



# Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2023/24

Belinda Fleischmann

Inhalte basieren auf Programmierung und Deskriptive Statistik von Dirk Ostwald, lizenziert unter CC BY-NC-SA 4.0

Datum	Einheit	Thema
11.10.23	Einführung	(1) Einführung
18.10.23	R Grundlagen	(2) R und Visual Studio Code
25.10.23	R Grundlagen	(2) R und Visual Studio Code
01.11.23	R Grundlagen	(3) Vektoren
08.11.23	R Grundlagen	(4) Matrizen
15.11.23	R Grundlagen	(5) Listen und Dataframes
22.11.23	R Grundlagen	(6) Datenmanagement
<b>29.11.23</b>	<b>R Grundlagen</b>	<b>(7) Häufigkeitsverteilungen</b>
06.12.23	R Grundlagen	(8) Verteilungsfunktionen und Quantile
13.12.23	Deskriptive Statistik	(9) Maße der zentralen Tendenz
20.12.23	<i>Leistungsnachweis Teil 1</i>	
20.12.23	Deskriptive Statistik	(10) Maße der Datenvariabilität
	Weihnachtspause	
10.01.24	Deskriptive Statistik	(11) Anwendungsbeispiel (Deskriptive Statistik)
17.01.24	Inferenzstatistik	(12) Anwendungsbeispiel (Parameterschätzung, Konfidenzintervalle)
24.01.24	Inferenzstatistik	(13) Anwendungsbeispiel (Hypothesentest)
25.01.24	<i>Leistungsnachweis Teil 2</i>	

## (7) Häufigkeitsverteilungen

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Übungen und Selbstkontrollfragen

## **Motivation**

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Übungen 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

Übungen 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	Seitwertsch.	Seitwertsch.

**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, das gilt auch für Gruppe 16 (Veränderungen der Schlafgewohnheiten) oder Gruppe 18 (Veränderungen des Appetits).

**1.) Traurigkeit**

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

**2.) Pessimismus**

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

**3.) Versagensgefühle**

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

**4.) Verlust von Freude**

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

**5.) Schuldgefühle**

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

**11.) Unruhe**

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

**12.) Interessenverlust**

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

**13.) Entscheidungsfähigkeit**

- Ich bin bei Entscheidungen wie immer.
- Es fällt mir schwerer als sonst, Entscheidungen zu treffen.
- Es fällt mir sehr viel schwerer als sonst, Entscheidungen zu treffen.
- Ich habe Mühe, überhaupt Entscheidungen zu treffen.

**14.) Wertigkeit**

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

**15.) Energieverlust**

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

**16.) Veränderungen der Schlafgewohnheiten**

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

**17.) Reizbarkeit**

- Ich bin nicht reizbarer als sonst.
- Ich bin reizbarer als sonst.
- Ich bin viel reizbarer als sonst.
- Ich fühle mich dauernd gereizt.

**18.) Veränderungen des Appetits**

- Mein Appetit hat sich nicht verändert.
- Mein Appetit ist etwas schlechter als sonst.
- Mein Appetit ist etwas größer als sonst.
- Mein Appetit ist viel schlechter als sonst.
- Mein Appetit ist viel größer als sonst.
- Ich habe überhaupt keinen Appetit.
- Ich habe ständig Heißhunger.

**19.) Konzentrationschwierigkeiten**

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

**20.) Ermüdung oder Erschöpfung**

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

**21.) Verlust an sexuellem Interesse**

- Mein Interesse an Sexualität hat sich in letzter Zeit nicht verändert.
- Ich interessiere mich weniger für Sexualität als früher.
- Ich interessiere mich jetzt viel weniger für Sexualität.
- 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

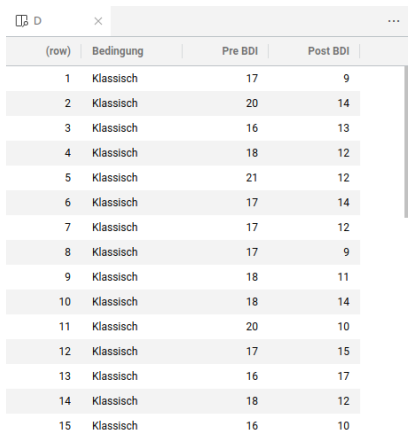
## Einlesen des Datensatzes mit read.table()

```
pfad_zu_datei <- file.path(pfad_zu_Datenordner, "psychotherapie_datensatz.csv")  
  
# z.B. könnte pfad_zu_datei so aussehen:  
# "/home/belindame_f/ovgu/progr-und-deskr-stat-23/Daten/psychotherapie_datensatz.csv"  
  
D <- read.table(pfad_zu_datei, 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

## 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

Übungen 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

Erzeugen der absoluten Häufigkeitsverteilung mit `table()`

Erzeugen der relativen Häufigkeitsverteilung durch Division mit  $n$

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

# Visualisierung Häufigkeitsverteilungen

## Visualisierung der absoluten Häufigkeitsverteilung mit `barplot()`

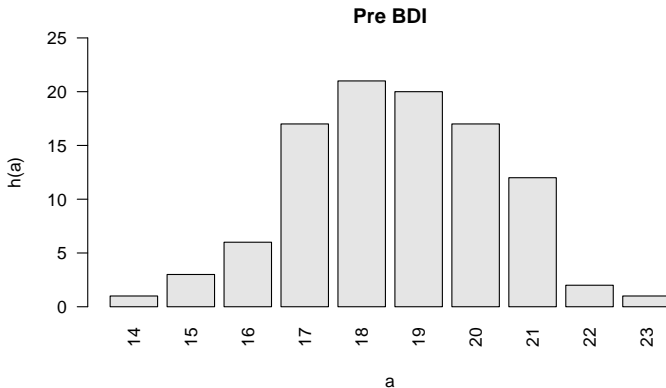
```
h          <- H$h          # h(a) Werte
names(h)   <- H$a          # barplot braucht a Werte als names
dev.new()  # Abbildungsinitialisierung
barplot(   # Balkendiagramm
  h,       # absolute Häufigkeiten
  col      = "gray90",     # Balkenfarbe
  xlab     = "a",          # x Achsenbeschriftung
  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()`

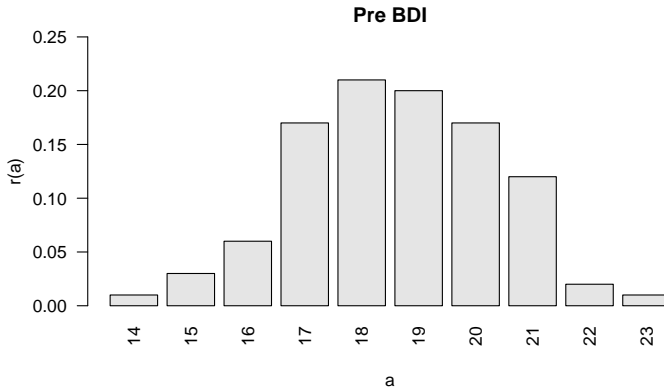
```
dev.copy2pdf(                                     # PDF Kopierfunktion
  file      = file.path(pfad_zu_abbildungs_ordner, "pds_7_ha_prebdi.pdf"), # Dateiname
  width     = 7,                                  # Breite (inch)
  height    = 4,                                  # Höhe (inch)
)
```



## Absolute Häufigkeitsverteilung aller Pre-BDI Werte



## Relative Häufigkeitsverteilung aller Pre-BDI Werte



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

**Histogramme**

Übungen 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)

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

# Berechnung und Visualisierung von Histogrammen

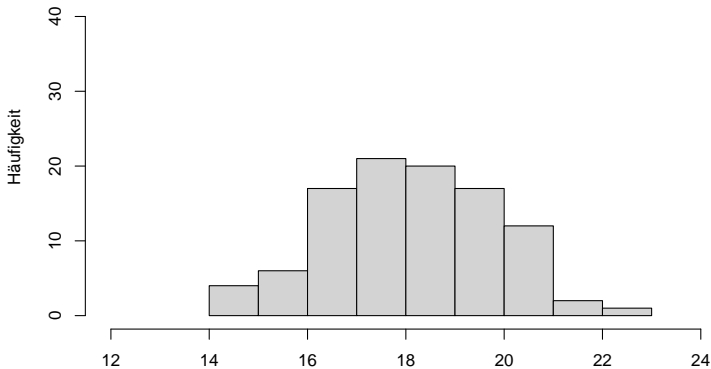
## Berechnung und Visualisierung von Histogrammen mit hist()

- 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 (unten)
x_max  <- 25                 # x Achsengrenze (oben)
y_min  <- 0                  # y Achsengrenze (oben)
y_max  <- 30                 # y Achsengrenze (unten)
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)           # 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)           # 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)           # Klassen [b_{j-1}, b_j[
```



## Berechnung und Visualisierung von Histogrammen mit hist()

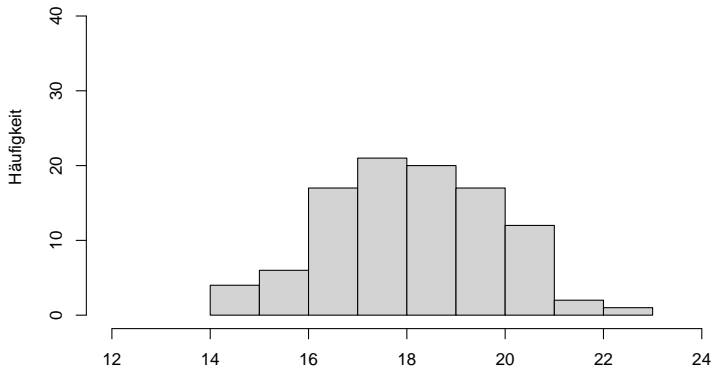
- Die Klassen  $[b_{j-1}, b_j], j = 1, \dots, k$ , die in der Variable **b** gespeichert sind, werden als Argument mit **breaks** festgelegt
- **breaks** ist der atomic vector  $c(b_0, b_1, \dots, b_k)$  mit Länge  $k + 1$

```
# 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
```

Gewünschte Klassenbreite  $h := 1$

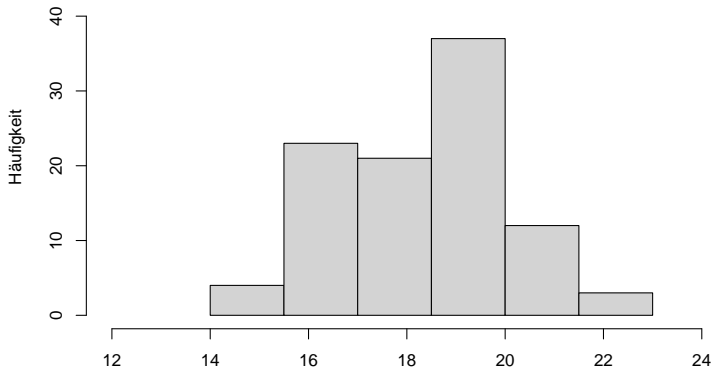
**Pre-BDI,  $k = 9$ ,  $h = 1.00$**



## Histogramme - Beispiel

Gewünschte Klassenbreite  $h := 1.5$

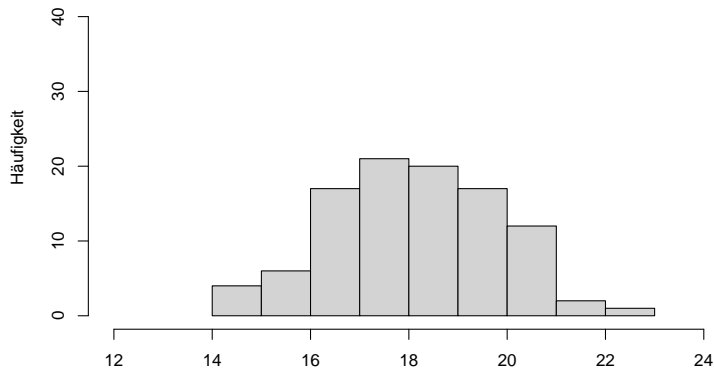
**Pre-BDI,  $k = 6$ ,  $h = 1.50$**



## Histogramme - Beispiel

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

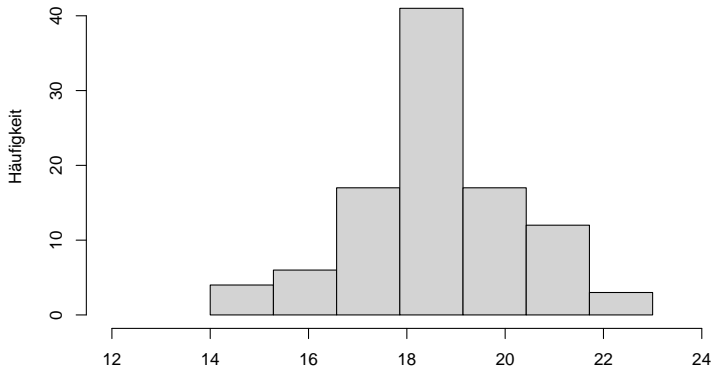
**Pre-BDI, k = 10, h = 1.00**



## Histogramme - Beispiel

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

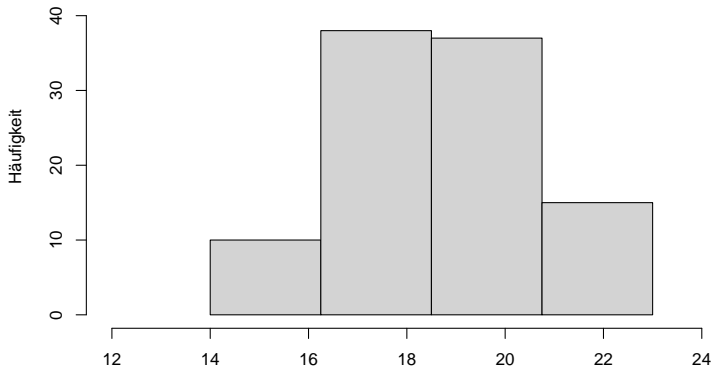
**Pre-BDI,  $k = 8$ ,  $h = 1.29$**



## Histogramme - Beispiel

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

**Pre-BDI, k = 5, h = 2.25**



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

**Übungen und Selbstkontrollfragen**

1. Definieren Sie die Begriffe der absoluten und relativen Häufigkeitsverteilungen.
2. Visualisieren Sie die Häufigkeitsverteilungen der Post-BDI Daten.
3. Visualisieren Sie die Häufigkeitsverteilungen der Differenzen von Post- und Pre-BDI Daten.
4. Visualisieren Sie die Häufigkeitsverteilungen der Differenzen von Post- und Pre-BDI Daten getrennt nach den experimentellen Bedingungen "Klassisch" und "Online". Nutzen Sie dazu Ihr Wissen zu den Prinzipien der Indizierung in R.
5. Beschreiben Sie die in der vorherigen Aufgabe erstellten Häufigkeitsverteilungen.
6. Definieren Sie den Begriff des Histogramms.
7. Erläutern Sie die Bedeutung der Klassenanzahl für das Erscheinungsbild eines Histogramms.
8. Visualisieren Sie Histogramme der Daten wie in Aufgabe 4. mit einer Klassenbreite von 3, dem Excelstandard, der Sturges Klassenanzahl und der Scott Klassenanzahl.
9. Beschreiben Sie die in der vorherigen Aufgabe erstellten Histogramme.



## 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>.