



Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2024/25

Belinda Fleischmann

(7) Häufigkeitsverteilungen

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Programmierübungen und Selbstkontrollfragen

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Programmierübungen und Selbstkontrollfragen

Definition und Ziele der Deskriptive Statistik

- Die Deskriptive Statistik ist die *beschreibende* Statistik.
- Ziel der Deskriptiven Statistik ist es, Daten übersichtlich darzustellen.
- Deskriptive Statistik ist insbesondere bei großen Datensätzen sinnvoll.
- Die Deskriptive Statistik berechnet zusammenfassende Maße aus Daten.

Typische Methoden der Deskriptiven Statistik

- Häufigkeitsverteilungen und Histogramme
- Verteilungsfunktionen und Quantile
- Maße der zentralen Tendenz und der Datenvariabilität
- Zusammenhangsmaße

Die Deskriptive Statistik benutzt keine probabilistischen Modelle, aber die Methoden der Deskriptiven Statistik ergeben nur vor dem Hintergrund probabilistischer Modelle Sinn.

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

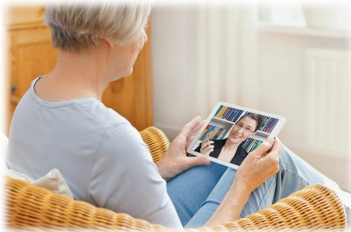
Histogramme

Programmierübungen und Selbstkontrollfragen

Evidenzbasierte Evaluation von Psychotherapieformen bei Depression

Welche Therapieform ist bei Depression wirksamer?

Online Psychotherapie



Klassische Psychotherapie



Beispieldatensatz

Evidenzbasierte Evaluation von Psychotherapieformen bei Depression

Becks Depressions-Inventar (BDI) zur Depressionsdiagnostik

BDI-II Fragebogen

Name	Alter	Geschlecht	Prüfung	Datum

Anleitung: Dieser Fragebogen enthält 21 Gruppen von Aussagen. Bitte lesen Sie jede dieser Gruppen von Aussagen sorgfältig durch und wählen Sie sich dann in jeder Gruppe eine Aussage heraus, die am besten beschreibt, wie Sie sich in den letzten zwei Wochen, einschließlich heute, gefühlt haben. Kreuzen Sie die Zahl neben der Aussage an, die Sie sich herausgerufen haben (0, 1, 2 oder 3). Falls in einer Gruppe mehrere Aussagen gleichermaßen auf Sie zutreffen, kreuzen Sie die Aussage mit der höchsten Zahl an. Antworten Sie bitte darauf, dass Sie in jeder Gruppe nicht mehr als eine Aussage ankreuzen, die gilt auch für Gruppe 16 (Veränderungen der Schlafgewohnheiten) oder Gruppe 18 (Veränderungen des Appetits).

1.) Traurigkeit

- 0 Ich bin nicht traurig.
- 1 Ich bin oft traurig.
- 2 Ich bin ständig traurig.
- 3 Ich bin so traurig oder unglücklich, dass ich es nicht aushalte.

2.) Pessimismus

- 0 Ich sehe nicht mutlos in die Zukunft.
- 1 Ich sehe mutlos in die Zukunft als sonst.
- 2 Ich bin mutlos und erwarte nicht, dass meine Situation besser wird.
- 3 Ich glaube, dass meine Zukunft hoffnungslos ist und nur noch schlechter wird.

3.) Versagensgefühle

- 0 Ich fühle mich nicht als Versager.
- 1 Ich habe häufiger Versagensgefühle.
- 2 Wenn ich zurückblicke, sehe ich eine Menge Fehlentscheidungen.
- 3 Ich habe das Gefühl, ich werde ein völliger Versager zu sein.

4.) Verlust von Freude

- 0 Ich kann die Dinge genauso gut genießen wie früher.
- 1 Ich kann die Dinge nicht mehr so genießen wie früher.
- 2 Dinge, die mir früher Freude gemacht haben, kann ich kaum mehr genießen.
- 3 Dinge, die mir früher Freude gemacht haben, kann ich überhaupt nicht mehr genießen.

5.) Schuldgefühle

- 0 Ich habe keine besonderen Schuldgefühle.
- 1 Ich habe oft Schuldgefühle wegen Dingen, die ich getan habe oder hätte tun sollen.
- 2 Ich habe die meiste Zeit Schuldgefühle.
- 3 Ich habe ständig Schuldgefühle.

6.) Bestrafungsgefühle

- 0 Ich habe nur das Gefühl, für etwas bestraft zu sein.
- 1 Ich habe das Gefühl, vielleicht bestraft zu werden.
- 2 Ich erwarte, bestraft zu werden.
- 3 Ich habe das Gefühl, bestraft zu sein.

7.) Selbstablehnung

- 0 Ich habe von mir genauso viel wie immer.
- 1 Ich habe Vertrauen in mich verloren.
- 2 Ich bin von mir enttäuscht.
- 3 Ich lehne mich völlig ab.

8.) Selbstvorwürfe

- 0 Ich kritisiere oder tadle mich nicht mehr als sonst.
- 1 Ich bin mir gegenüber kritischer als sonst.
- 2 Ich kritisieren mich für all meine Mängel.
- 3 Ich gebe mir die Schuld für alles Schlechte, was passiert.

9.) Selbstmordgedanken

- 0 Ich denke nicht daran, mir etwas anzutun.
- 1 Ich denke manchmal an Selbstmord.
- 2 Ich möchte mich umbringen, aber ich würde mich umbringen, wenn ich die Gelegenheit dazu hätte.
- 3 Ich gebe mir die Schuld für alles Schlechte, was passiert.

10.) Weinen

- 0 Ich weine nicht öfter als früher.
- 1 Ich weine jetzt mehr als früher.
- 2 Ich weine beim geringsten Anlass.
- 3 Ich möchte gern weinen, aber ich kann nicht.

11.) Unruhe

- 0 Ich bin nicht unruhiger als sonst.
- 1 Ich bin unruhiger als sonst.
- 2 Ich bin so unruhig, dass es mir schwerfällt, still zu sitzen.
- 3 Ich bin so unruhig, dass ich mich ständig bewege oder etwas tun muss.

12.) Interessenverlust

- 0 Ich habe das Interesse an anderen Menschen oder an Tätigkeiten nicht verloren.
- 1 Ich habe weniger Interesse an anderen Menschen oder an Dingen als sonst.
- 2 Ich habe das Interesse an anderen Menschen oder Dingen zum größten Teil verloren.
- 3 Es fällt mir schwer, mich überhaupt für irgend etwas zu interessieren.

13.) Entscheidungsfähigkeit

- 0 Ich bin so entscheidungsfreudig wie immer.
- 1 Es fällt mir schwerer als sonst, Entscheidungen zu treffen.
- 2 Es fällt mir sehr viel schwerer als sonst, Entscheidungen zu treffen.
- 3 Ich habe Mühe, überhaupt Entscheidungen zu treffen.

14.) Wertigkeit

- 0 Ich fühle mich nicht wertlos.
- 1 Ich habe mich für weniger wertvoll und nützlich als sonst.
- 2 Vergleichen mit anderen Menschen fühle ich mich viel weniger wert.
- 3 Ich fühle mich völlig wertlos.

15.) Energieverlust

- 0 Ich habe so viel Energie wie immer.
- 1 Ich habe weniger Energie als sonst.
- 2 Ich habe so wenig Energie, dass ich kaum noch etwas schaffe.
- 3 Ich habe keine Energie mehr, um überhaupt noch etwas zu tun.

16.) Veränderungen der Schlafgewohnheiten

- 0 Meine Schlafgewohnheiten haben sich nicht verändert.
- 1 Ich schlafe etwas mehr als sonst.
- 2 Ich schlafe etwas weniger als sonst.
- 3 Ich schlafe viel mehr als sonst.
- 4 Ich schlafe viel weniger als sonst.
- 5 Ich schlafe fast den ganzen Tag.
- 6 Ich wache 1-2 Stunden früher auf als gewöhnlich und kann dann nicht mehr einschlafen.

17.) Reizbarkeit

- 0 Ich bin nicht reizbarer als sonst.
- 1 Ich bin reizbarer als sonst.
- 2 Ich bin viel reizbarer als sonst.
- 3 Ich fühle mich dauernd gereizt.

18.) Veränderungen des Appetits

- 0 Mein Appetit hat sich nicht verändert.
- 1 Mein Appetit ist etwas schlechter als sonst.
- 2 Mein Appetit ist etwas größer als sonst.
- 3 Mein Appetit ist viel schlechter als sonst.
- 4 Mein Appetit ist viel größer als sonst.
- 5 Ich habe überhaupt keinen Appetit.
- 6 Ich habe ständig Heißhunger.

19.) Konzentrationschwierigkeiten

- 0 Ich kann mich so gut konzentrieren wie immer.
- 1 Ich kann mich nicht mehr so gut konzentrieren wie sonst.
- 2 Es fällt mir schwer, mich längere Zeit auf irgend etwas zu konzentrieren.
- 3 Ich kann überhaupt nicht mehr konzentrieren.

20.) Ermüdung oder Erschöpfung

- 0 Ich fühle mich nicht müde oder erschöpfter als sonst.
- 1 Ich fühle mich schneller müde oder erschöpfter als sonst.
- 2 Für viele Dinge, die ich üblicherweise tue, bin ich zu müde oder erschöpft.
- 3 Ich bin so müde oder erschöpft, dass ich fast nichts mehr tun kann.

21.) Verlust an sexuellem Interesse

- 0 Mein Interesse an Sexualität hat sich in letzter Zeit nicht verändert.
- 1 Ich interessiere mich weniger für Sexualität als früher.
- 2 Ich interessiere mich jetzt viel weniger für Sexualität.
- 3 Ich habe das Interesse an Sexualität völlig verloren.

Summe Seite 1:

Übertrag Seite 1: Summe Seite 2:

0 - 8 keine Depression

9 - 13 minimale Depression

14 - 19 leichte Depression

20 - 28 mittelschwere Depression

29 - 63 schwere Depression

Beispiel: Evaluation von Psychotherapieformen bei Depression

Experimentelle Bedingung
(Gruppen von $n = 50$)

Psychotherapie

Klassisch

Pre-BDI



Post-BDI

Online

Pre-BDI



Post-BDI

Beispieldatensatz

Einlesen des Datensatzes mit read.table()

```
file_path <- file.path(data_dir_path, "psychotherapie_datensatz.csv")

# z.B. könnte file_path so aussehen:
# "/home/belindame_f/ovgu/progr-und-deskr-stat-23/Daten/psychotherapie_datensatz.csv"

D <- read.table(file_path, sep = ",", header = T)
```

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

VSCode Interactive Viewers

Table Viewer

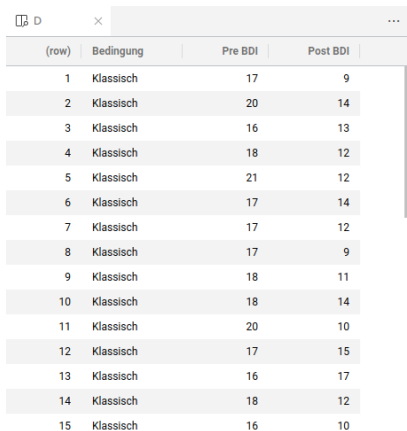
The screenshot shows the VS Code interface with a file named `table_viewer_demo.R` open. The code defines a data frame `D` with 18 rows and 3 columns: `Bedingung`, `Pro.BdI`, and `Post.BdI`. The Table Viewer on the right displays this data frame as a table. The table has 18 rows, with the third row highlighted. The columns are labeled `Bedingung`, `Pro.BdI`, and `Post.BdI`. The data in the table is as follows:

Bedingung	Pro.BdI	Post.BdI
1. Klausur	17	9
2. Klausur	20	14
3. Klausur	16	13
4. Klausur	18	12
5. Klausur	21	12
6. Klausur	17	14
7. Klausur	17	12
8. Klausur	17	9
9. Klausur	18	11
10. Klausur	18	14
11. Klausur	20	10
12. Klausur	17	15
13. Klausur	16	17
14. Klausur	18	12
15. Klausur	16	10
16. Klausur	18	13
17. Klausur	17	9
18. Klausur	14	13
19. Klausur	18	12

Mit dem Befehl `View()` oder im R **WORKSPACE** → **Global Environment** über das View Symbol  neben entsprechendem Objekt

[VS Code Wiki - Interactive viewers](#)

Datensatzübersicht mit View()



(row)	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
9	Klassisch	18	11
10	Klassisch	18	14
11	Klassisch	20	10
12	Klassisch	17	15
13	Klassisch	16	17
14	Klassisch	18	12
15	Klassisch	16	10

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Programmierübungen und Selbstkontrollfragen

Definition (Absolute und relative Häufigkeitsverteilungen)

$x := (x_1, \dots, x_n)$ mit $x_i \in \mathbb{R}$ sei ein *Datensatz* (manchmal auch “Urliste” genannt) und $A := \{a_1, \dots, a_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden verschiedenen Zahlenwerte (manchmal auch “Merkmalsausprägungen” genannt). Dann heißt die Funktion

$$h : A \rightarrow \mathbb{N}, a \mapsto h(a) := \text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i = a \quad (1)$$

die *absolute Häufigkeitsverteilung* der Zahlenwerte von x und die Funktion

$$r : A \rightarrow [0, 1], a \mapsto r(a) := \frac{h(a)}{n} \quad (2)$$

die *relative Häufigkeitsverteilung* der Zahlenwerte von x .

Bemerkungen

- Absolute und relative Häufigkeitsverteilungen fassen Datensätze zusammen
- Absolute und relative Häufigkeitsverteilungen können einen ersten Datenüberblick geben

Berechnung der Häufigkeitsverteilungen

Erzeugen der absoluten Häufigkeitsverteilung mit `table()`

Erzeugen der relativen Häufigkeitsverteilung durch Division mit n

```
x <- D$Pre.BDI           # Double vector der Pre BDI Werte
n <- length(x)           # Anzahl der Datenwerte (100)
H <- as.data.frame(table(x)) # absolute Häufigkeitsverteilung (dataframe)
names(H) <- c("a", "h")    # Spaltenbenennung
H$r <- H$h/n              # relative Häufigkeitsverteilung
```

a	h	r
14	1	0.01
15	3	0.03
16	6	0.06
17	17	0.17
18	21	0.21
19	20	0.20
20	17	0.17
21	12	0.12
22	2	0.02
23	1	0.01

Visualisierung Häufigkeitsverteilungen

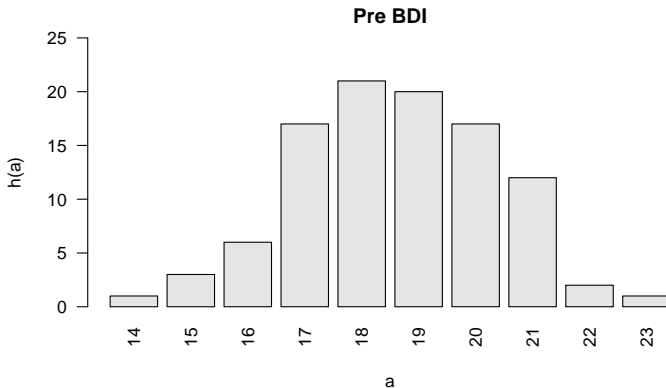
Visualisierung der absoluten Häufigkeitsverteilung mit `barplot()`

```
h          <- H$h          # h(a) Werte
names(h)   <- H$a          # barplot braucht a Werte als names
dev.new()  # Abbildungsinitialisierung
barplot(   # Balkendiagramm
  h,       # absolute Häufigkeiten
  col      = "gray90",     # Balkenfarbe
  xlab     = "a",          # x Achsenbeschriftung
  ylab     = "h(a)",       # y Achsenbeschriftung
  ylim     = c(0,25),      # y Achsengrenzen
  las      = 2,            # x Tick Orientierung
  main     = "Pre BDI"     # Titel
)
```

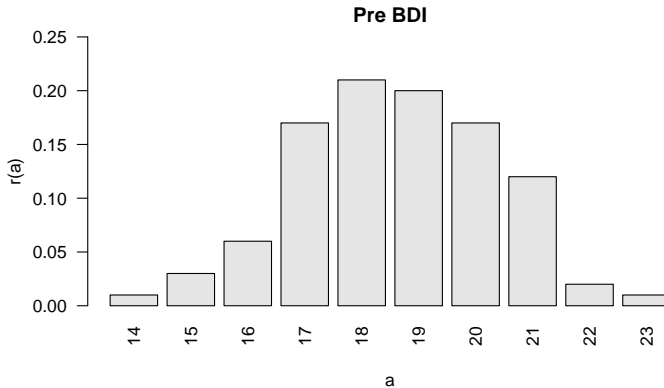
Speichern von Abbildungen mit `dev.copy2pdf()`

```
dev.copy2pdf(                                     # PDF Kopierfunktion
  file      = file.path(pfad_zu_abbildungs_ordner, "pds_7_ha_prebdi.pdf"), # Dateiname
  width     = 7,                                   # Breite (inch)
  height    = 4,                                   # Höhe (inch)
)
```

Absolute Häufigkeitsverteilung aller Pre-BDI Werte



Relative Häufigkeitsverteilung aller Pre-BDI Werte



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Programmierübungen und Selbstkontrollfragen

Definition (Histogramm)

Ein *Histogramm* ist ein Diagramm, in dem zu einem Datensatz $x = (x_1, \dots, x_n)$ mit verschiedenen Zahlenwerten $A := \{a_1, \dots, a_m\}$, $m \leq n$ über benachbarten Intervallen $[b_{j-1}, b_j[$, welche *Klassen* oder *Bins* genannt werden, für $j = 1, \dots, k$ Rechtecke mit

Breite $d_j = b_j - b_{j-1}$

Höhe $h(a)$ oder $r(a)$ mit $a \in [b_{j-1}, b_j[$

abgebildet sind, wobei $b_0 := \min A$ und $b_k := \max A$ angenommen werden soll.

Bemerkungen

- Das Aussehen eines Histogramms ist stark von der Anzahl k der Klassen abhängig.
- Mit der Aufrundungsfunktion $\lceil \cdot \rceil$ sind konventionelle Werte für k

$k := \lceil (b_k - b_0)h \rceil$ h ist die gewünschte Klassenbreite

$k := \lceil \sqrt{n} \rceil$ Excelstandard

$k := \lceil \log_2 n + 1 \rceil$ Implizite Normalverteilungsannahme (Sturges, 1926)

$h := 3.49 S_n / \sqrt[3]{n}$ Min. MSE Dichteschätzung bei Normalverteilung (Scott, 1979)

Berechnung und Visualisierung von Histogrammen

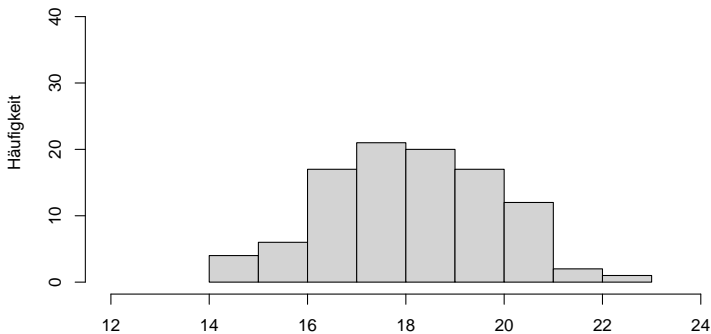
Berechnung und Visualisierung von Histogrammen mit hist()

- Die Klassen $[b_{j-1}, b_j[, j = 1, \dots, k$ werden als Argument breaks festgelegt
- breaks ist der atomic vector $c(b_0, b_1, \dots, b_k)$ mit Länge $k + 1$
- Per default benutzt hist() eine Modifikation der Sturges Empfehlung $k = \lceil \log_2 n + 1 \rceil$
- hist() bietet eine Vielzahl weiterer Spezifikationsmöglichkeiten

```
# Default Histogramm
x      <- D$Pre.BDI           # Datensatz
x_min  <- 12                  # x Achsengrenze (links)
x_max  <- 25                  # x Achsengrenze (rechts)
y_min  <- 0                   # y Achsengrenze (unten)
y_max  <- 45                  # y Achsengrenze (oben)
hist(      # Histogramm
  x,      # Datensatz
  xlim = c(x_min, x_max), # x Achsengrenzen
  ylim = c(y_min, y_max), # y Achsengrenzen
  ylab = "Häufigkeit",    # y-Achsenbezeichnung
  xlab = "",              # x-Achsenbezeichnung
  main = "Pre-BDI, R Default" # Titel
)
```

Visualisierung

Pre-BDI, R Default



Histogramm mit hist() und default breaks - Beispiel

Ausgabe der Ergebnisse

```
# Default Histogramm
Ergebnisse <- hist(      # hist() output in Variable speichern
  x,                      # Datensatz
  plot = FALSE
)

print(Ergebnisse)       # Ausgabe der Ergebnisse
```

\$breaks

[1] 14 15 16 17 18 19 20 21 22 23

\$counts

[1] 4 6 17 21 20 17 12 2 1

\$density

[1] 0.04 0.06 0.17 0.21 0.20 0.17 0.12 0.02 0.01

\$mids

[1] 14.5 15.5 16.5 17.5 18.5 19.5 20.5 21.5 22.5

\$xname

[1] "x"

\$equidist

[1] TRUE

attr(,"class")

[1] "histogram"

Histogramme mit alternativen Klassengrößen

Berechnung von Klassenanzahlen und breaks Argument

```
# Histogramm mit gewünschter Klassenbreite
h  <- 1                                # gewünschte Klassenbreite
b_0 <- min(x)                          # b_0
b_k <- max(x)                          # b_k
k  <- ceiling((b_k - b_0)/h)           # Anzahl der Klassen
b  <- seq(b_0, b_k, by = h)            # Klassen [b_{j-1}, b_j[

# Excelstandard
n  <- length(x)                        # Anzahl Datenwerte
k  <- ceiling(sqrt(n))                 # Anzahl der Klassen
b  <- seq(b_0, b_k, len = k + 1)       # Klassen [b_{j-1}, b_j[
h  <- b[2] - b[1]                      # Klassenbreite

# Sturges
n  <- length(x)                        # Anzahl Datenwerte
k  <- ceiling(log2(n)+1)               # Anzahl der Klassen
b  <- seq(b_0, b_k, len = k + 1)       # Klassen [b_{j-1}, b_j[
h  <- b[2] - b[1]                      # Klassenbreite

# Scott
n  <- length(x)                        # Anzahl Datenwerte
S  <- sd(x)                            # Stichprobenstandardabweichung
h  <- ceiling(3.49*S/(n^(1/3)))         # Klassenbreite
k  <- ceiling((b_k - b_0)/h)           # Anzahl der Klassen
b  <- seq(b_0, b_k, len = k + 1)       # Klassen [b_{j-1}, b_j[
```

Berechnung und Visualisierung von Histogrammen mit hist()

- Die Klassen $[b_{j-1}, b_j], j = 1, \dots, k$, die in der Variable **b** gespeichert sind, werden als Argument mit **breaks** festgelegt
- **breaks** ist der atomic vector $c(b_0, b_1, \dots, b_k)$ mit Länge $k + 1$

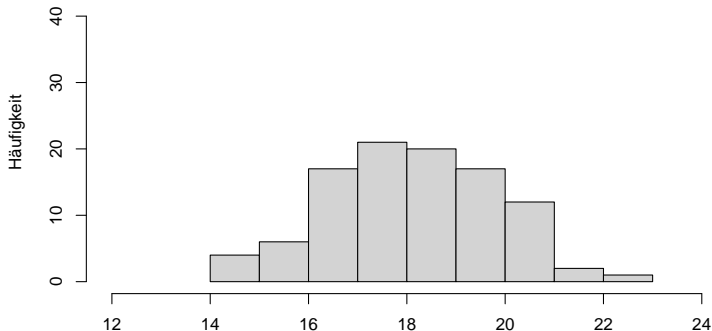
```
# Default Histogramm
x      <- D$Pre.BDI
x_min  <- 12
x_max  <- 25
y_min  <- 0
y_max  <- 30
hist(
  x,
  breaks= b,
  xlim  = c(x_min, x_max),
  ylim  = c(y_min, y_max),
  ylab  = "Häufigkeit",
  xlab  = "",
  main  = sprintf("Pre-BDI, k = %.0f, h = %.2f", k, h))
)
```

```
# Datensatz
# x Achsengrenze (unten)
# x Achsengrenze (oben)
# y Achsengrenze (oben)
# y Achsengrenze (unten)
# Histogramm
# Datensatz
# breaks
# x Achsengrenzen
# y Achsengrenzen
# y-Achsenbezeichnung
# x-Achsenbezeichnung
# Titel
```

Histogramme - Beispiel

Gewünschte Klassenbreite $h := 1$

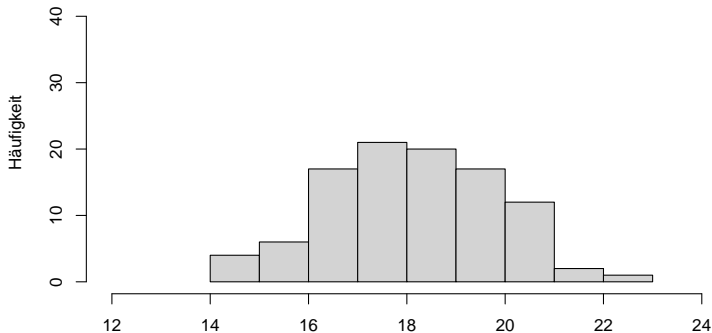
Pre-BDI, $k = 9$, $h = 1.00$



Histogramme - Beispiel

Gewünschte Klassenbreite $h := 1.5$

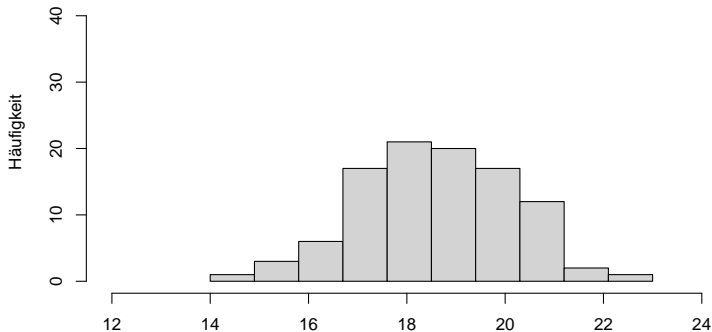
Pre-BDI, $k = 9$, $h = 1.00$



Histogramme - Beispiel

Excelstandard $k := \lceil \sqrt{n} \rceil$

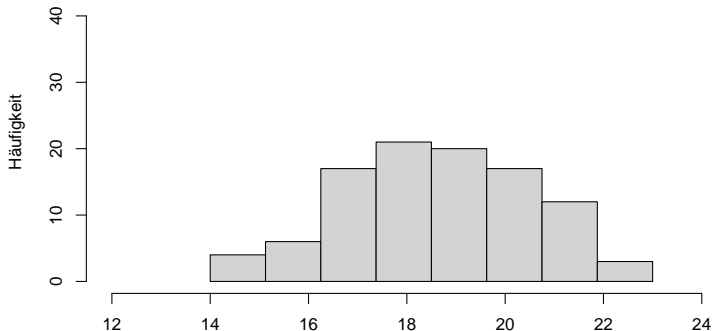
Pre-BDI, k = 10, h = 0.90



Histogramme - Beispiel

nach Sturges (1926) , $k := \lceil \log_2 n + 1 \rceil$

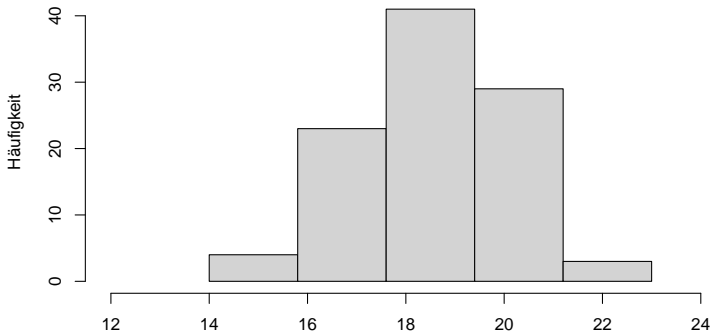
Pre-BDI, k = 8, h = 1.12



Histogramme - Beispiel

nach Scott (1979) , $h := 3.49S_n/\sqrt[3]{n}$

Pre-BDI, k = 5, h = 1.80



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Programmierübungen und Selbstkontrollfragen

1. Visualisiere die Häufigkeitsverteilungen der Post-BDI Daten.
2. Visualisiere die Häufigkeitsverteilungen der Differenzen von Post- und Pre-BDI Daten.
3. Visualisiere die Häufigkeitsverteilungen der Differenzen von Post- und Pre-BDI Daten getrennt nach den experimentellen Bedingungen "Klassisch" und "Online". Nutze dazu dein Wissen zu den Prinzipien der Indizierung in R.
4. Beschreibe die in der vorherigen Aufgabe erstellten Häufigkeitsverteilungen.
5. Erstelle Histogramme der Differenzen von Post- und Pre-BDI Daten mithilfe verschiedener Methoden zur Bestimmung der Klassengrößen: mit einer Klassenbreite von 3, nach dem Excelstandard, mit der Sturges Klassenanzahl und mit der Scott Klassenanzahl.
6. Beschreibe die in der vorherigen Aufgabe erstellten Histogramme.

1. Definiere die Begriffe der absoluten und relativen Häufigkeitsverteilungen.
2. Definiere den Begriff des Histogramms.
3. Erläutere die Bedeutung der Klassenanzahl für das Erscheinungsbild eines Histogramms.

References

Scott, David W. 1979. "On Optimal and Data-Based Histograms," 6.

Sturges, Herbert A. 1926. "The Choice of a Class Interval." *Journal of the American Statistical Association* 21 (153): 65–66. <https://doi.org/10.1080/01621459.1926.10502161>.