Einführung in die Datenanalyse mit R: Datenaufbereitung

Marco Wähner


Gebäude 37.03.03.14


marco.waehner@hhu.de


Sprechstunde nach Vereinbarung

Kurze Wiederholung

R in a Nutshell

- R speichert alles in Objekten
  - Objekte haben Klassen und Formate
- Funktionen
  - Automatisierung bestimmter Aufgaben
  - Benötigen eine Eingabe und ggf. weitere Argumente
- Packages und Base R
  - Bestimmte Funktionen werden durch Base R zur Verfügung gestellt
  - Funktionen werden über Packages erweitert
- RProjects
  - Organisieren den Workflow

Ablauf

Freitag, 8.04.2022	
Zeit	**Inhalt**
9:00 – 10:30	Deskriptivstatistik I
10:30 – 10:45	Pause
10:45 – 12:15	Deskriptivstatistik II / R Markdown
12:15 – 13:30	Pause
13:30 – 15:00	Datenvisualisierung I
15:00 – 15:30	Pause
15:30 – 17:00	Inferenzstatistik I

Deskriptivstatistik I (Kapitel 5)

Deskriptivstatistik

- Häufigkeitsverteilungen
- Maße der zentralen Tendenz
- Streuungsmaße

Deskriptivstatistik: Häufigkeiten

- Absolute Häufigkeit

```
# Häufigkeitsverteilung Ost/West
table(allbus2018$eastwest)
##
##    1    2
## 2387 1090
```

- Relative Häufigkeit

```
prop.table(table(allbus2018$eastwest))
##
##         1         2
## 0.6865114 0.3134886
```

- Prozentuale Häufigkeit

```
prop.table(table(allbus2018$eastwest))*100
##
##        1        2
## 68.65114 31.34886
```

Deskriptivstatistik: Maße der zentralen Tendenz

- Arithmetische Mittel:
  - Das arithmetische Mittel kennzeichnet den “Schwerpunkt einer Verteilung” (Diaz-Bone, 2019, S. 45)
  - Vergleich na.rm=TRUE und na.rm=FALSE

```
mean(allbus2018$age, na.rm = TRUE)
## [1] 51.67713
```

- Median:
  - „[Der Median]unterteilt die Reihe in zwei Hälften: die eine Hälfte der Ausprägungen ist kleiner als (oder höchstens gleich groß wie) der Median, die andere Hälfte der Ausprägungen ist größer als (oder zumindest gleich groß wie) der Median“ (Diaz-Bone, 2019, S. 45-46)

```
median(allbus2018$age, na.rm = TRUE)
## [1] 53
```

Deskriptivstatistik: Maße der zentralen Tendenz

- Modus:
  - Die häufigste Ausprägung einer mindestens nominalskalierten Variable ist der Modus. Der Modus ist ein „typischer“ Wert für eine Verteilung

```
modal_tabelle <- table(allbus2018$age)
names(modal_tabelle)[which(modal_tabelle==max(modal_tabelle))]
## [1] "55"
```

Streuungsmaße

- Varianz
  - Maß für die Streuung einer Verteilung um ihren Mittelwert.

```
var(allbus$age, na.rm = TRUE)
## [1] 311.2478
```

- Standardabweichung
  - Wurzel aus der Varianz

```
sd(allbus$age, na.rm = TRUE)
## [1] 17.64222
```

- R-Package „summarytools“
  - Einfacher und schneller Überblick zur Verteilung einer Variablen
  - Zur Dokumentation
  - Funktionen zur Häufigkeitsverteilung und univariaten Statistik

```
#Häufigkeitsverteilung: Ost/West
freq(df$eastwest)
```

```
## Frequencies
## df$eastwest
## Label: ERHEBUNGSGEBIET (WOHNGEBIET): WEST - OST
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1   2387     68.65          68.65     68.65          68.65
##           2   1090     31.35         100.00     31.35         100.00
##        <NA>      0                               0.00         100.00
##       Total   3477    100.00         100.00    100.00         100.00
```

Praxis

Hands-On

Aufgabe (Lösung)

Kontakt

Vielen Dank!

Marco Wähner

 Gebäude 37.03.03.14

 marco.waehner@hhu.de

 Sprechstunde nach Vereinbarung