

Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2024/25

Belinda Fleischmann

Datum	Einheit	Thema	Form
15.10.24	R Grundlagen	(1) Einführung	Seminar
22.10.24	R Grundlagen	(2) R und Visual Studio Code	Seminar
29.10.24	R Grundlagen	(2) R und Visual Studio Code	Übung
05.11.24	R Grundlagen	(3) Vektoren, (4) Matrizen	Seminar
12.11.24	R Grundlagen	(5) Listen und Dataframes	Seminar
	Leistungsnachweis 1		
19.11.24	R Grundlagen	(6) Datenmanagement	Seminar
26.11.24	R Grundlagen	(2)-(6) R Grundlagen	Übung
03.12.24	Deskriptive Statistik	(7) Häufigkeitsverteilungen	Seminar
10.12.24	Deskriptive Statistik	(8) Verteilungsfunktionen und Quantile	Seminar
	Leistungsnachweis 2		
17.12.24	Deskriptive Statistik	(9) Maße der zentralen Tendenz und Datenvariabilität	Seminar
	Weihnachtspause		
07.01.25	R Grundlagen	(10) Strukturiertes Programmieren: Kontrollfluss, Debugging	Seminar
14.01.25	Deskriptive Statistik	(11) Anwendungsbeispiel	Übung
	Leistungsnachweis 3		
21.01.25	Deskriptive Statistik	(11) Anwendungsbeispiel	Seminar
28.01.25	Deskriptive Statistik	(11) Anwendungsbeispiel, Q&A	Seminar

(11) Anwendungsbeispiel

(Foliensatz mit Programmierübungen)

Beispieldatensatz

Datenvorverarbeitung

Deskriptive Statistiken

Visualisierung

Beispieldatensatz

 ${\sf Datenvor verar beitung}$

Deskriptive Statistiken

Visualisierung

Forschungsfrage

Evidenzbasierte Evaluation von Psychotherapieformen bei Depression

Welche Therapieform ist bei Depression wirksamer?

Online Psychotherapie



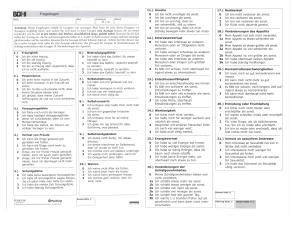
Klassische Psychotherapie



Beispiel: Evaluation von Psychotherapieformen bei Depression



Becks Depressions-Inventar (BDI) zur Depressionsdiagnostik



- 0 8 keine Depression
- 9 13 minimale Depression
- 14 19 leichte Depression
- 20 28 mittelschwere Depression
- 29 63 schwere Depression

Beispieldatensatz

Einlesen des Datensatzes mit read.table()

```
file_path <- file.path(data_dir_path, "psychotherapie_datensatz.csv")
# file_path könnte beispielsweise so aussehen:
# "/home/username/uni/progr-und-deskr-stat-24/Daten/psychotherapie_datensatz.csv"
D <- read.table(file_path, sep = ",", header = TRUE)</pre>
```

Daten der ersten acht Proband:innen jeder Gruppe

	Bedingung	Pre.BDI	Post.BDI
1	Klassisch	17	9
2	Klassisch	20	14
3	Klassisch	16	13
4	Klassisch	18	12
5	Klassisch	21	12
6	Klassisch	17	14
7	Klassisch	17	12
8	Klassisch	17	9
51	Online	22	16
52	Online	19	15
53	Online	21	13
54	Online	18	15
55	Online	19	13
56	Online	17	16
57	Online	20	13
58	Online	19	16

Bonus: Datensimulation

```
# Seed setzen
set.seed(5)
                                                  # Startwert für den Zufallsgenerator setzen
# Simulationsparameter
      <- 50
                                                  # Proband:innnen pro Gruppe
mu <- c(
                                                  # Erwartungswertparameter
 18. 12.
                                                  # Pre und Post der Gruppe Klassisch
 19, 14)
                                                  # Pre und Post der Gruppe Online
sigsqr <- 3
                                                  # Varianzparameter (gleich für alle Gruppen)
# Datensimulation
D <- data.frame(
        "Bedingung" = c(
         rep("Klassisch", n), rep("Online", n)), # n-mal "Klassisch", n-mal "Online"
        "Pre BDT" = c(
         round(rnorm(n, mu[1], sqrt(sigsqr))), # n Zufallswerte aus Normalveritung mit mu[1]
         round(rnorm(n, mu[3], sgrt(sigsgr)))), # n Zufallswerte aus Normalveritung mit mu[3]
        "Post BDT" = c(
         round(rnorm(n, mu[2], sqrt(sigsqr))), # n Zufallswerte aus Normalveritung mit mu[2]
         round(rnorm(n, mu[4], sqrt(sigsqr))))
                                                # n Zufallswerte aus Normalveritung mit mu[4]
# Datenspeicherung
fname <- file.path(data_path, "psychotherapie_datensatz.csv")</pre>
write.csv(D, file = fname)
```

Beispieldatensatz

Datenvorverarbeitung

Deskriptive Statistiken

Visualisierung

Datenvorverarbeitung

Überlegungen für die Datenvorverarbeitung

- Studienfokus ist die Veränderung der Depressionsymptomatik durch Therapieformen.
- Für jede Proband:in ergibt sich diese Veränderung als Differenz zwischen Post.BDI und Pre.BDI.
- Eine Reduktion der Depressionssymptomatik ergibt dabei einen negativen Wert.
- Es ist intuitiver, Verbesserungen mit positiven Zahlen zu repräsentieren.
- Als Quantifizierung des Therapieeffekts bei Proband:in i bietet sich also folgendes Maß an

$$\Delta \mathsf{BDI}[\mathsf{i}] := -(\mathsf{Post.BDI}[\mathsf{i}] - \mathsf{Pre.BDI}[\mathsf{i}]) \tag{1}$$

ullet Wir betrachten in der Folge also das ΔBDI Maß mit folgenden Interpretationen

$\Delta \mathrm{BDI} > 0$	Verminderung der Depressionsymptomatik	Wirksame Therapie
$\Delta \mathrm{BDI} = 0$	Keine Veränderung der Depressionsymptomatik	Wirkungslose Therapie
$\Delta \mathrm{BDI} < 0$	Verstärkung der Depressionsymptomatik	Schädigende Therapie

Programmierübung 1) Datenvorverarbeitung

Aufgabe: Berechnung einer neuen Variablen im Datensatz

Füge dem Datensatz eine neue Spalte Delta.BDI hinzu, die den Unterschied Δ zwischen dem Präund Post-Wert darstellt. Nutze dafür die entsprechenden Spalten Pre.BDI und Post.BDI. Achte darauf, dass das Delta so berechnet wird, dass ein positiver Wert eine Verbesserung (Reduktion der Depressionswerte) widerspiegelt.

Nachfolgend wird am Beispiel der Daten der ersten acht Proband:innen gezeigt, wie das Ergebnis in etwa aussehen sollte.

	Bedingung	Pre.BDI	Post.BDI	Delta.BDI
1	Klassisch	17	9	8
2	Klassisch	20	14	6
3	Klassisch	16	13	3
4	Klassisch	18	12	6
5	Klassisch	21	12	9
6	Klassisch	17	14	3
7	Klassisch	17	12	5
8	Klassisch	17	9	8
51	Online	22	16	6
52	Online	19	15	4
53	Online	21	13	8
54	Online	18	15	3
55	Online	19	13	6
56	Online	17	16	1
57	Online	20	13	7
58	Online	19	16	3

Beispieldatensatz

 ${\sf Datenvor verar beitung}$

Deskriptive Statistiken

Visualisierung

Programmierübung 2) Bedingungsunabhängige Deskriptive Statistiken

Aufgabe 2a): Berechnung bedingungsunabhängiger deskriptiver Statistiken

Berechne die deskriptiven Statistiken für die Variable Delta.BDI. Erstelle hierfür einen neuen Dataframe, der die folgenden statistischen Kennzahlen enthält: die Stichprobengröße, das Maximum, das Minimum, den Median, den Mittelwert, die Varianz und die Standardabweichung. Der resultierende Dataframe sollte in etwa so aussehen:

```
n Max Min Median Mean Var Std
1 100 12 -1 6 5.54 5.826667 2.413849
```

Programmierübung 2) Bedingungsabhängige Deskriptive Statistiken

Aufgabe b): Berechnung bedingungsabhängiger deskriptiver Statistiken

Berechne die deskriptiven Statistiken für die Variable Delta.BDI, getrennt nach den zwei Therapiebedingungen Klassisch und Online. Erstelle dazu einen Dataframe deskr_stat mit zwei Zeilen (jeweils eine pro Bedingung) und den statistischen Kennzahlen als Spalten. Gehe wie folgt vor.

- 1. Lege eine Variable th_bed an, die die Namen der beiden Therapiebedingungen enthält.
- Initialisiere einen leeren Dataframe mit den genannten Kennzahlen als Spalten und den Therapiebedingungen als Zeilen.
- 3. Verwende eine for-Schleife, um über th bed zu iterieren.
- 4. Filter in jeder Iteration die Daten für die entsprechende Bedingung und berechne die Statistiken für Delta.BDI und trage die berechneten Werte in den entsprechenden Zeilen des Dataframes ein.

Der resultierende Dataframe sollte in etwa so aussehen:

```
        n Max
        Min Median Mean
        Var
        Std

        Klassisch
        50
        12
        -1
        6
        6.16
        7.075918
        2.660060

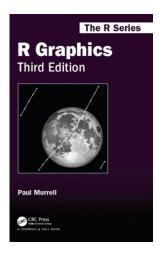
        Online
        50
        9
        1
        5
        4.92
        3.911837
        1.977836
```

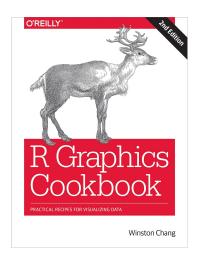
Beispieldatensatz

 ${\sf Datenvor verar beitung}$

Deskriptive Statistiken

Visualisierung





Murrell (2019)

Online-Buch, Chang (2013)

R Funktionalitäten für Abbildungen

Base Graphics

- Erstellung und bedarfsgerechte Anpassung von Abbildungen
- Eher low-level, fine tuning orientiert

Lattice und ggplot2

- Erstellung und bedarfsgerechte Anpassung von Abbildungen
- Eher high level, an der eigenen Philosophie orientiert

Base Graphics, lattice und ggplot2 können ähnliche Abbildungen generieren

LaTeX Typesetting ist in allen Paketen unterentwickelt

R Funktionalitäten für Abbildungen

Base Graphics

- Erstellung und bedarfsgerechte Anpassung von Abbildungen
- Eher low-level, fine tuning orientiert

Lattice und ggplot2

- Erstellung und bedarfsgerechte Anpassung von Abbildungen
- Eher high level, an der eigenen Philosophie orientiert

Base Graphics, lattice und ggplot2 können ähnliche Abbildungen generieren

LaTeX Typesetting ist in allen Paketen unterentwickelt

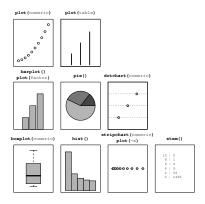


Figure 2.5 High-level base graphics plotting functions for producing plots of a single variable. Where the function can be used to produce more than one type of plot, the relevant data type is shown (in gray). For example, plot(mumeric) means that this is what the plot() produces when it is given a single numeric argument.

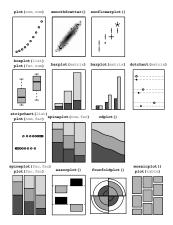


Figure 2.6 High-level base graphics plotting functions for producing plots of two variables. Where the function can be used to produce more than one type of plot, the relevant data type is shown (in gray). For example plot(um, #cz) represents calling the plot() function with a numeric vector as the first argument and a factor as the second argument.

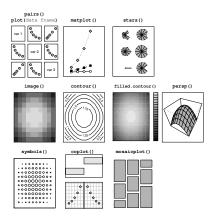


Figure 2.7 High-level base graphics plotting functions for producing plots of many variables. Where the function can be used to produce more than one type of plot, the relevant data type is shown (in gray).

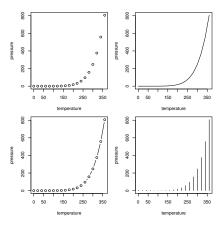


Figure 2.2
Four variations on a scatterplot. In each case, the plot is produced by a call to the plot() function with the same data, all that changes is the value of the type argument. At top-left, type="p" to give points (data symbols), at top-right, type="1" to give lines, at bottom-left, type="b" to give both, and at bottom-right, type="h" to give histogram-like vertical lines.

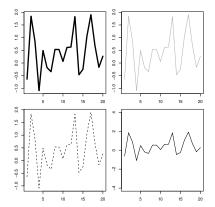


Figure 2.9
Standard arguments for high-level functions. All four plots are produced by calls to the plot() function with the same data, but with different standard plot function arguments specified: the top-left plot makes use of the lud argument to control line thickness; the top-left plot to sets the col argument to control line top-left plot makes use of the luty argument to control line type; and the bottom-right plot uses the vilu argument to control line type; and the bottom-right plot uses the vilu argument to control the scale on the v-axis.

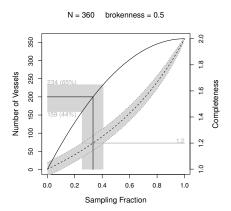


Figure 1.3
A customized scatterplot produced using R. This is created by starting with a simple scatterplot and augmenting it by adding an additional y-axis and several additional sets of lines, polygons, and text labels.

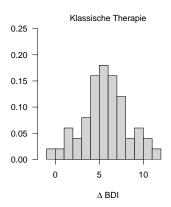
Code Outline

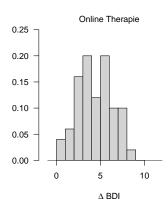
```
# Initialisierung einer neuen Abbildung
dev.new()
# Abbildungsparameter
par(
 z.B. Arrangement von Panels, Begrenzungsstile, Schriftfonts, etc
# Higher-level Abbildungsfunktion wie plot(), hist(), barplot(), ...
plot(
 z.B. x- und v-Daten, Achsenlimits, Achsenbeschriftungen, Titel, Farben, etc.
 Jeder Aufruf einer higher-level Graphikfunktion belegt ein neues Subpanel!
# Hinzufügen weiterer Daten mit lower-level Abbildungsfunktionen zum aktuellen Panel
z.B. points(), lines(), abline()
# Weitere Graphikannotation zu aktuellem Panel
z.B. legend(), text()
# Speichern der Abbildung (Größenverhältnisse erst hier final festgelegt)
z.B. dev.copy2pdf()
```

Histogramme

```
# Histogrammparameter
                                                                 # gewünschte Klassenbreite
h
                                                                 # b_0
b 0
         <- min(D$Delta.BDI)
b k
      <- max(D$Delta.BDI)
                                                                 # b 0
         \leftarrow ceiling((b_k - b_0) / h)
                                                                 # Anzahl der Klassen
        \leftarrow seq(b_0, b_k, by = h)
                                                                 # Klassen [b_{j-1}, b_j[
vlimits \leftarrow c(0, .25)
                                                                 # v-Achsenlimits
xlimits \leftarrow c(-2, 14)
                                                                 # x-Achsenlimits
therapie <- c("Klassisch" , "Online")
                                                                 # Therapiebedingungen
       <- c("Klassische Therapie", "Online Therapie")
                                                                 # Abbildungslabel
# Abbildungsparameter
                                                                 # für Details siehe ?par
par(
  mfcol
            = c(1, 2),
                                                                 # 1 x 2 Panelstruktur
                                                                 # Serif-freier Fonttyp
  family
            = "sans".
  pty
            = "m",
                                                                 # Maximale Abbildungsregion
            = "1",
                                                                 # L förmige Box
  btv
  las
                                                                 # Horizontale Achsenbeschriftung
  xaxs
            = "i",
                                                                 # x-Achse bei y = 0
  vaxs
            = "i".
                                                                 # v-Achse bei x = 0
  font.main = 1,
                                                                 # Non-Bold Titel
  cex
         = 1
                                                                 # Textvergrößerungsfaktor
  cex.main = 1
                                                                 # Titeltextvergrößerungsfaktor
# Iteration über Therapiebedingungen
for(i in 1:2){
  hist(
    D$Delta.BDI[D$Bedingung == therapie[i]],
                                                                 # Delta.BDI Werte von Therapiebedingung i
    breaks = b.
                                                                 # Histogrammklassen
                                                                 # normierte relative Häufigkeit
    frea = F.
    xlim = xlimits,
                                                                 # x-Achsenlimits
                                                                 # v-Achsenlimits
    vlim = vlimits.
    xlab = TeX("$\\Delta$ BDI"),
                                                                 # x-Achsenbeschriftung
    ylab = "",
                                                                 # y-Achsenbeschriftung
    main = labs[i])
                                                                 # Titelbeschriftung
# PDF Speicherung
dev.copv2pdf(
              = file.path(abb_dir, "pds_11_histogramm.pdf"),
  file
  width
              = 8,
              = 4
  height
```

Histogramme





Aufgabe 3a): Balkendiagramme

Erstelle ein Balkendiagramm mit Fehlerbalken, das die Mittelwerte und Standardabweichungen für die beiden Therapiebedingungen (Klassisch und Online) der Variable Delta.BDI zeigt. Gehe wie folgt vor.

- 1. Extrahiere die Mittelwerte und Standardabweichungen aus dem Dataframe deskr stat.
- 2. Verwende die Funktion barplot(), um die Mittelwerte als Balken darzustellen.
- Füge Fehlerbalken hinzu, die die Standardabweichungen anzeigen, indem du die Funktion arrows() verwendest.
- 4. Wähle passende Achsenbeschriftungen und einen Titel für das Diagramm.

Als Bearbeitungshilfe seien folgende Befehle gegeben. Beachte, dass eine oder mehrere hierin verwendeten Variablen zuerst noch erstellt werden müssen, und dem barplot()-Befehl noch Funktionsargumente fehlen, um die Aufgabe vollständig zu erfüllen (z.B. Achsenlimits und -beschriftung).

```
x <- barplot(
                                                  # Speichern der der x-Ordinaten (?barplot für Details)
                                                  # Mittelwerte als Balkenhöhe
 height = mw,
arrows(
 χO
        = x,
                                                  # arrow start x-ordinate
                                                 # arrow start y-ordinate
 v0 = mw - sd
                                                  # arrow end x-ordinate
 x1 = x
 v1
       = mw + sd.
                                                 # arrow end y-ordinate
                                                 # Pfeilspitzen beiderseits
 code = 3.
 angle = 90,
                                                 # Pfeilspitzenwinkel -> Linie
 length = 0.05
                                                 # Linielänge
```

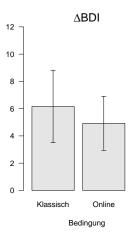
Aufgabe 3 b): Boxplots

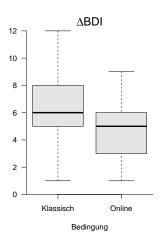
Erstelle Boxplots, die die Verteilung der Delta.BDI-Werte für die beiden Therapiebedingungen Klassisch und Online visualisieren. Verwende dazu die Funktion boxplot() und das Funktionsargument formula, um mit einem einzigen Plot-Befehl beide Boxplots in einem Diagramm nebeneinander darzustellen.

Nutze den Befehl ?boxplot, um die Argumente der Funktion genauer zu verstehen. Schau auch in die **Examples** hinein, um ein besseres Verständnis für die Anwendung der Funktion unter Hinzunahme des Arguments formula zu bekommen.

Der resultierenden Diagramme sollten in etwa so aussehen:

Bedingungsabhängige Visualisierung deskriptiver Statistiken





Literatur

Chang, Winston. 2013. *R Graphics Cookbook*. Beijing Cambridge Farnham Köln Sebastopol Tokyo: O'Reilly. Murrell, Paul. 2019. *R Graphics*. Third edition. The R Series. Boca Raton: CRC Press, Taylor & Francis Group.