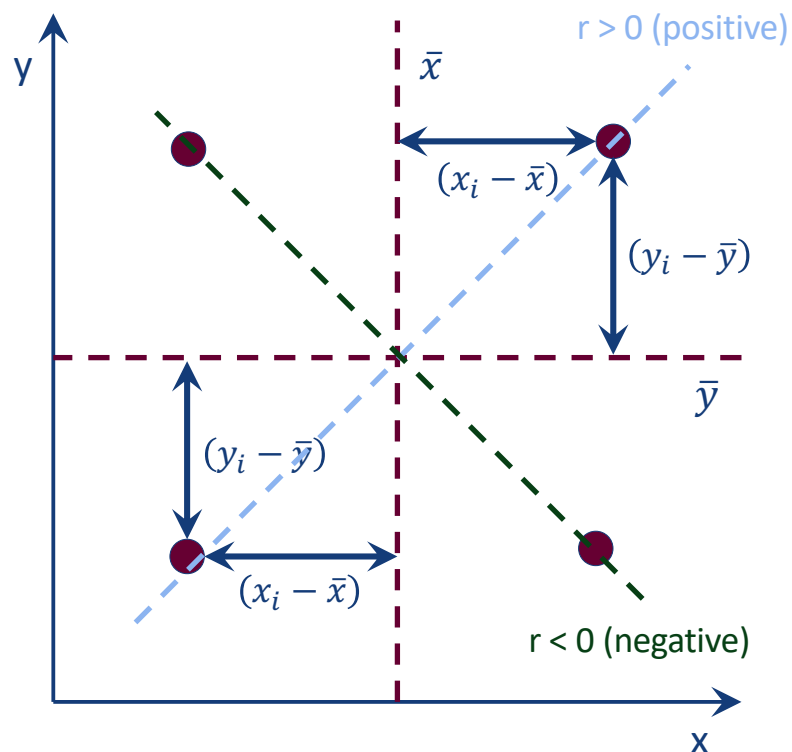
- **Verwendung:**
 - Die Korrelationsanalyse untersucht den (linearen) Zusammenhang zwischen zwei oder mehr metrischen Variablen.
- **Fragestellung:**
 - Wie stark ist der lineare Zusammenhang zwischen den Variablen – und in welcher Richtung besteht er?
→ Wird der Wert von Variable A erhöht / gesenkt oder bleibt er gleich, wenn sich der Wert von Variable B ändert?
- **Vorgehen:**
 - 1) Ist der p-value < 0.05 (< 0.01 , < 0.001)? → Signifikanz?
 - 2) Wie hoch ist der Korrelationskoeffizient?
- **Mögliche Interpretation des Korrelationskoeffizienten:**

$0 < r \leq +1.0$	gleichgerichteter Zusammenhang
$-1.0 \geq r < 0$	entgegengesetzter Zusammenhang

$0.2 < r \leq 0.5$	geringe Korrelation
$0.5 < r \leq 0.7$	mittlere Korrelation
$0.7 < r \leq 1.0$	hohe Korrelation

Korrelation: Steckbrief

- Statistische Erklärung des Korrelationskoeffizienten



Kovarianz (gemeinsame Streuung zweier Variablen):

$$\text{cov} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Varianz (Streuung einer Variablen):

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

Korrelationskoeffizient (Anteil der gemeinsamen Streuung zweier Variablen an der gesamten Streuung):

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\text{var}_x * \text{var}_y}$$

Korrelation: Übungsaufgaben & R-Befehle

Aufgaben:

- Gibt es einen nicht-zufälligen Zusammenhang zwischen der Beurteilung der Musik und der Moderation?
- Gibt es einen nicht-zufälligen Zusammenhang zwischen der Beurteilung von Comedy und Gewinnspielen?

- | | |
|---|--|
| <ul style="list-style-type: none"> Stammsender SenderCluster | Variablen |
| <ul style="list-style-type: none"> Verlaesslicher_Service Informative_Nachrichten Unterhaltsame_Nachrichten Aktuelle_Musik Alte_Musik Anregende_Musik Beruhigende_Musik Informative_Moderation Unterhaltsame_Moderation Sympathische_Stimme Anspruchsvolle_Comedy Lustige_Comedy Spannende_Gewinnspiele Attraktive_Preise | <ul style="list-style-type: none"> Berichterstattung Musik Moderation Comedy Gewinnspiele Gesamtprogramm Geschlecht Alter Altersgruppen Aktuelle_Musik_A |

Prof. Dr. Godbersen
Chi² test
Prof. Dr. Godbersen

Correlation

cor.test(variable1~variable2, data = dataset)

→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:

```
dataset$new_variable <- dataset$variable1 + dataset$variable2
```

Changing the variable format:

```
dataset$new_variable <- as.numeric(dataset$variable)
```

→ factor to numeric

```
(1) dataset$variable <- as.factor(dataset$variable)
(2) levels(dataset$variable) <- c("attribute1", "attribute1")
```

→ numeric to factor

Creating new datasets (is equal, is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value",]
new_dataset <- dataset[dataset$variable != "value",]
new_dataset <- dataset[dataset$variable > value,]
```

General command structure

```
command1~x, z)
```

Frequencies

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
fvstats(~variable, data = dataset)
```

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Summary

```
summary(modelkey)
```

Shapiro-Wilk test

```
shapiro.test(dataset$variable)
```

Mann-Whitney & Wilcoxon test

```
wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

Correlation

```
cor.test(variable1~variable2, data = dataset)
```

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

Principal Component analysis (package: psych)

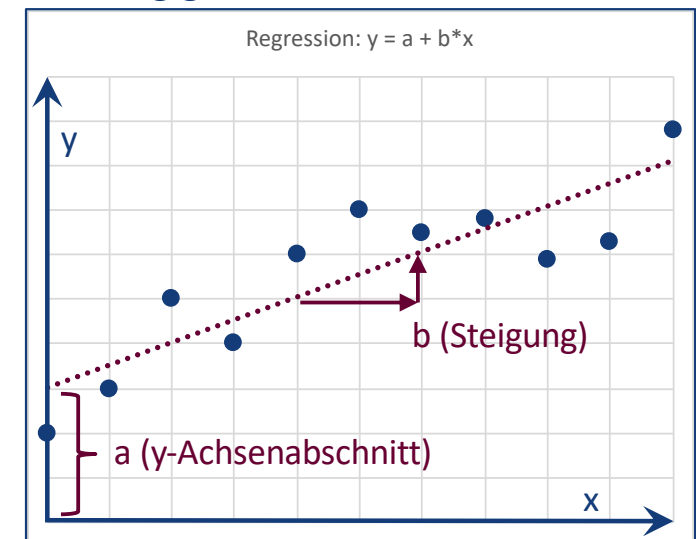
```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

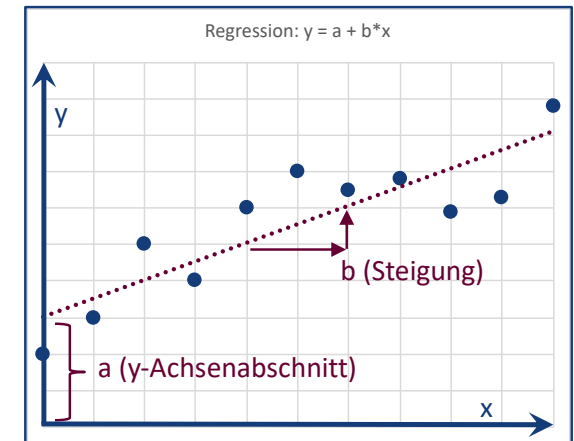
1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

- **Verwendung:**
 - Die lineare Regression untersucht den linearen Einfluss einer metrischen Variable (Einfachregression) oder mehrerer metrischer Variablen (Multiple Regression) auf eine abhängige metrische Variable.
- **Fragestellung:**
 - Inwieweit kann der Wert einer abhängigen Variablen durch die Werte einer (Einfachregression) oder mehrere (Multiple Regression) unabhängigen Variablen erklärt werden?
- **Regressionsgerade (Ergebnis der Analyse):**
 - $y = a + b \cdot x$
 - y: Abhängige Variable („Wirkung“)
 - x: Unabhängige Variable („Ursache“)
 - a: y-Achsenabschnitt
 - b: Regressionskoeffizient („um wieviel steigt y, wenn x sich um eine Einheit erhöht“)
- **Statistische Analyse („im Hintergrund“):**
 - Im Rahmen der Regressionsanalyse wird die Gerade ermittelt, bei der die quadrierten Abweichungen der gemessenen Werte von der Geraden am geringsten ist (Methode der kleinsten Quadrate)



- Vorgehen:**

- 1) Sind die p-values < 0.05 (< 0.01 , < 0.001)?
 - Liegt ein signifikanter Einfluss von x_i auf y vor (p-values)?
- 2) Wie hoch sind die Regressionskoeffizienten (b)?
 - Wie stark ist der Einfluss von x_i auf y ?
- 3) Welchen Wert hat das Bestimmtheitsmaß (R^2)?
 - Wie hoch ist der Erklärungsgehalt des Modells?
 - R^2 = Anteil der Varianz der abhängigen Variable, den die unabhängigen Variablen erklären



```
Call:
lm(formula = Gesamtprogramm ~ Moderation, data = radio)

Residuals:
    Min       1Q   Median       3Q      Max
-48.169  -6.168   1.293   8.611  50.672

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.32790    2.26833   21.75  <2e-16 ***
Moderation    0.42062    0.03162   13.30  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.36 on 273 degrees of freedom
Multiple R-squared:  0.3933, Adjusted R-squared:  0.3911
F-statistic: 177 on 1 and 273 DF, p-value: < 2.2e-16
```

1) p-value

2) Regressionskoeffizient

3) Bestimmtheitsmaß (R^2)

Regression: Übungsaufgaben & R-Befehle

Aufgaben:

- Welchen Einfluss hat die Musik auf das Gesamtprogramm?
- Welchen Einfluss hat die Comedy auf das Gesamtprogramm?
- Welchen Einfluss haben die fünf Programmelemente auf das Gesamtprogramm?

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)
cronbach(factorX)



- Stammsender
- SenderCluster

Variablen

- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

- **Berichterstattung**
- **Musik**
- **Moderation**
- **Comedy**
- **Gewinnspiele**
- **Gesamtprogramm**
- Geschlecht
- Alter
- Altersgruppen
- Aktuelle_Musik_A

1 Einleitung	
2 Installation & Grundstruktur von R Studio	
2.1 Installation von R, R Studio & relevanten Paketen	
2.2 Grundstruktur von R Studio	
3 R-Befehle & Übungsdaten	
3.1 R-Befehle	
3.2 Übungsdaten	
4 Datenmanagement	
4.1 Datenvorbereitung & Laden von Daten	
4.2 Daten bearbeiten in R	
5 Deskriptive Statistik	
5.1 Skalenniveaus	
5.2 Häufigkeiten	
5.3 Lageparameter & Streuungsmaße	
	6 Inferenzstatistische Verfahren
	6.1 Überblick & Systematisierung der Verfahren
	6.2 Chi-Quadrat-Test
	6.3 t-Test
	6.4 ANOVA
	6.5 Shapiro-Wilk-Test
	6.6 Wilcoxon-Test
	6.7 Korrelation
	6.8 Regression
	6.9 Hauptkomponentenanalyse
	7 Graphiken & Export von Ergebnissen nach Excel
	7.1 Graphiken in R
	7.2 Export von Ergebnissen nach Excel

Hauptkomponentenanalyse (& Cronbachs α): Steckbrief

- **Anwendung:**
 - Die Hauptkomponentenanalyse wird verwendet, um Fragen/Variablen/Indikatoren zu wenigen Dimensionen zusammenzufassen.
- **Fragestellung:**
 - Können multidimensionale metrische Daten auf wenige wichtige Hauptkomponenten verdichtet werden?
- **Grundidee:**
 - Die Items einer Komponente korrelieren hoch miteinander.
 - Die Items unterschiedlicher Komponenten korrelieren nicht hoch miteinander.

Indikatoren (Items) der Mitarbeiterbindung	Ladungen (>0,5) der Items auf die Hauptkomponenten der Bindung				Zuverlässigkeit der Skalen		
	affektiv	kognitiv	normativ	vertraglich	Cronbachs Alpha	Composite Reliability	AVE
Ich empfinde es als angenehm, bei meinem Arbeitgeber zu arbeiten.	0,76						
Ich fühle mich mit meinem Arbeitgeber persönlich verbunden.	0,81						
Ich fände es persönlich schade, wenn die Beschäftigung bei meinem Arbeitgeber beendet würde.	0,82						0,62
Ich kann mich mit meinem Arbeitgeber und den Produkten/Dienstleistungen identifizieren.	0,81						
Meine persönlichen Kontakte zu meinem Arbeitsumfeld sind für mich von Bedeutung.	0,61						
In gewisser Weise bindet mich der für einen Wechsel benötigte Zeitaufwand an meinem Arbeitgeber.		0,79					
Ich bin auf meinen Arbeitgeber angewiesen, weil es zurzeit keine gleichwertigen Alternativen am Markt gibt.		0,71					
Ich empfinde eine Bindung an meinen Arbeitgeber, weil bei einem Wechsel der von mir investierte Aufwand an Wert verlieren würde.		0,71					
Ich fühle mich an meinen Arbeitgeber gebunden, weil ein Wechsel mit Wechselkosten einhergehen würde.		0,67					
Es wäre nicht fair, die Beziehung zu meinem Arbeitgeber aufzukündigen, weil er sich stets um mich als Arbeitnehmer bemüht hat.			0,63				
Aufgrund der langen Beziehung mit meinem Arbeitgeber fühle ich mich zu einer gewissen Rücksichtnahme verpflichtet.			0,76		0,80	0,83	0,55
Ich fühle mich in der Angestelltenbeziehung mit dem Arbeitgeber zur Fairness verpflichtet.			0,81				
Moralische Verpflichtungen gegenüber dem Arbeitgeber spielen für mich auch eine Rolle.			0,82				
Ich bin vertraglich an meinen Arbeitgeber gebunden.				0,93	1	1	1

hohe Korrelation

geringe Korrelation

Quelle: Gansser, O. & Godbersen, H. (2017). Mitarbeiterbindung durch Betriebliches Gesundheitsmanagement. Ergebnisse einer empirischen Studie und Leitfaden für die Praxis. zfo Zeitschrift Führung + Organisation, Jg. 86 (2), 108-116.

Hauptkomponentenanalyse (& Cronbachs α): Steckbrief

Prof. Dr.
Godbersen

- Vorgehen

- 0) Vorbereiten der Datenbasis (neuer Datensatz)
- 1) Prüfung, ob eine Hauptkomponentenanalyse durchgeführt werden kann/soll mittels KMO (Kaiser-Meyer-Olkin-Test (misst den Anteil der Varianz zwischen Variablen, der auf gemeinsame Varianz zurückzuführen ist))
 - Auswahl: So viele Komponenten wie Eigenwerte > 1
- 2) Bestimmung der Anzahl der Komponenten anhand der Eigenwerte
- 3) Bestimmung der Faktorladungen (Korrelation eines Items mit einer Komponente) & Zuordnung der Items zu Komponenten
 - Voraussetzung: Ladung > 0.5
 - Entscheidung: Zuordnung des Items zu der Komponente, auf der es am höchsten lädt
- 4) Interpretation der Komponenten
- 5) Testen der einzelnen Komponenten mittels Cronbachs α (Reliabilität; durchschnittliche Korrelation aller Split-half-tests))

KMO – Interpretation:

- 0.00 to 0.49 - unacceptable
- 0.50 to 0.59 - miserable
- **0.60** to 0.69 - mediocre
- **0.70** to 0.79 - middling
- 0.80 to 0.89 - meritorious
- 0.90 to 1.00 - marvellous

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = pca_radio_data)
Overall MSA = 0.85

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	2.359856	1.2124941	1.16061942	1.00739666	0.98060991	0.84432456	0.78977668	0.73846855	0.63632144
Proportion of Variance	0.397780	0.1050101	0.09621696	0.07248915	0.06868541	0.05092028	0.04455337	0.03895256	0.02892178
Cumulative Proportion	0.397780	0.5027901	0.59900708	0.67149623	0.74018164	0.79110192	0.83565530	0.87460785	0.90352964

	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14
Standard deviation	0.63167541	0.51744946	0.50533062	0.48631269	0.43811269
Proportion of Variance	0.02850099	0.01912528	0.01823993	0.01689286	0.01311269
Cumulative Proportion	0.93203063	0.95115591	0.96939584	0.98628870	1.00000000

4 Komponenten

Principal Components Analysis
Call: principal(r = pca_radio_data, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix

	RC1	RC4	RC3	RC2	h2	u2	com
Verlaesslicher_Service	0.16	0.78	0.18	-0.10	0.67	0.33	1.2
Informative_Nachrichten	0.12	0.90	0.03	0.02	0.83	0.17	1.0
Unterhaltsame_Nachrichten	0.48	0.47	0.04	0.35	0.58	0.42	2.8
Aktuelle_Musik	0.19	-0.02	-0.02	0.77	0.62	0.38	1.1
Alte_Musik	0.29	0.16	0.65	-0.12	0.55	0.45	1.1

Zuordnung der Items

Cronbachs α – akzeptierbare Werte:

- > 0.6 ("weiches" Kriterium)
- > 0.7 ("hartes" Kriterium)

```
> cronbach(comp.Unterhaltung)
$sample.size
[1] 275

$number.of.items
[1] 5

$alpha
[1] 0.8511712
```

Hauptkomponentenanalyse (& Cronbachs α): Übungsaufgaben & R-Befehle

Aufgaben:

- Zu welchen Komponenten lassen sich die Programmeigenschaften zusammenfassen? („Wie können aus den Programmeigenschaften neue übergeordnete Variablen gebildet werden?“)
- Wie ist die Reliabilität der neuen Komponenten zu beurteilen?

Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX)                plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")
```

Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2...)
cronbach(factorX)
```

- Stammsender
- SenderCluster

Variablen

- **Verlaesslicher_Service**
 - **Informative_Nachrichten**
 - **Unterhaltsame_Nachrichten**
 - **Aktuelle_Musik**
 - **Alte_Musik**
 - **Anregende_Musik**
 - **Beruhigende_Musik**
 - **Informative_Moderation**
 - **Unterhaltsame_Moderation**
 - **Sympathische_Stimme**
 - **Anspruchsvolle_Comedy**
 - **Lustige_Comedy**
 - **Spannende_Gewinnspiele**
 - **Attraktive_Preise**
- Berichterstattung
 - Musik
 - Moderation
 - Comedy
 - Gewinnspiele
 - Gesamtprogramm
 - Geschlecht
 - Alter
 - Altersgruppen
 - Aktuelle_Musik_A

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

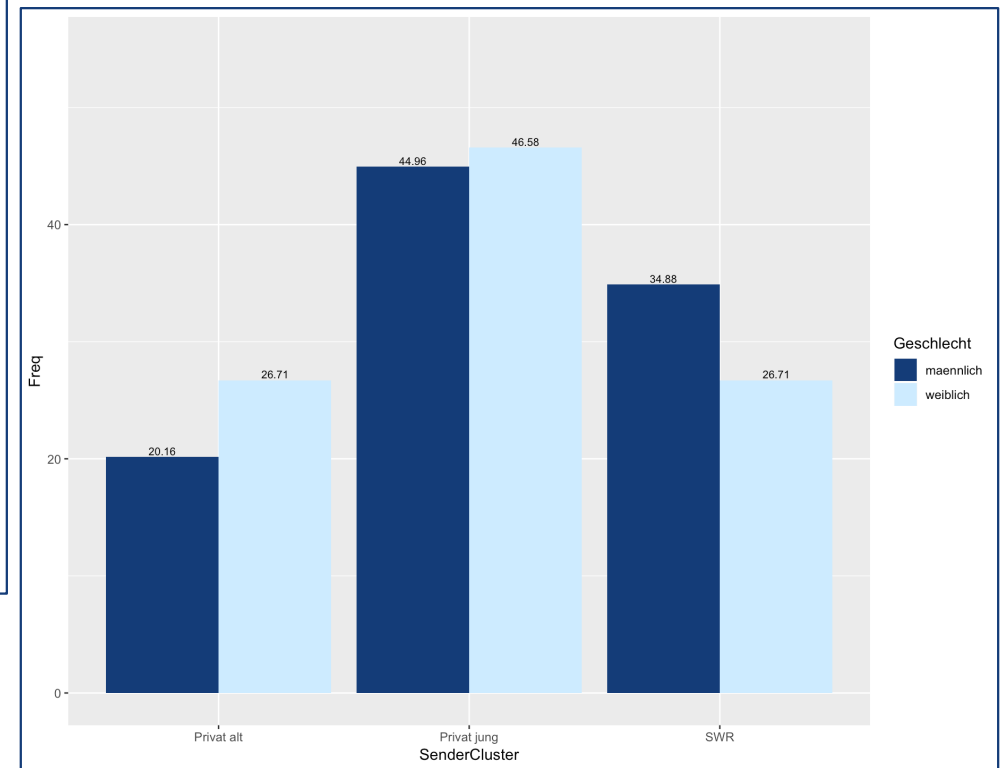
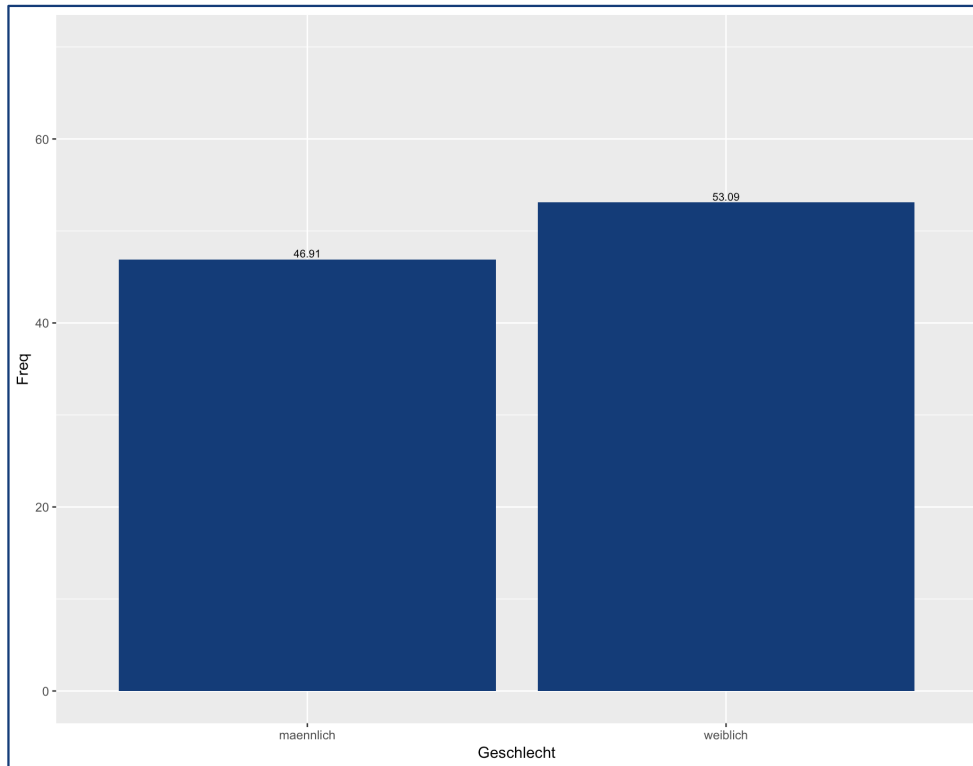
6.8 Regression

6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel



Graphiken in R: Übungsaufgaben & R-Befehle

Aufgaben:

- Erstellen Sie ein Säulendiagramm, das die relativen Häufigkeiten (%) der Variable Geschlecht wiedergibt.
- Erstellen Sie ein Säulendiagramm, dass die relativen Häufigkeiten (%) der Variable SenderCluster in Abhängigkeit vom Geschlecht wiedergibt.

- Stammsender
- **SenderCluster**
- Verlaesslicher_Service
- Informative Nachrichten

Variablen

Bar graph – frequencies with 1 variable

```
table <- tally(~variable, format = "proportion", data = dataset)
```

```
frame <- as.data.frame(table)
```

```
View(frame)
```

```
ggplot(frame, aes(variable, Freq)) + geom_bar(position = "dodge", stat = "identity", fill = "#143C78") +  
scale_y_continuous(limits = c(min, max)) + geom_text(aes(label = sprintf("%0.2f", frame$Freq)), position =  
position_dodge(width = 0.9), vjust = -0.25, colour = "black", size = 2.7)
```

- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele

Bar graph – frequencies with 2 variables

```
table <- tally(variable_1~variable_2, format= "proportion", data = dataset)
```

```
frame <- as.data.frame(table)
```

```
View(frame)
```

```
ggplot(frame, aes(variable_1, Freq, fill = variable_2)) + geom_bar(position = "dodge", stat = "identity") +  
scale_fill_manual(values = c("#143C78", "#CDEBFF")) + scale_y_continuous(limits = c(min, max)) + geom_text(aes(label =  
sprintf("%0.2f", frame$Freq)), position = position_dodge(width = 0.9), vjust = -0.25, colour = "black", size = 2.7)
```

- Gesamtprogramm
- **Geschlecht**
- Alter
- Altersgruppen
- Aktuelle_Musik_A

1 Einleitung

2 Installation & Grundstruktur von R Studio

2.1 Installation von R, R Studio & relevanten Paketen

2.2 Grundstruktur von R Studio

3 R-Befehle & Übungsdaten

3.1 R-Befehle

3.2 Übungsdaten

4 Datenmanagement

4.1 Datenvorbereitung & Laden von Daten

4.2 Daten bearbeiten in R

5 Deskriptive Statistik

5.1 Skalenniveaus

5.2 Häufigkeiten

5.3 Lageparameter & Streuungsmaße

6 Inferenzstatistische Verfahren

6.1 Überblick & Systematisierung der Verfahren

6.2 Chi-Quadrat-Test

6.3 t-Test

6.4 ANOVA

6.5 Shapiro-Wilk-Test

6.6 Wilcoxon-Test

6.7 Korrelation

6.8 Regression

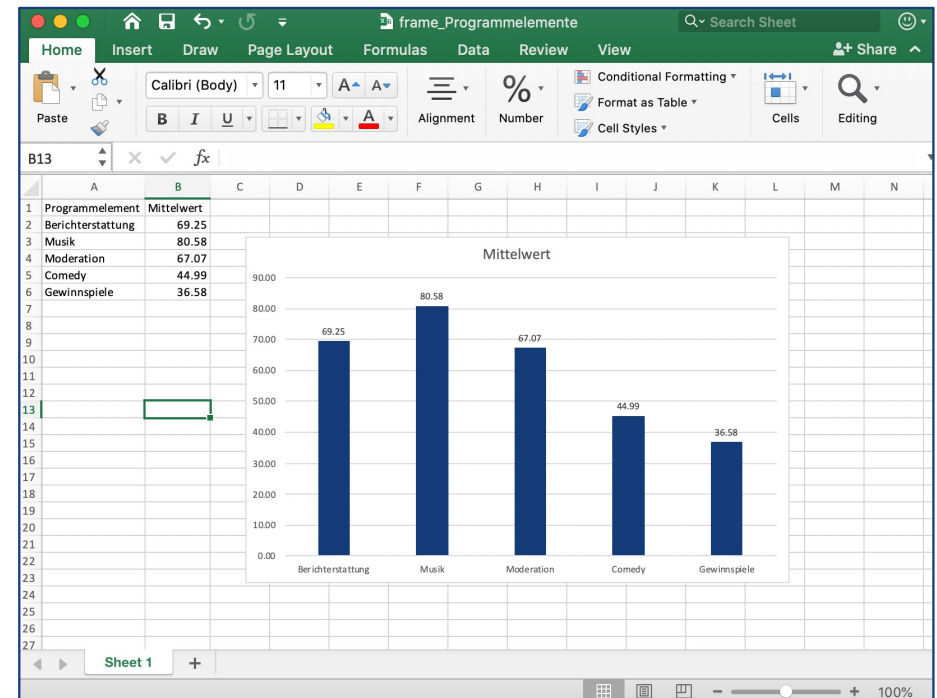
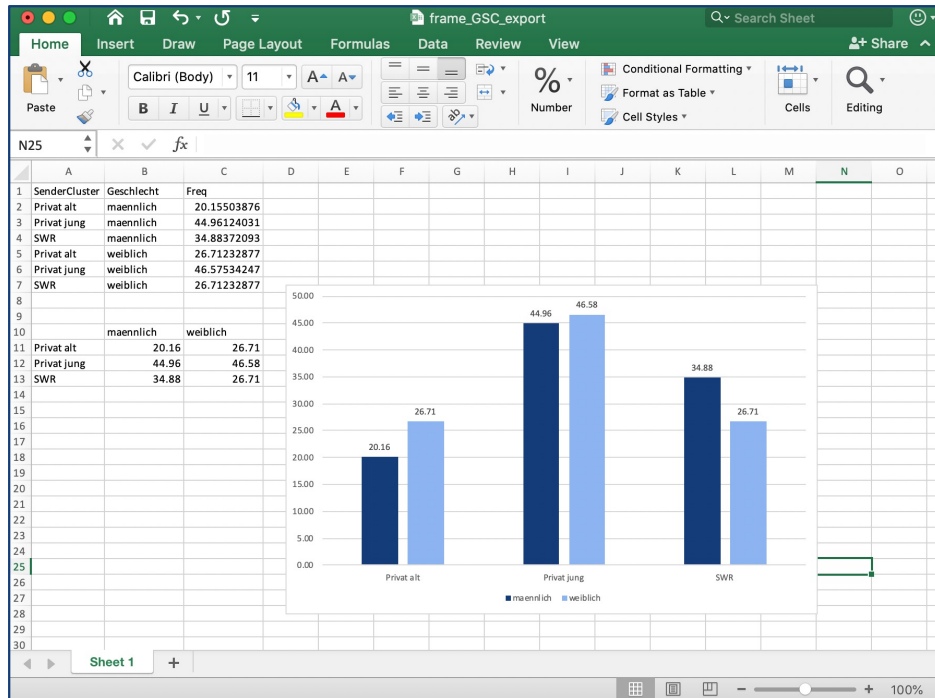
6.9 Hauptkomponentenanalyse

7 Graphiken & Export von Ergebnissen nach Excel

7.1 Graphiken in R

7.2 Export von Ergebnissen nach Excel

Export von Ergebnissen nach Excel: Ergebnisse



Export von Ergebnissen nach Excel: Übungsaufgaben & R-Befehle

Aufgaben:

- Exportieren Sie die relativen Häufigkeiten (%) der Variable SenderCluster in Abhängigkeit vom Geschlecht nach Excel.
- Exportieren Sie die Mittelwerte der Bewertungen der Programmelemente nach Excel.

Export R results to Excel (package: openxlsx)

(1.1) Preparing frequencies

```
frame <- data.frame(tally(variable_1~variable_1, data = dataset))
```

(1.2) Preparing mean, median etc.

```
frame <- data.frame(new_variable_name_1 =  
c(mean(~variable_1, data = dataset), mean(~variable_2, data =  
dataset)), new_variable_name_2 = c(mean(~variable_3, data =  
dataset), mean(~variable_4, data = dataset)))
```

(2) Export to Excel

```
write.xlsx(frame, "Desktop/R_operations/name.xlsx", asTable =  
FALSE)
```

→ Please note: The name of the R frame & the directory / name of the xlsx-file have to be adjusted

- | | Variablen |
|-----------------------------|----------------------------|
| • Stammsender | |
| • SenderCluster | |
| • Verlaesslicher_Service | |
| • Informative_Nachrichten | |
| • Unterhaltsame_Nachrichten | |
| • Aktuelle_Musik | |
| • Alte_Musik | • Berichterstattung |
| • Anregende_Musik | • Musik |
| • Beruhigende_Musik | • Moderation |
| • Informative_Moderation | • Comedy |
| • Unterhaltsame_Moderation | • Gewinnspiele |
| • Sympathische_Stimme | |
| • Anspruchsvolle_Comedy | • Gesamtprogramm |
| • Lustige_Comedy | |
| • Spannende_Gewinnspiele | • Geschlecht |
| • Attraktive_Preise | • Alter |
| | • Altersgruppen |
| | • Aktuelle_Musik_A |

Prof. Dr. Hendrik Godbersen

www.godbersen.online