Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2025/26

Belinda Fleischmann und Dirk Ostwald

	Gruppe 1/2	Gruppe 3	Format	Thema
1	Mi, 15.10.	Do, 16.10.	Seminar	(1) Grundbegriffe der Informatik
2	Mi, 22.10.	Do, 23.10.	Seminar	(2) Arithmetik und Variablen
3	Mi, 29.10.	Do, 30.10.	Übung	(2) Arithmetik und Variablen
4	Mi, 05.11.	Do, 06.11.	Seminar	(3) Vektoren und Matrizen
5	Mi, 12.11.	Do, 13.11.	Übung	(3) Vektoren und Matrizen
6	Mi, 19.11.	Do, 20.11.	Seminar	(4) Listen und Dataframes
7	Mi, 26.11.	Do, 27.11.	Übung	(4) Listen und Dataframes
8	Mi, 03.12.	Do, 04.12.	Seminar	(5) Datenmanagement
9	Mi, 10.12.	Do, 11.12.	Übung	(5) Datenmanagement
	Mo, 15.12		Abgabe	*Individualleistung (1/2)*
10	Mi, 17.12.	Do, 18.12.	Seminar	(6) Strukturiertes Progr.: Kontrollfluss, Debugging
11	**Mi, 07.01.**	**Do, 08.01.**	**Seminar**	**(7) Häufigkeitsverteilungen**
12	Mi, 14.01.	Do, 15.01.	Übung	(7) Häufigkeitsverteilungen
13	Mi, 21.01.	Do, 22.01.	Seminar	(8) Maße der zentralen Tendenz und Datenvariabilität
14	Mi, 28.01.	Do, 29.01.	Übung	(8) Maße der zentralen Tendenz und Datenvariabilität
	Mo, 02.02		Abgabe	*Individualleistung (2/2)*

(7) Häufigkeitsverteilungen

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Definition und Ziele der Deskriptiven Statistik

- Die Deskriptive Statistik ist die *beschreibende* Statistik.
- Ziel ist es, Daten der menschlichen Betrachtung zugänglich zu machen.
- Reine Zahlendarstellungen sind bei großen Datensätzen schwer zu erfassen.
- Deskriptivstatistiken bieten graphische Darstellungen und zusammenfassende Maßzahlen, die charakteristische Aspekte eines Datensatzes wie seinen "durchschnittlicher" Wert oder seine Variabilität beschreiben.

Typische Methoden der Deskriptiven Statistik

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

Bezug zu Wahrscheinlichkeitstheorie

- Die Deskriptive Statistik basiert nicht auf der Wahrscheinlichkeitstheorie.
- Allerdings ergeben viele Aspekte der Deskriptivstatistik erst vor dem Hintergrund wahrscheinlichkeitstheoretischer Betrachtungen Sinn.
- Umgekehrt sind viele Begriffe der Wahrscheinlichkeitstheorie durch deskriptivstatistische Konzepte motiviert.
- Auch wenn diese Einheit ohne wahrscheinlichkeitstheoretischen Hintergrund auskommt, erschließt sich seine volle Bedeutung erst im Kontext der Begriffe der Wahrscheinlichkeitstheorie.

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

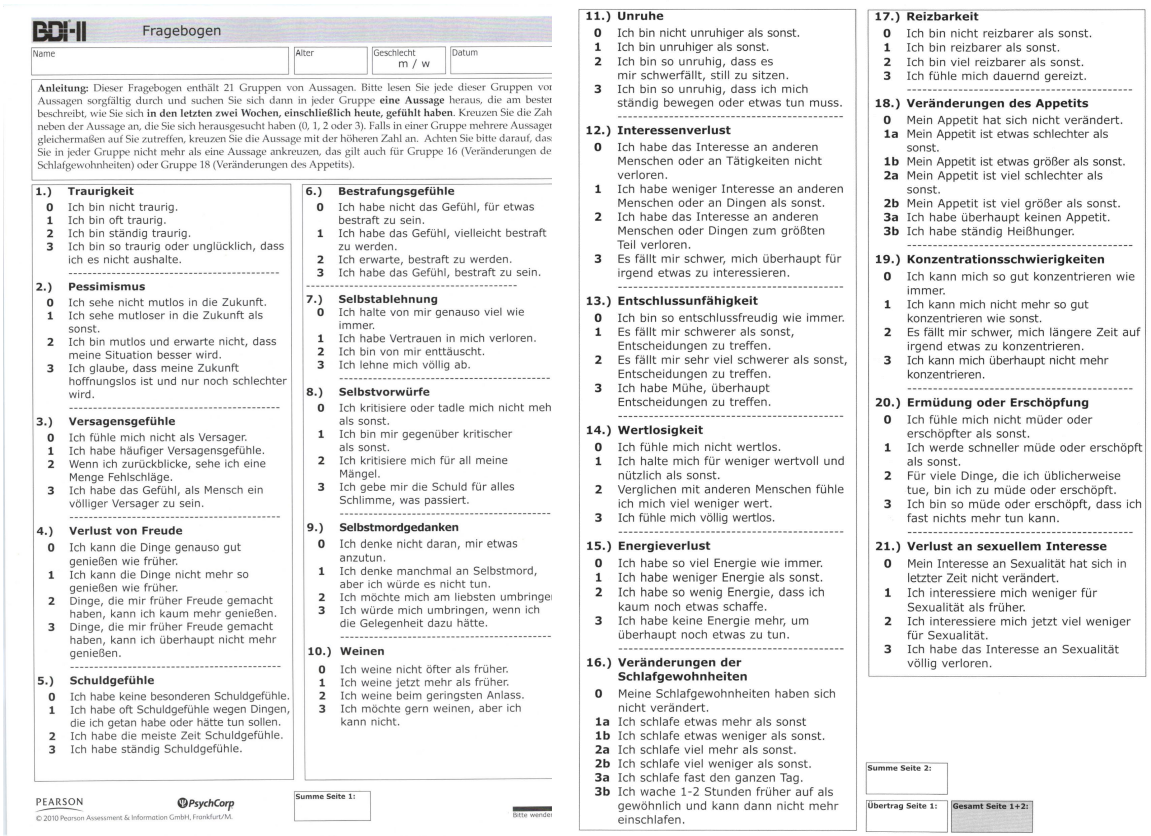
Evidenzbasierte Evaluation von Psychotherapieformen bei Depression

Online Psychotherapie

Klassische Psychotherapie

Evidenzbasierte Evaluation von Psychotherapieformen bei Depression

Becks Depressions-Inventar (BDI-II) zur Depressionsdiagnostik



BDI-II Fragebogen

Name | Alter | Geschlecht m / w | Datum

Anleitung: Dieser Fragebogen enthält 21 Gruppen von Aussagen. Bitte lesen Sie jede dieser Gruppen von Aussagen sorgfältig durch und suchen 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 herausgesucht haben (0, 1, 2 oder 3). Falls in einer Gruppe mehrere Aussagen gleichermaßen auf Sie zutreffen, kreuzen Sie die Aussage mit der höheren Zahl an. Achten 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
- 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 mutloser 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 Fehlschläge.
- 3 Ich habe das Gefühl, als Mensch 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 nicht 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 halte 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 kritisiere mich für all meine Mängel.
- 3 Ich gebe mir die Schuld für alles Schlimme, was passiert.

9.) Selbstmordgedanken
- 0 Ich denke nicht daran, mir etwas anzutun.
- 1 Ich denke manchmal an Selbstmord, aber ich würde es nicht tun.
- 2 Ich möchte mich am liebsten umbringen.
- 3 Ich würde mich umbringen, wenn ich die Gelegenheit dazu hätte.

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.

Summe Seite 1:

PEARSON PsychCorp © 2010 Pearson Assessment & Information GmbH, Frankfurt/M.

Bitte wenden

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 bewegen 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.) Entschlussunfähigkeit
- 0 Ich bin so entschlussfreudig 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.) Wertlosigkeit
- 0 Ich fühle mich nicht wertlos.
- 1 Ich halte mich für weniger wertvoll und nützlich als sonst.
- 2 Verglichen 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.
- 1a Ich schlafe etwas mehr als sonst
- 1b Ich schlafe etwas weniger als sonst.
- 2a Ich schlafe viel mehr als sonst.
- 2b Ich schlafe viel weniger als sonst.
- 3a Ich schlafe fast den ganzen Tag.
- 3b 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.
- 1a Mein Appetit ist etwas schlechter als sonst.
- 1b Mein Appetit ist etwas größer als sonst.
- 2a Mein Appetit ist viel schlechter als sonst.
- 2b Mein Appetit ist viel größer als sonst.
- 3a Ich habe überhaupt keinen Appetit.
- 3b Ich habe ständig Heißhunger.

19.) Konzentrationsschwierigkeiten
- 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 mich überhaupt nicht mehr konzentrieren.

20.) Ermüdung oder Erschöpfung
- 0 Ich fühle mich nicht müder oder erschöpfter als sonst.
- 1 Ich werde schneller müde oder erschöpft 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 2:

Übertrag Seite 1: | Gesamt Seite 1+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()`

```
dateipfad <- file.path(datenordner, "psychotherapie_datensatz.csv")
D         <- read.csv(dateipfad)
```

Daten der ersten acht Proband:innen jeder Gruppe

```
head(D[D$Bedingung == "Klassisch", ], 8)
```

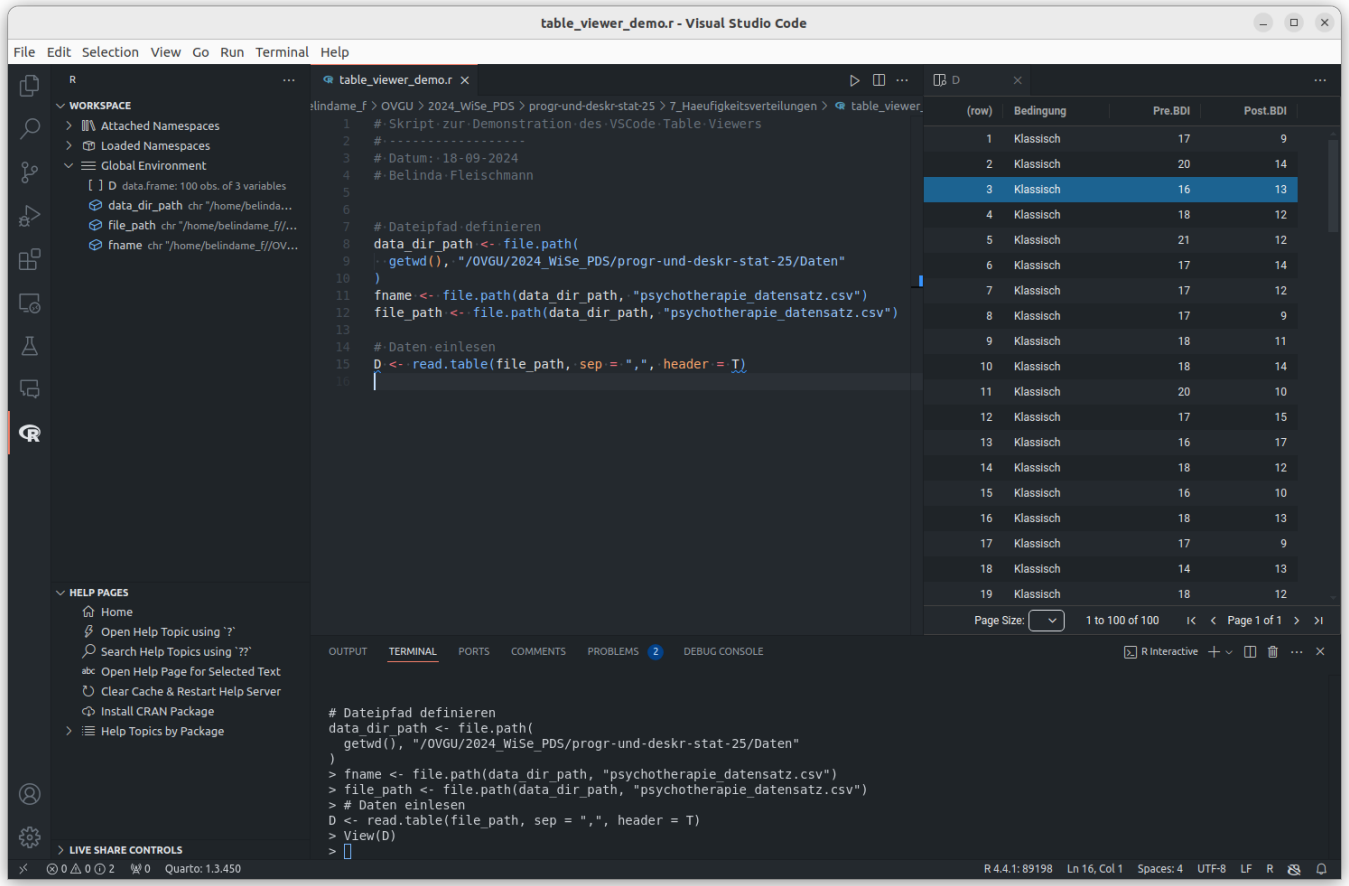
```
  Bedingung Pre.BDI Post.BDI
1 Klassisch      17        9
2 Klassisch      20       14
3 Klassisch      16       13
4 Klassisch      18       12
5 Klassisch      21       12
6 Klassisch      17       14
7 Klassisch      17       12
8 Klassisch      17        9
```

```
head(D[D$Bedingung == "Online", ], 8)
```

```
   Bedingung Pre.BDI Post.BDI
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



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

VS Code Wiki - Interactive viewers

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

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

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

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

Hinzufügen einer ΔBDI Spalte zum Dataframe

```
D$Delta.BDI <- -(D$Post.BDI - D$Pre.BDI)                    # \Delta BDI Maß
```

Daten der ersten acht Proband:innen jeder Gruppe

```
  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

   Bedingung Pre.BDI Post.BDI Delta.BDI
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
```

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Definition (Absolute und relative Häufigkeitsverteilungen)

$y := (y_1, ..., y_n)$ sei ein Datensatz und $W := \{w_1, ..., w_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden Werte. Dann heißt die Funktion

$$h : W \to \mathbb{N}, w \mapsto h(w) := \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i = w \tag{2}$$

die *absolute Häufigkeitsverteilung* der Werte von y und die Funktion

$$r : W \to [0, 1], w \mapsto r(w) := \frac{h(w)}{n} \tag{3}$$

heißt die *relative Häufigkeitsverteilung* der Werte von y.

Bemerkungen

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

Berechnung der Häufigkeitsverteilungen

Die **absoluten Häufigkeiten** einer kategorialen Variable können in R mit der Funktion `table()` bestimmt werden.

```
y        <- D$Pre.BDI                  # double vector der Pre BDI Werte
n        <- length(y)                  # Anzahl der Datenwerte (hier: 100)
H        <- as.data.frame(table(y))    # absolute Häufigkeitsverteilung als Dataframe
names(H) <- c("w", "h")                # Benennen der Spalten im Dataframe
```

Die **relativen Häufigkeiten** erhält man, indem man die absoluten Häufigkeiten durch die Stichprobengröße n teilt.

```
H$r      <- H$h / n                    # Neue Spalte mit relativen Häufigkeiten
print(H)                               # Ausgabe des Dataframes
```

```
    w  h    r
1  14  1 0.01
2  15  3 0.03
3  16  6 0.06
4  17 17 0.17
5  18 21 0.21
6  19 20 0.20
7  20 17 0.17
8  21 12 0.12
9  22  2 0.02
10 23  1 0.01
```

Visualisierung der absoluten Häufigkeitsverteilung mit `barplot()`

```r
h          <- H$h              # Kopie der absoluten Häufigkeiten in einen Vektor
names(h)   <- H$w              # Den Häufigkeitswerten die entsprechenden Kategorien zuweisen
barplot(                       # Balkendiagramm
  h,                           # absolute Haeufigkeiten
  col  = "gray90",             # Balkenfarbe
  xlab = "w",                  # x Achsenbeschriftung
  ylab = "h(w)",               # y Achsenbeschriftung
  ylim = c(0, 25),             # y Achsengrenzen
  las  = 2,                    # x Tick Orientierung (2: perpendicular to the axis)
  main = "Pre BDI"             # Titel
)
```

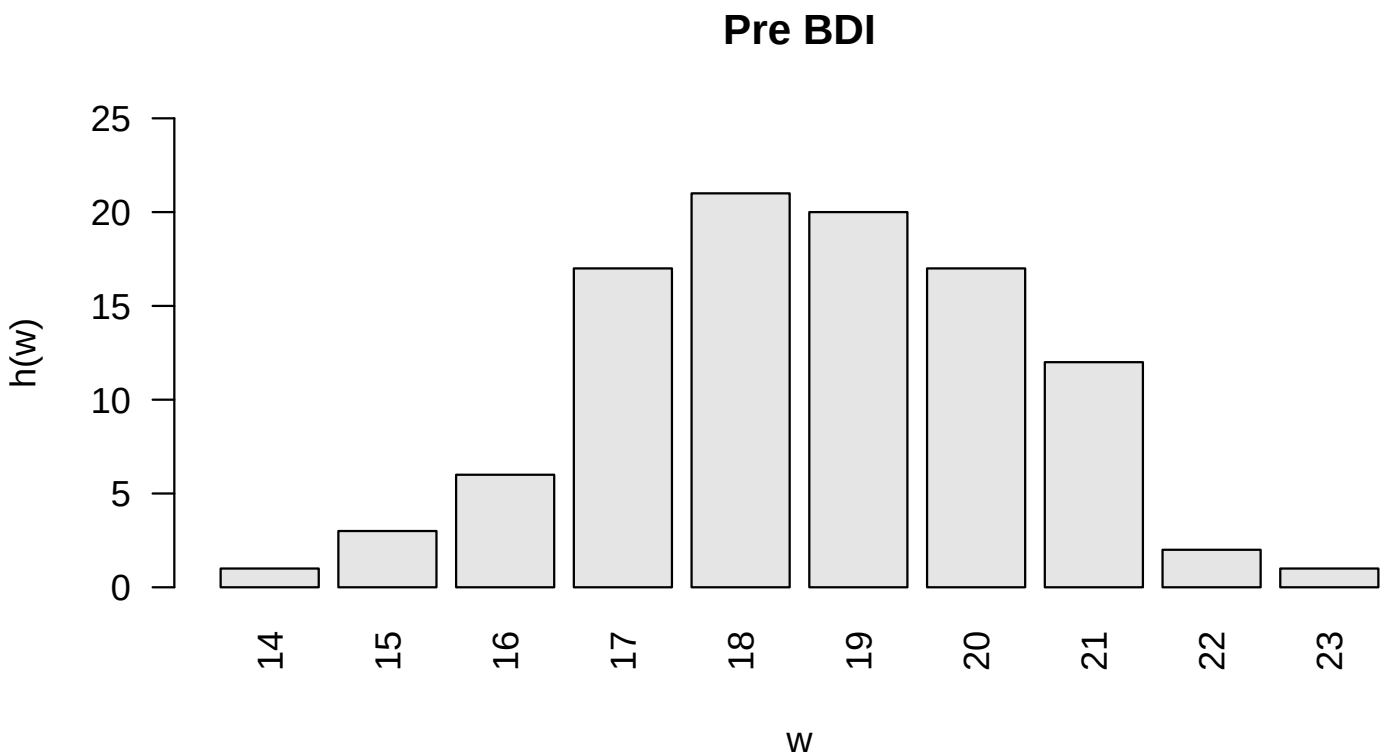
Anmerkungen

- Allgemeine **R Graphics** Plot Parameter können mit `?par` und Plotspezifische Parameter auf der jeweiligen Plot Help Page (z.B. `?barplot`) nachgeschlagen werden

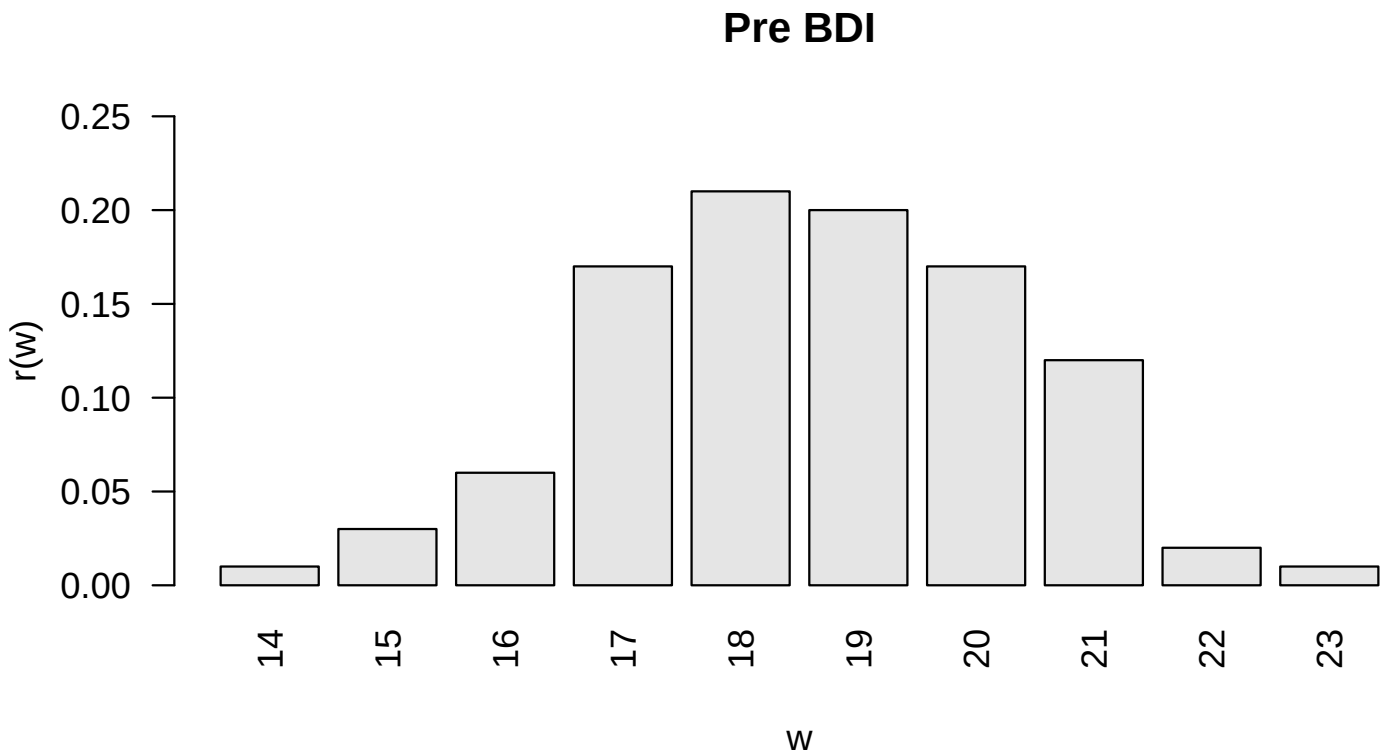
Speichern von Abbildungen mit `dev.copy2pdf()`

```
# Variante 1: Dateinamen im selben Verzeichnis wie das Skript erstellen
skriptordner      <- sys.frame(1)$ofile  # Verzeichnis des aktuellen Skripts ermitteln
abb_ausgabepfad   <- file.path(          # Absoluter Pfad zur Zieldatei im Skriptverzeichnis
  skriptordner,
  "pds_7_ah_prebdi.pdf"
)

# Variante 2: Dateinamen in einem Unterverzeichnis "Abbildungen" im Arbeitsverzeichnis erstellen
abbildungsordner  <- file.path(          # Absoluter Pfad zum Unterverzeichnis "Abbildungen"
  getwd(),                               # Im Beispiel hier: im aktuellen working directory
  "VL_Abbildungen"
)
abb_ausgabepfad   <- file.path(          # Absoluter Pfad zur Zieldatei im Abbildungsverzeichnis
  abbildungsordner,
  "pds_7_ah_prebdi.pdf"
)

# Speichern der Abbildung als PDF
dev.copy2pdf(                            # PDF Kopierfunktion
  file   = abb_ausgabepfad,              # Dateiname
  width  = 7,                            # Breite (inch)
  height = 4                             # Höhe (inch)
)
```

Absolute Häufigkeitsverteilung aller Pre-BDI Werte

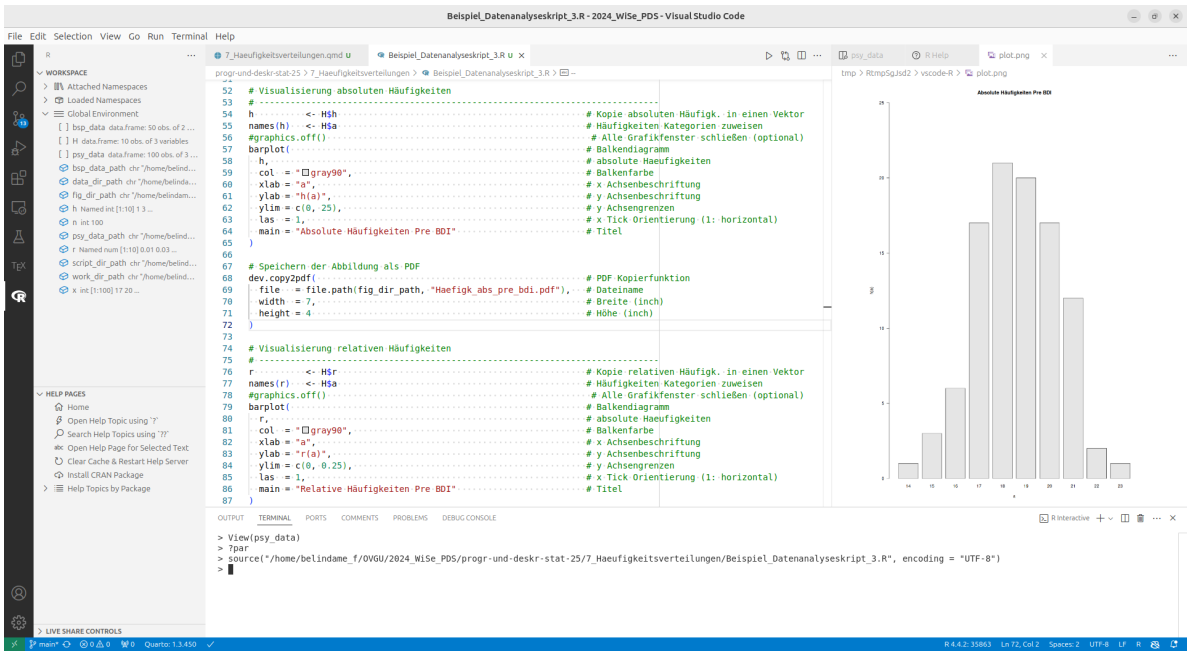


Relative Häufigkeitsverteilung aller Pre-BDI Werte



Der base VSCode-R plot viewer (`png` device)

Wenn Visualisierungen in R erstellt werden, öffnet sich standardmäßig ein Grafikfenster (graphical device). Dieses Fenster wird verwendet, um die Plots anzuzeigen.



In Theorie können Grafikfenster mit `dev.new()` neu geöffnet und mit `dev.off()` geschlossen werden. Allerdings ist dieses Verhalten in VS Code nicht immer konsistent. Mit `graphics.off()` können jedoch alle noch offenen Grafikfenster geschlossen werden, bevor eine neue erstellt werden soll.

Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Definition (Histogramm)

Ein *Histogramm* ist ein Diagramm, in dem zu einem Datensatz $y = (y_1, ..., y_n)$ mit Werten $W := \{w_1, ..., w_m\}$ mit $m \leq n$ für $j = 1, ..., k$ über benachbarten Intervallen $[b_{j-1}, b_j[$ mit $b_0 := \min W$ und $b_k := \max W$ Rechtecke mit der

Breite $d_j := b_j - b_{j-1}$ und der
Höhe $h(w)$ oder $r(w)$ für $w \in [b_{j-1}, b_j[$

abgebildet sind. Dabei werden die Intervalle $[b_{j-1}, b_j[$ *Klassen* (engl. *bins*) genannt.

Bemerkungen

- Das Aussehen eines Histogramms ist stark von der Anzahl k der Klassen abhängig.
- Im Folgenden sind einige konventionelle Methoden zur Bestimmung von k aufgeführt ($\lceil \cdot \rceil$ ist die Aufrundungsfunktion).

$k := \lceil \frac{b_k - b_0}{d} \rceil$	Bestimmung der Anzahl der Klassen aus gewünschter Klassenbreite d
$k := \lceil \sqrt{n} \rceil$	Klassenanzahl in Excel
$k := \lceil \log_2 n + 1 \rceil$	Implizite Normalverteilungsannahme (Sturges, 1926)
$d := 3.49 \cdot \frac{S}{\sqrt[3]{n}}$	Berechnung d nach Scott (1979): Min. MSE Dichteschätzung bei Normalverteilung

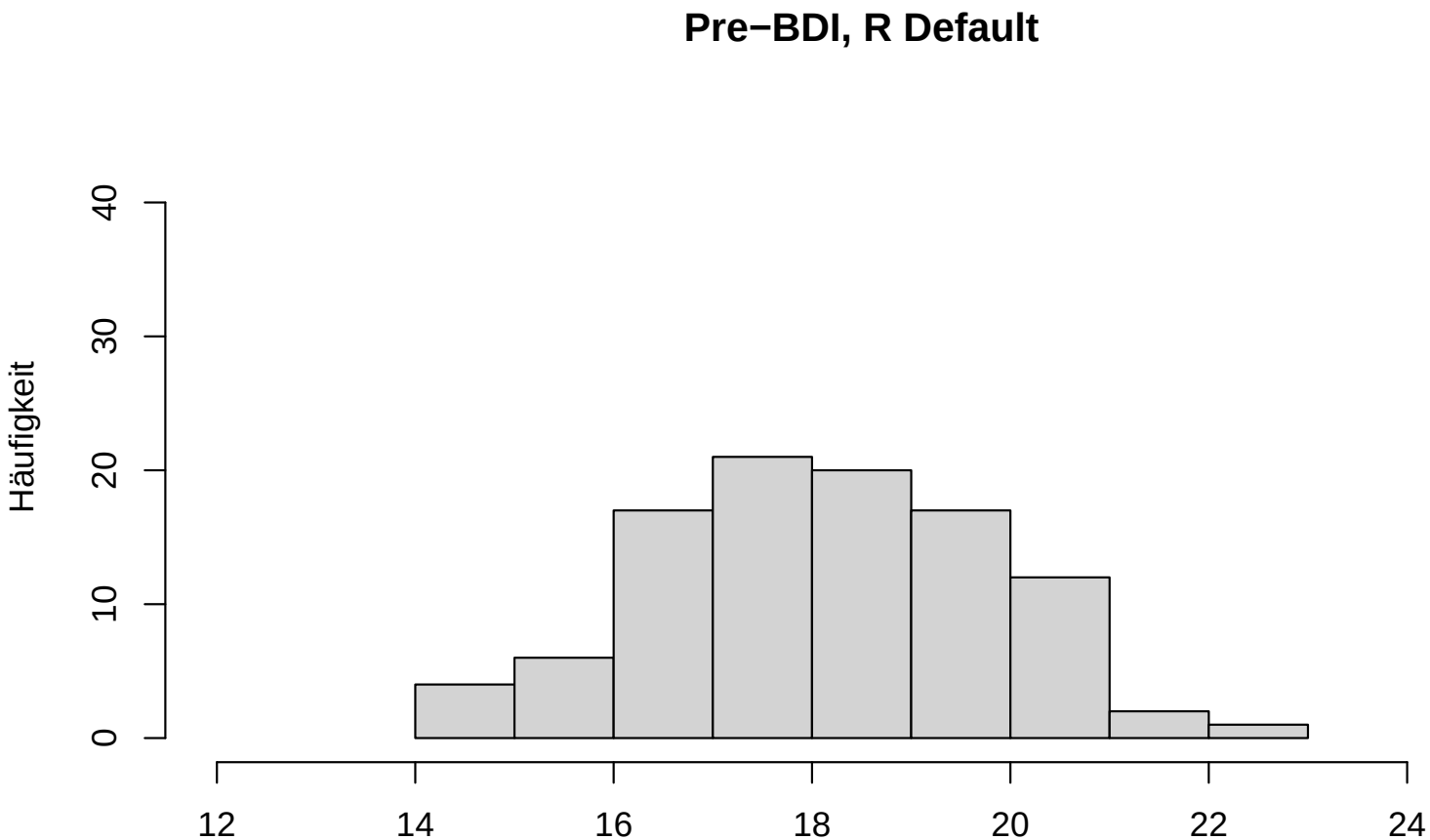
Berechnung und Visualisierung von Histogrammen mit `hist()`

- Die Klassen $[b_{j-1}, b_j[, j = 1, ..., k$ werden als Argument `breaks` festgelegt
- `breaks` ist der R Vector c$(b_0, b_1, ..., b_k)$ 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

```r
# Default Histogramm
y     <- 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
  y,                          # 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
)
```

Histogramm mit `hist()` und default `breaks` - Beispiel

Visualisierung



Histogramm mit `hist()` und default `breaks` - Beispiel

Ausgabe der Ergebnisse

```r
# Histogrammdaten berechnen, ohne Plot auszugeben
histogram_data <- hist(        # Speichert die Ergebnisse von hist()
  y,                           # Datensatz
  plot = FALSE                 # Verhindert die direkte Darstellung des Plots
)

print(histogram_data)          # 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] "y"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

Histogramme mit alternativen Klassengrößen

Berechnung von Klassenanzahlen und `breaks` Argument

```
# Histogramm mit gewuenschter Klassenbreite
d   <- 1                          # gewuenschte Klassenbreite
b_0 <- min(y)                     # b_0
b_k <- max(y)                     # b_k
k   <- ceiling((b_k - b_0) / d)   # Anzahl der Klassen
b   <- seq(b_0, b_k, len = k + 1) # Klassen [b_{j-1}, b_j[

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

# Sturges
n   <- length(y)                  # 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[
d   <- b[2] - b[1]                # Klassenbreite

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

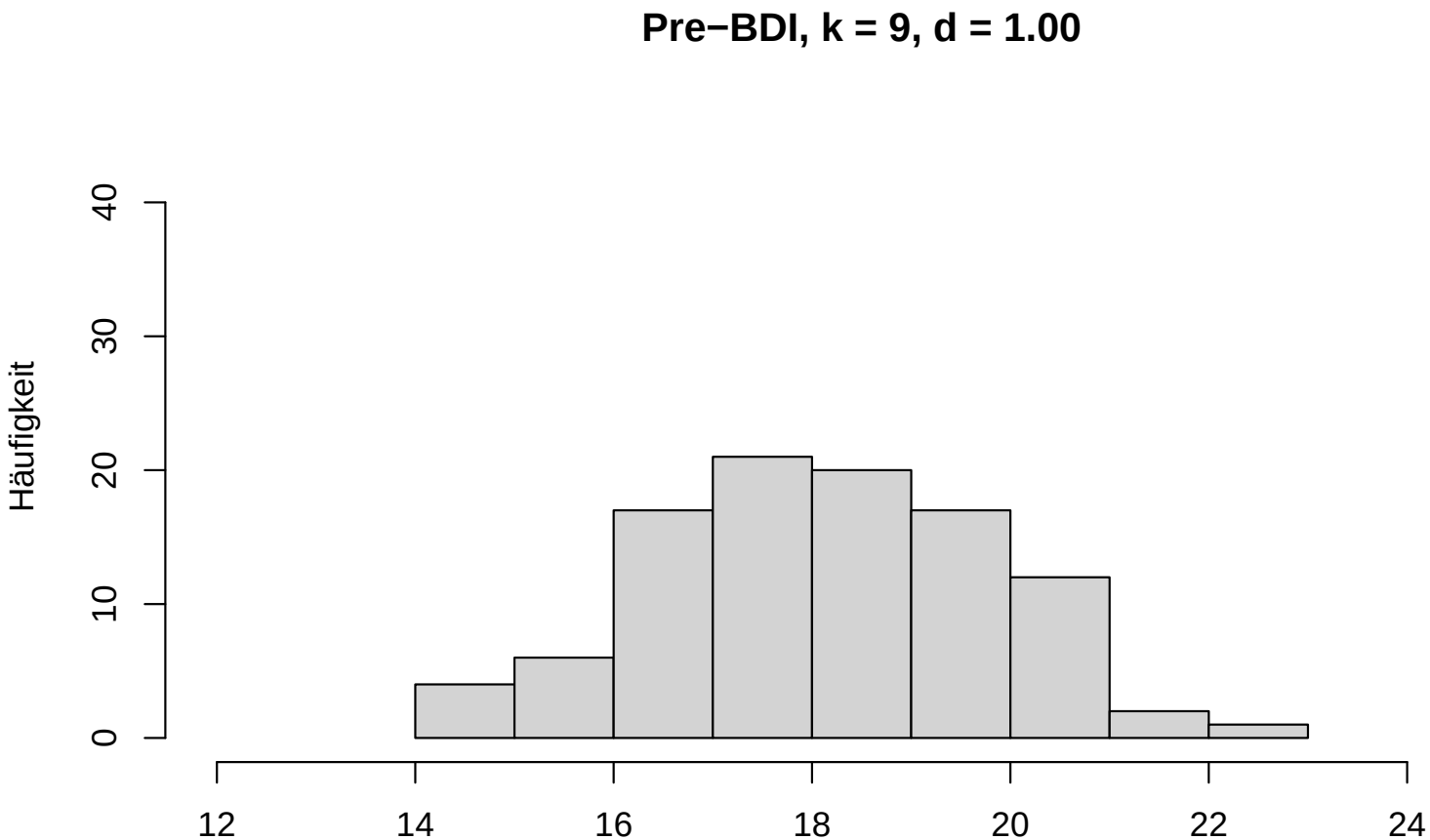
Berechnung und Visualisierung von Histogrammen mit `hist()`

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

```
# Default Histogramm
y      <- 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
  y,                                                         # Datensatz
  breaks = b,                                                # breaks
  xlim   = c(x_min, x_max),                                  # x Achsengrenzen
  ylim   = c(y_min, y_max),                                  # y Achsengrenzen
  ylab   = "Häufigkeit",                                     # y-Achsenbezeichnung
  xlab   = "",                                               # x-Achsenbezeichnung
  main   = sprintf("Pre-BDI, k = %.0f, d = %.2f", k, d)      # Titel
)
```

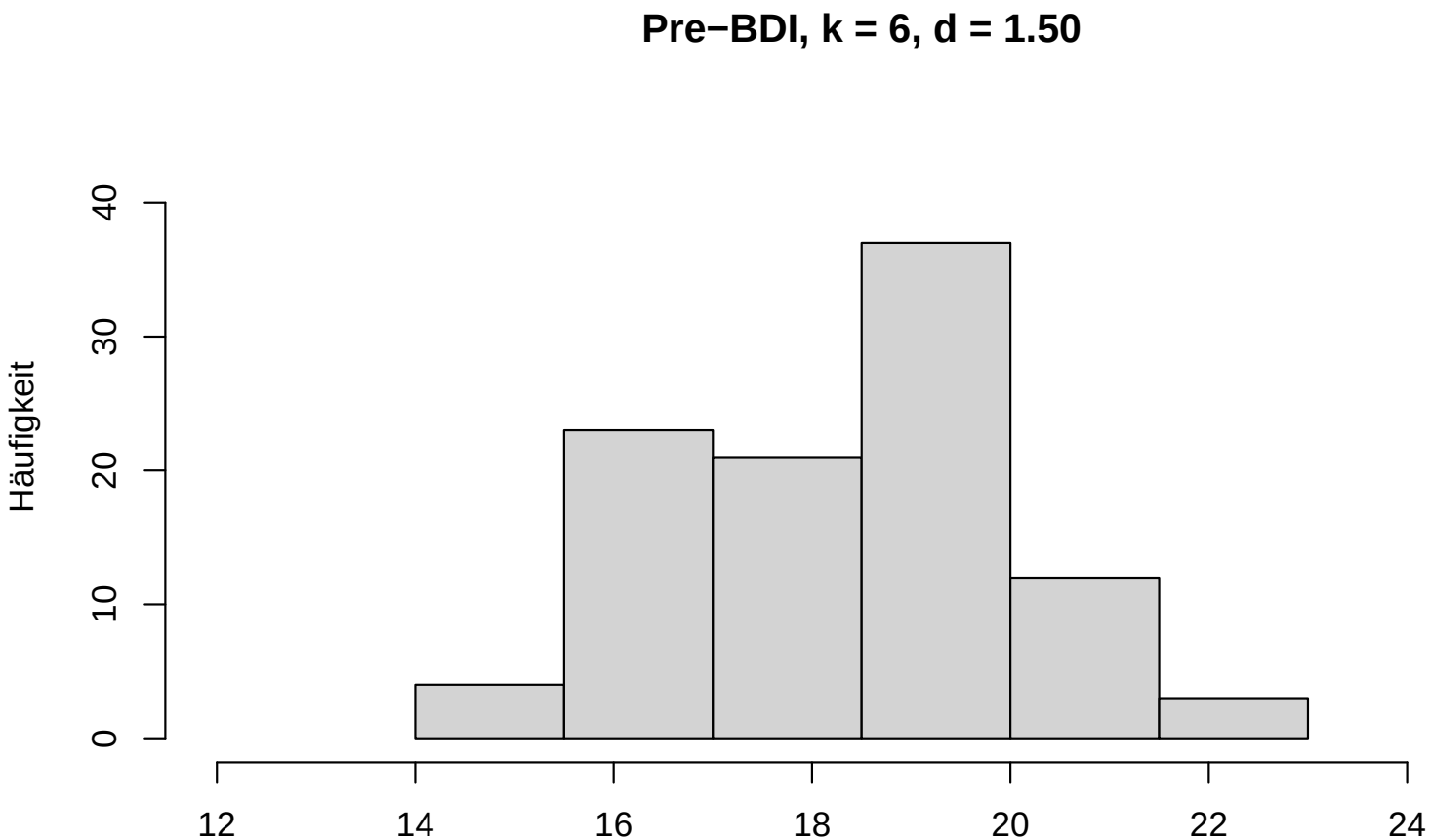
Histogramme - Beispiel

Gewünschte Klassenbreite $d := 1$



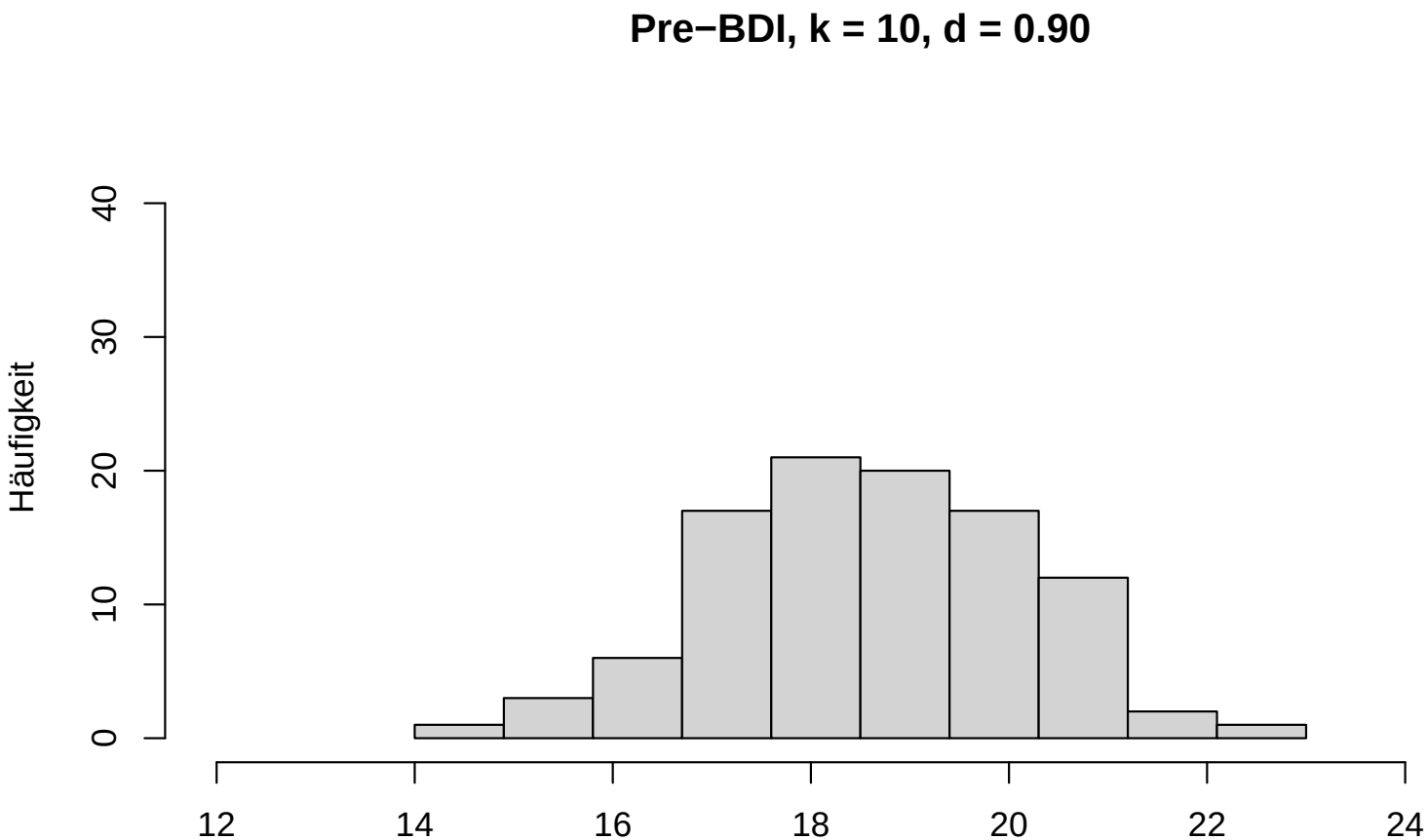
Histogramme - Beispiel

Gewünschte Klassenbreite $d := 1.5$



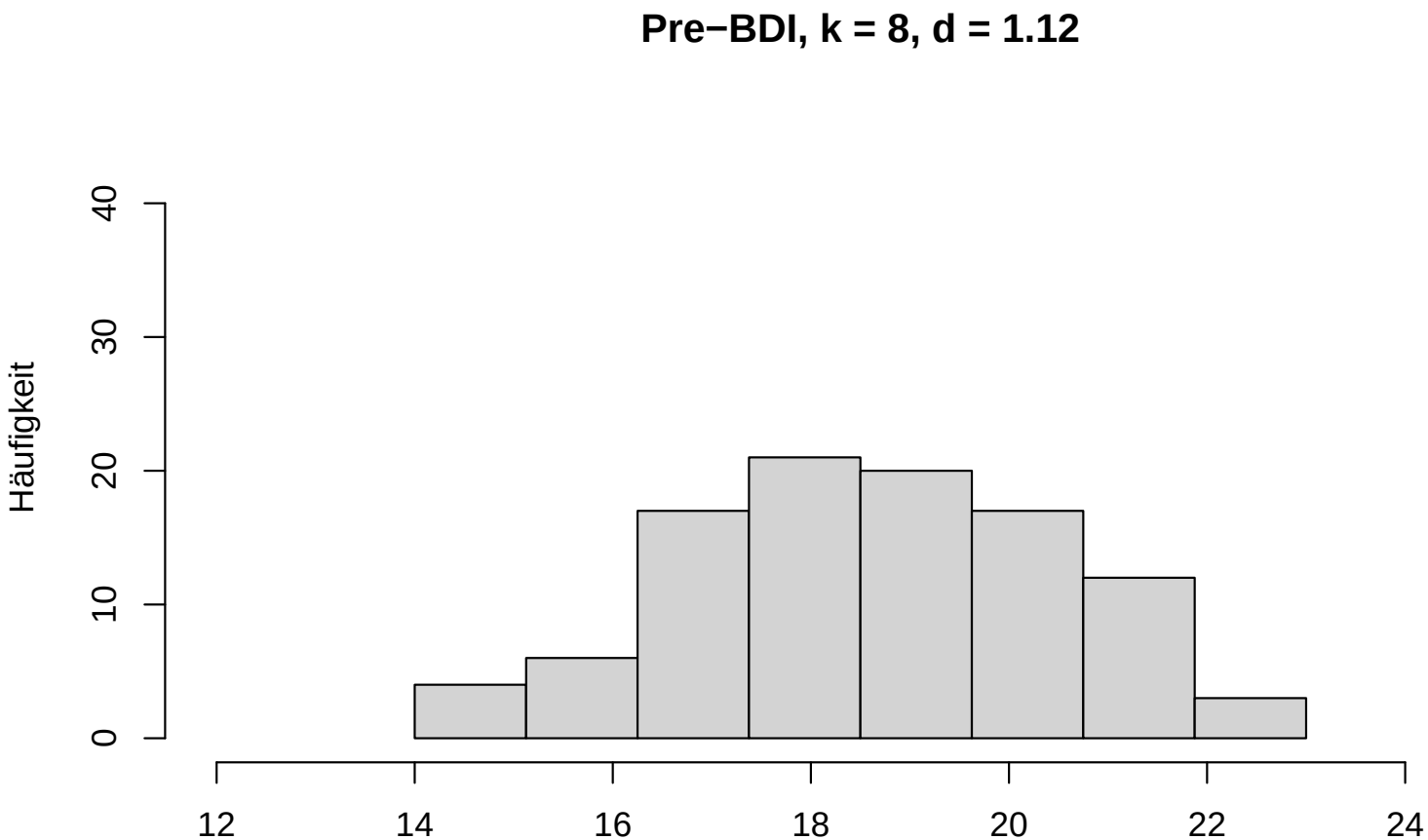
Histogramme - Beispiel

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



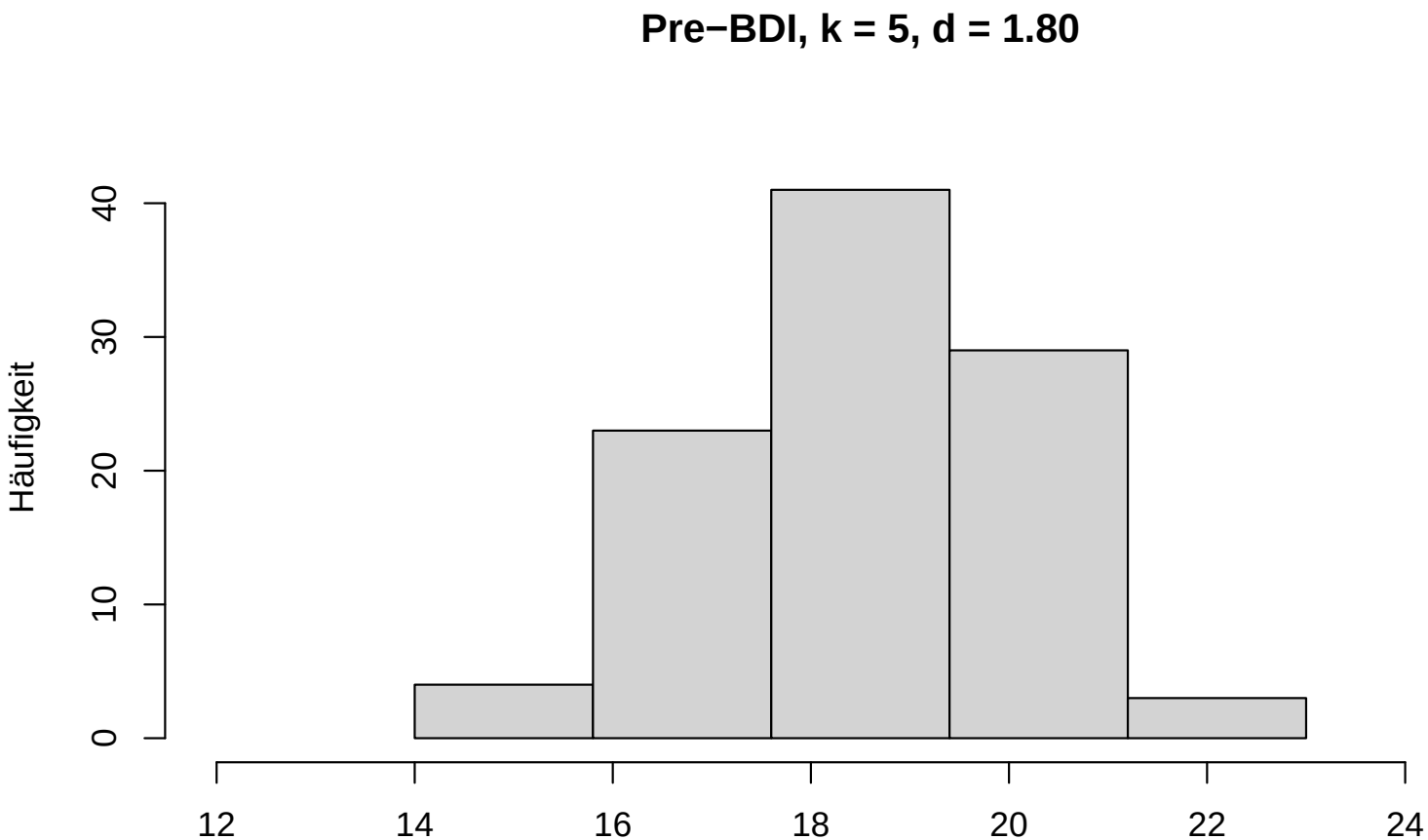
Histogramme - Beispiel

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



Histogramme - Beispiel

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



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Definition (Kumulative absolute und relative Häufigkeitsverteilungen)

$y = (y_1, ..., y_n)$ sei ein Datensatz, $W := \{w_1, ..., w_k\}$ mit $k \leq n$ die im Datensatz vorkommenden Werte und h und r die absoluten und relativen Häufigkeitsverteilungen der Werte von y, respektive. Dann heißt die Funktion

$$H : W \to \mathbb{N}, w \mapsto H(w) := \sum_{w' \leq w} h(w') \tag{4}$$

die *kumulative absolute Häufigkeitsverteilung* der Werte von y und die Funktion

$$R : W \to [0, 1], w \mapsto R(w) := \sum_{w' \leq w} r(w') \tag{5}$$

die *kumulative relative Häufigkeitsverteilung* der Werte von y.

Bemerkung

- Mit den Definitionen der absoluten und relativen Häufigkeitsverteilungen gilt also

$$H(w) = \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq w \tag{6}$$

und

$$R(w) = \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq w \text{ geteilt durch } n. \tag{7}$$

Evaluation kumulativer absoluter und relativer Häufigkeitsverteilungen

In R können kumulative Summen mit `cumsum()` berechnet werden.

Evaluation am Beispiel der Pre.BDI-Werte:

```
# Einlesen des Beispieldatensatzes und Abbildungsverzeichnisdefinition
psy_dateipfad <- file.path(datenordner, "psychotherapie_datensatz.csv")
D             <- read.csv(psy_dateipfad)

# Evaluation der absoluten und relativen Häufigkeitsverteilugen von Pre.BDI
y             <- D$Pre.BDI                  # Kopie der Pre.BDI Werte in einen double Vektor
n             <- length(y)                  # Anzahl der Datenwerte
H             <- as.data.frame(table(y))    # absolute Häufigkeitsverteilung als Dataframe
names(H)      <- c("w", "h")                # Benennen der Spalten im Dataframe
H$r           <- H$h / n                    # Neue Spalte mit relativen Häufigkeiten

# Evaluation der kumulativen absoluten und relativen Häufigkeitsverteilungen
H$H           <- cumsum(H$h)                # Neue Spalte mit kumulativen absoluten Häufigkeiten
H$R           <- cumsum(H$r)                # Neue Spalte mit kumulativen relativen Häufigkeiten
print(H)                                    # Ausgabe des Dataframes
```

```
    w  h    r   H    R
1  14  1 0.01   1 0.01
2  15  3 0.03   4 0.04
3  16  6 0.06  10 0.10
4  17 17 0.17  27 0.27
5  18 21 0.21  48 0.48
6  19 20 0.20  68 0.68
7  20 17 0.17  85 0.85
8  21 12 0.12  97 0.97
9  22  2 0.02  99 0.99
10 23  1 0.01 100 1.00
```

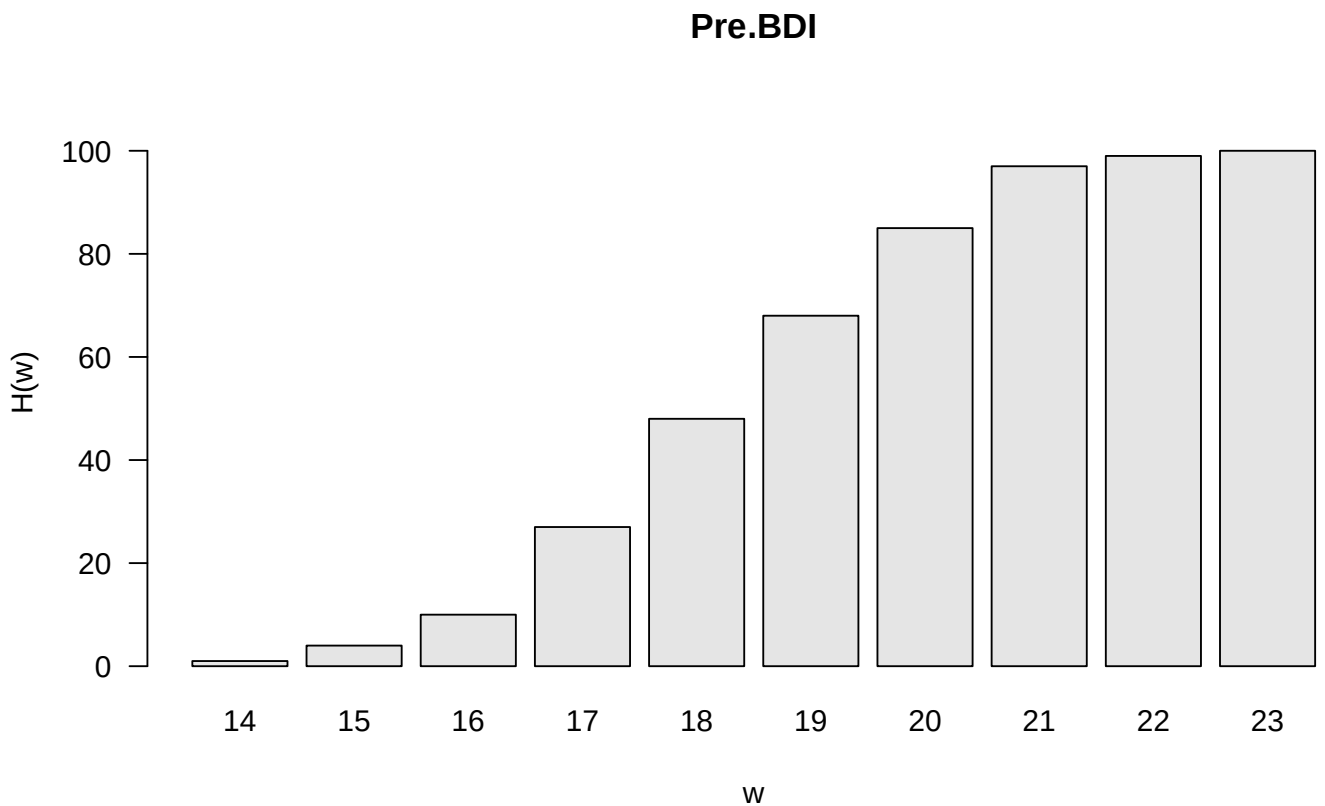
Visualisierung kumulativer Häufigkeiten

Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte

```r
# Vorbereitung der zu visualisierenden Daten
Hw        <- H$H                # Kopie der kumulativen absoluten Häufigkeiten in einen Vektor
names(Hw) <- H$w                # Den Häufigkeitswerten die entsprechenden Kategorien zuweisen

# Visualisierung der kumulativen absoluten Häufigkeitsverteilung
barplot(                        # Balkendiagramm
  Hw,                           # H(w) Werte (kumulativen absoluten Häufigkeiten) als input
  col  = "gray90",              # Balkenfarbe
  xlab = "w",                   # x Achsenbeschriftung
  ylab = "H(w)",                # y Achsenbeschriftung
  ylim = c(0, 110),             # y Achsenlimits
  las  = 1,                     # Achsenticklabelorientierung (1: horizontal)
  main = "Pre.BDI"              # Titel
)

# PDF Speicherung
dev.copy2pdf(
  file   = file.path(abbildungsordner, "pds_8_kh.pdf"),
  width  = 8,
  height = 5
)
```

Visualisierung kumulativer Häufigkeiten

Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte



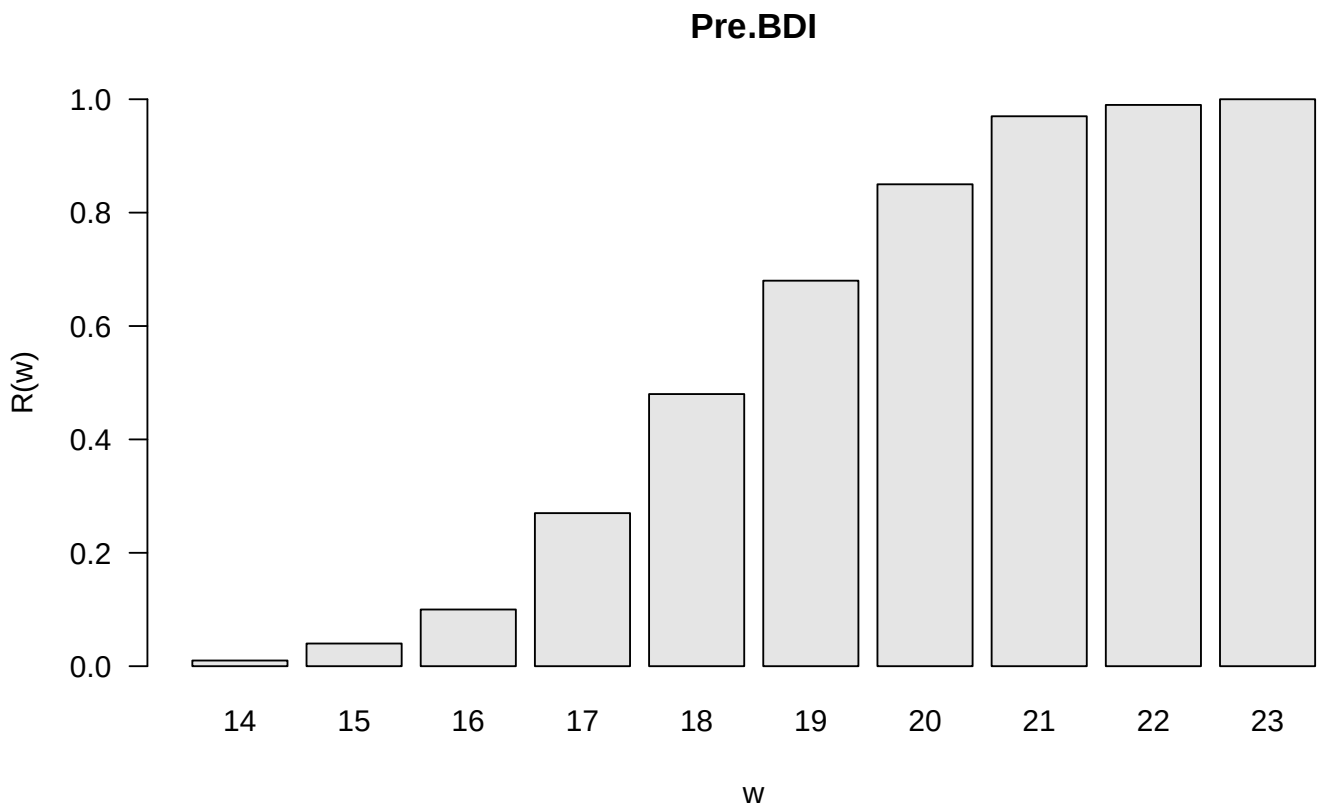
Visualisierung kumulativer Häufigkeiten

Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte

```r
# Vorbereitung der zu visualisierenden Daten
Rw        <- H$R            # R(w) Werte
names(Rw) <- H$w            # barplot braucht w Werte als names

# Visualisierung der kumulativen relativen Häufigkeitsverteilung
barplot(                    # Balkendiagramm
  Rw,                       # R(w) Werte
  col  = "gray90",          # Balkenfarbe
  xlab = "w",               # x Achsenbeschriftung
  ylab = "R(w)",            # y Achsenbeschriftung
  ylim = c(0, 1),           # y Achsenlimits
  las  = 1,                 # Achsenticklabelorientierung
  main = "Pre.BDI"          # Titel
)

# PDF Speicherung
dev.copy2pdf(
  file   = file.path(abbildungsordner, "pds_8_kr.pdf"),
  width  = 8,
  height = 5
)
```

Visualisierung kumulativer Häufigkeiten

Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte



Definition (Empirische Verteilungsfunktion)

$y = (y_1, ..., y_n)$ sei ein Datensatz. Dann heißt die Funktion

$$F : \mathbb{R} \to [0, 1], x \mapsto F(x) := \frac{\text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq x}{n} \tag{8}$$

die empirische Verteilungsfunktion (EVF) von y.

Bemerkungen

- Die empirische Verteilungsfunktion wird auch *empirische kumulative Verteilungsfunktion* genannt.
- Die Definitionsmenge der EVF ist im Gegensatz zu Häufigkeitsverteilungen $\mathbb{R}$ und nicht W
- Die EVF verhält sich zu kumulativen Häufigkeitsverteilungen wie Histogramme zu Häufigkeitsverteilungen.
- Typischerweise sind empirische Verteilungsfunktionen Treppenfunktionen.
- Die (visuelle) Umkehrfunktion der EVF kann zur Bestimmung von Quantilen genutzt werden.

Evaluation und Visualisierung der EVF mit `ecdf()`

Die Funktion `ecdf()` berechnet die empirische Verteilungsfunktion (EVF) eines Datensatzes.

Mit der Funktion `plot()` lässt sich die EVF, die von `ecdf()` erzeugt wurde, direkt visualisieren.

Empirische Verteilungsfunktion am Beispiel der Pre.BDI Werte

```
# Evaluation der EVF
y    <- D$Pre.BDI                                  # double vector der Pre.BDI Werte
evf  <- ecdf(y)                                    # Evaluation der EVF

# Evaluation der Funktionswerte der EVF
p_bis_17 <- evf(17)                                # Anteil der Werte <= 17
p_bis_19 <- evf(19)                                # Anteil der Werte <= 19

# Ausgabe der Ergebnisse
cat("Anteil der Werte kleiner oder gleich:\n",
    sprintf("17: %.2f\n", p_bis_17),
    sprintf("19: %.2f\n", p_bis_19))
```

```
Anteil der Werte kleiner oder gleich:
 17: 0.27
 19: 0.68
```

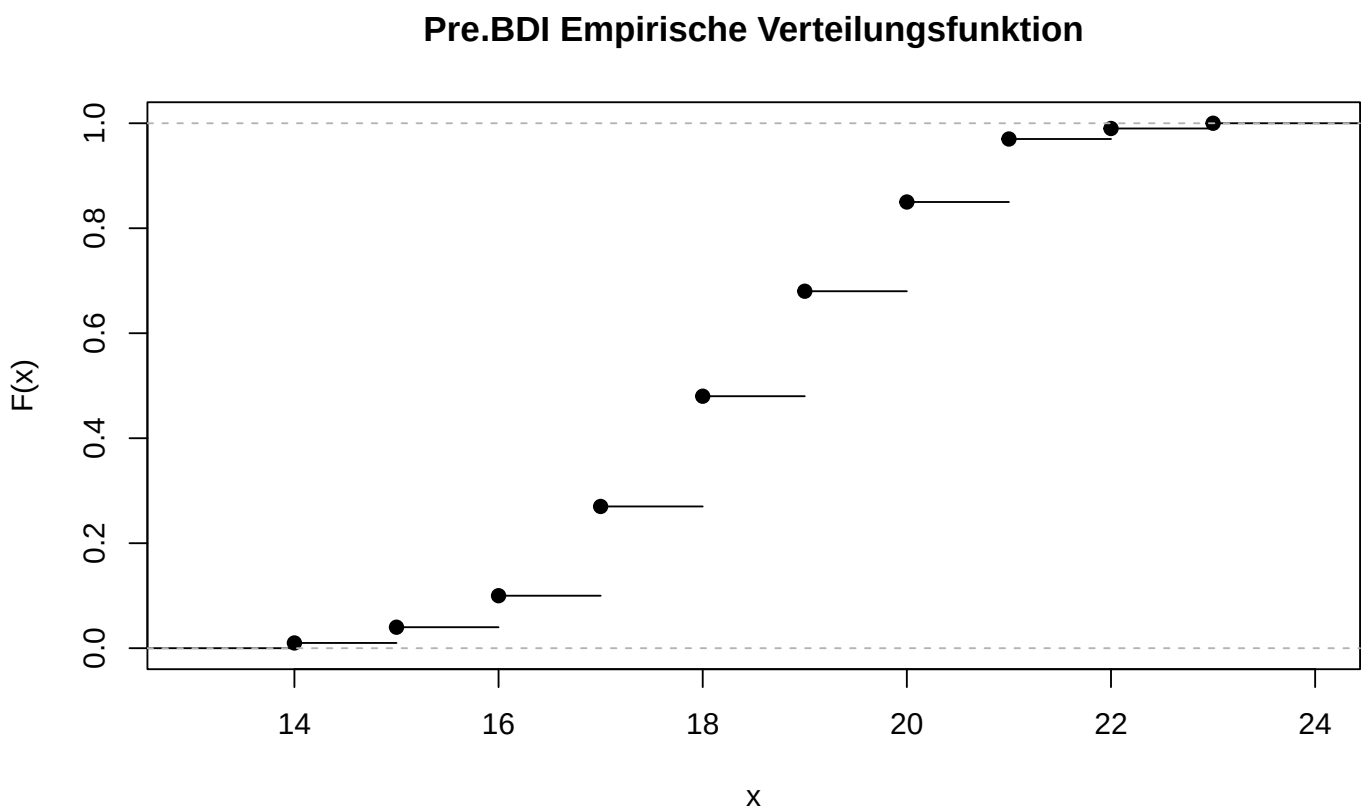
Evaluation und Visualisierung der EVF mit `ecdf()`

Mit der Funktion `plot()` lässt sich die EVF, die von `ecdf()` erzeugt wurde, direkt visualisieren.

Empirische Verteilungsfunktion am Beispiel der Pre.BDI Werte

```
# Visualisierung der kumulativen relativen Häufigkeitsverteilung
plot(
  evf,                                          # ecdf Objekt
  xlab = "x",                                   # x Achsenbeschriftung
  ylab = "F(x)",                                # y Achsenbeschriftung
  main = "Pre.BDI Empirische Verteilungsfunktion" # Titel
)

# PDF Speicherung
dev.copy2pdf(
  file   = file.path(abbildungsordner, "pds_8_ecdf.pdf"),
  width  = 8,
  height = 5
)
```

Empirische Verteilungsfunktion der Pre.BDI Werte



Motivation

Beispieldatensatz

Häufigkeitsverteilungen

Histogramme

Empirische Verteilungsfunktionen

Quantile und Boxplots

Definition (p-Quantil)

$y = (y_1, ..., y_n)$ sei ein Datensatz und

$$y_s = \left(y_{(1)}, y_{(2)}, ..., y_{(n)}\right) \text{ mit } \min_{1 \leq i \leq n} y_i = y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} = \max_{1 \leq i \leq n} y_i \tag{9}$$

sei der zugehörige aufsteigend sortierte Datensatz. Weiterhin bezeichne $\lfloor \cdot \rfloor$ die Abrundungsfunktion. Dann heißt für ein $p \in [0, 1]$ die Zahl

$$y_p := \begin{cases} y_{(\lfloor np+1 \rfloor)} & \text{falls } np \neq \mathbb{N} \\ \frac{1}{2}\left(y_{(np)} + y_{(np+1)}\right) & \text{falls } np \in \mathbb{N} \end{cases} \tag{10}$$

das p-*Quantil* des Datensatzes y.

Bemerkungen

- Mindestens $p \cdot 100\%$ aller Werte in y sind kleiner oder gleich y_p.
- Mindestens $(1-p) \cdot 100\%$ aller Werte in y sind größer als y_p .
- Das p-Quantil teilt den geordneten Datensatz im Verhältnis p zu $(1-p)$ auf.
- $y_{0.25}, y_{0.50}, y_{0.75}$ heißen *unteres Quartil*, *Median*, und *oberes Quartil*, respektive.
- $y_{j \cdot 0.10}$ für $j = 1, ..., 9$ heißen *Dezile*,
- $y_{j \cdot 0.01}$ für $j = 1, ..., 99$ heißen *Percentile*.

Beispiel p-Quantil

Datensatz und sortierter Datensatz

i	1	2	3	4	5	6	7	8	9	10
y_i	8.5	1.5	75	4.5	6.0	3.0	3.0	2.5	6.0	9.0
$y_{(i)}$	1.5	2.5	3.0	3.0	4.5	6.0	6.0	8.5	9.0	75

0.25-Quantil

Es ist $n = 10$ und es sei $p := 0.25$. Dann gilt $np = 10 \cdot 0.25 = 2.5 \notin \mathbb{N}$. Also folgt

$$y_{0.25} = y_{(\lfloor 2.5+1 \rfloor)} = y_{(3)} = 3.0 \tag{11}$$

0.80-Quantil

Es ist $n = 10$ und es sei $p := 0.80$. Dann gilt $np = 10 \cdot 0.80 = 8 \in \mathbb{N}$. Also folgt

$$y_{0.80} = \frac{1}{2}\left(y_{(8)} + y_{(8+1)}\right) = \frac{1}{2}\left(y_{(8)} + y_{(9)}\right) = \frac{8.5 + 9.0}{2} = 8.75. \tag{12}$$

(Henze 2018, Kapitel 5)

Quantilsbestimmung in R

“Manuelle” Quantilbestimmung anhand obiger Definition

```
y    <- c(8.5, 1.5, 12, 4.5, 6.0, 3.0, 3.0, 2.5, 6.0, 9.0)  # Beispieldaten (Henze, 2018)
n    <- length(y)                                            # Anzahl Datenwerte
y_s  <- sort(y)                                              # sortierter Datensatz
p    <- 0.25                                                 # Quantilsparameter
print(n * p)                                                 # np ist hier keine natürliche Zahl
```

```
[1] 2.5
```

```
y_p  <- y_s[floor(n * p + 1)]                                # 0.25 Quantil
cat("0.25 Quantil: ", round(y_p,2))                          # Ausgabe
```

```
0.25 Quantil:  3
```

```
p    <- 0.80                                                 # Quantilsparameter
print(n * p)                                                 # np hier eine natürliche Zahl
```

```
[1] 8
```

```
y_p  <- (1 / 2) * (y_s[n * p] + y_s[n * p + 1])              # 0.80 Quantil
cat("0.80 Quantil: ", round(y_p,2))                          # Ausgabe
```

```
0.80 Quantil:  8.75
```

Quantilsbestimmung mithilfe vordefinierter R-Funktionen

`quantile()` wertet Quantile anhand der Quantildefinition `type` aus.

Es gibt mindestens neun verschiedene Quantildefinitionen (vgl. Hyndman and Fan 1996)

```
y_p <- quantile(y, 0.80, type = 1)                    # 0.80 Quantil, Definition 1
print(y_p)
```

```
80%
8.5
```

```
y_p <- quantile(y, 0.80, type = 2)                    # 0.80 Quantil, Definition 2
print(y_p)
```

```
 80%
8.75
```

Visuelle Bestimmung des Quantils

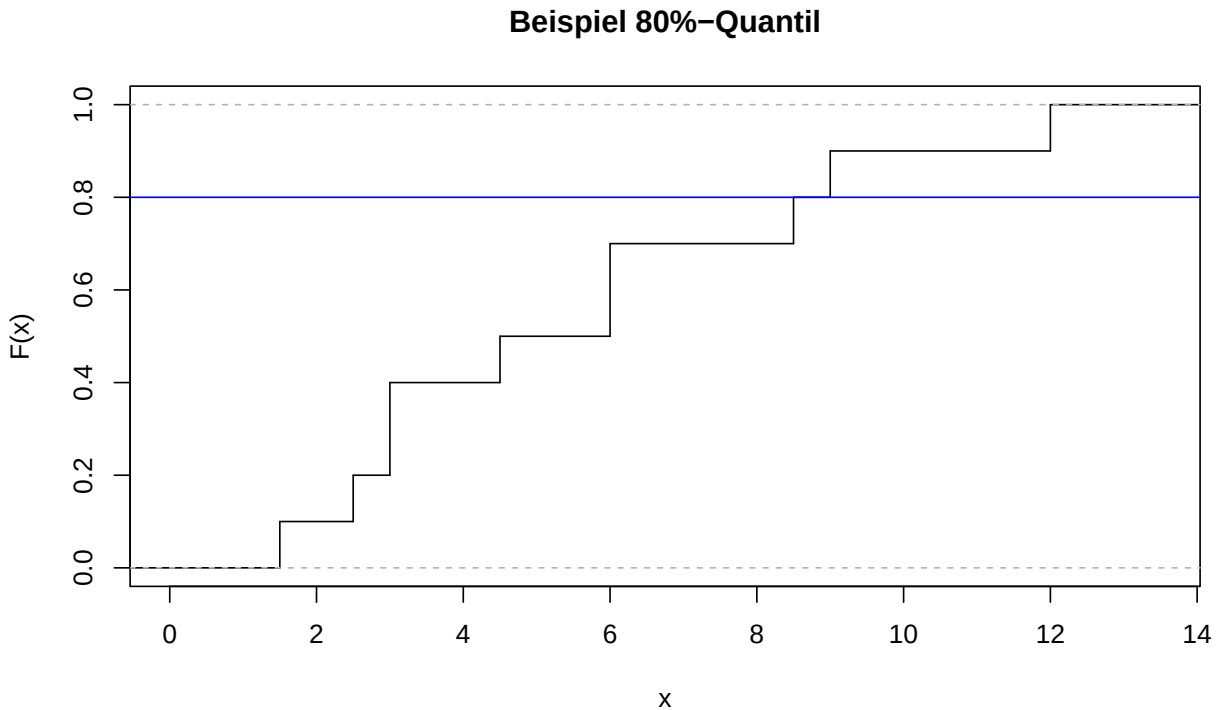
Kombination von `ecdf()`und `abline()` erlaubt prinzipiell die visuelle Bestimmung von Quantilen

EVF und 80%-Quantil der Beispieldaten aus Henze, 2018

```r
# Daten vorbereiten
y    <- c(8.5, 1.5, 12, 4.5, 6.0, 3.0, 3.0, 2.5, 6.0, 9.0)  # Beispieldaten (Henze, 2018)
evf <- ecdf(y)                                               # Evaluation der EVF

# Visualisierung der empirischen Verteilungsfunktion
plot(                                                        # plot() weiß mit ecdf object umzugehen
  evf,                                                       # ecdf Objekt
  xlab      = "x",                                           # x Achsenbeschriftung
  ylab      = "F(x)",                                        # y Achsenbeschriftung
  verticals = TRUE,                                          # vertikale Linien
  do.points = FALSE,                                         # keine Punkte
  main      = "Beispiel 80%-Quantil"                         # Titel
)

# Visualisierung einer Linie auf Höhe des Funktionswertes 0.80
abline(                                                      # horizontale Linie
  h   = 0.80,                                                # y Ordinate der Linie
  col = "blue"                                               # Farbe der Linie
)

# PDF Speicherung
dev.copy2pdf(
  file   = file.path(abbildungsordner, "pds_8_ecdf_abline_x.pdf"),
  width  = 8,
  height = 5
)
```

Visuelle Bestimmung des Quantils

EVF und 80%-Quantil der Beispieldaten aus Henze, 2018




```r
y_p_1 <- quantile(y, 0.80, type = 1)  # 0.80 Quantil, Definition 1
y_p_2 <- quantile(y, 0.80, type = 2)  # 0.80 Quantil, Definition 2
cat("0.80 Quantil mit type=1 bestimmt: ",y_p_1,
    "\n0.80 Quantil mit type=2 bestimmt: ", y_p_2)
```

```
0.80 Quantil mit type=1 bestimmt:  8.5
0.80 Quantil mit type=2 bestimmt:  8.75
```

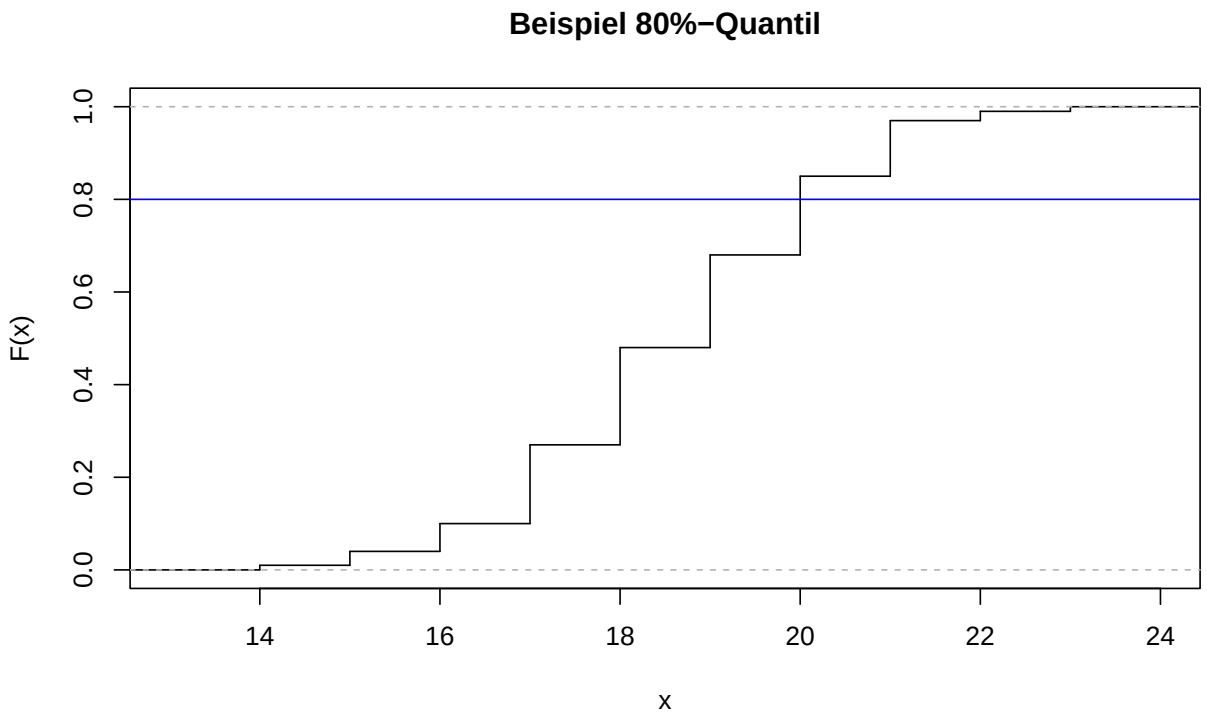
Visuelle Bestimmung des Quantils

EVF und 80%-Quantil der Pre.BDI-Werte aus dem Psychotherpiedatensatz

```
# Daten vorbereiten
y   <- D$Pre.BDI                          # Double vector der Pre.BDI Werte
evf <- ecdf(y)                            # Evaluation der EVF

# Visualisierung der empirischen Verteilungsfunktion
plot(                                     # plot() kann mit ecdf object umgehen
  evf,                                    # ecdf Objekt
  xlab      = "x",                        # x Achsenbeschriftung
  ylab      = "F(x)",                     # y Achsenbeschriftung
  verticals = TRUE,                       # vertikale Linien
  do.points = FALSE,                      # keine Punkte
  main      = "Beispiel 80%-Quantil"      # Titel
)

# Visualisierung einer Linie auf Höhe des Funktionswertes 0.80
abline(                                   # horizontale Linie
  h   = 0.80,                             # y Ordinate der Linie
  col = "blue"                            # Farbe der Linie
)

# PDF Speicherung
dev.copy2pdf(
  file   = file.path(abbildungsordner, "pds_8_ecdf_abline_prebdi.pdf"),
  width  = 8,
  height = 5
)
```

Visuelle Bestimmung des Quantils

EVF und 80%-Quantil der Pre.BDI-Werte aus dem Psychotherpiedatensatz




```
y_p_1 <- quantile(D$Pre.BDI, 0.80, type = 1)  # 0.80 Quantil, Definition 1
y_p_2 <- quantile(D$Pre.BDI, 0.80, type = 2)  # 0.80 Quantil, Definition 2
cat("0.80 Quantil mit type=1 bestimmt: ",y_p_1,
    "\n0.80 Quantil mit type=2 bestimmt: ", y_p_2)
```

```
0.80 Quantil mit type=1 bestimmt:  20
0.80 Quantil mit type=2 bestimmt:  20
```

Boxplots

Ein Boxplot visualisiert eine Quantil-basierte Zusammenfassung eines Datensatzes.

Typischerweise werden $\min y, y_{0.25}, y_{0.50}, y_{0.75}$ und $\max y$ visualisiert.

- $\min y$ und $\max y$ werden oft als "Whiskerendpunkte" dargestellt.
- $y_{0.25}$ und $y_{0.75}$ sind untere und obere Grenze der zentralen grauen Box.
- $y_{0.50}$ wird als Strich in der zentralen grauen Box abgebildet.

$d_Q := y_{0.75} - y_{0.25}$ heißt *Interquartilsabstand* und dient als Verteilungsbreitenmaß

`summary()` liefert wesentliche Kennzahlen

```
# Sechswertezusammenfassung
summary(D$Pre.BDI)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   17.00   19.00   18.61   20.00   23.00
```

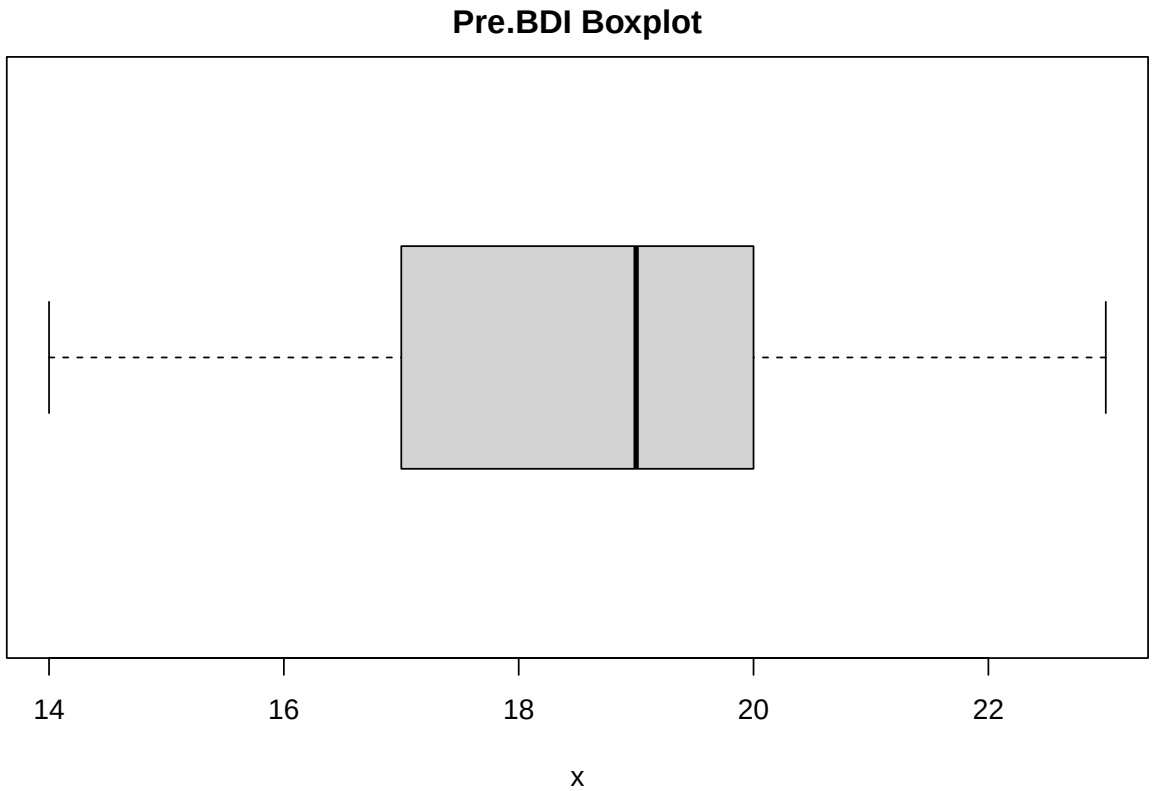
Erstellen von Boxplots in R

`boxplot()` erstellt einen Boxplot

Boxplot der Pre.BDI-Werte

```
# Boxplot erstellen
boxplot(                              # Boxplot
  D$Pre.BDI,                          # Datensatz
  horizontal  = TRUE,                 # horizontale Darstellung
  range       = 0,                    # Whiskers bis zu den min(y) und max(y)-Werten
  xlab        = "y",                  # x Achsenbeschriftung
  main        = "Pre.BDI Boxplot"     # Titel
)

# PDF Speicherung
dev.copy2pdf(
  file        = file.path(abbildungsordner, "pds_8_boxplot_prebdi.pdf"),
  width       = 8,
  height      = 5
)
```

Boxplot der Pre.BDI-Werte



Es gibt viele Boxplotvariationen (vgl. McGill, Tukey, and Larsen 1978). Es sollte immer erläutert werden, welche Kennzahlen dargestellt werden!

Referenzen

Henze, Norbert. 2018. *Stochastik für Einsteiger*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-22044-0.

Hyndman, Rob J., and Yanan Fan. 1996. "Sample Quantiles in Statistical Packages." *The American Statistician* 50 (4): 361. https://doi.org/10.2307/2684934.

McGill, Robert, John W. Tukey, and Wayne A. Larsen. 1978. "Variations of Box Plots." *The American Statistician* 32 (1): 12. https://doi.org/10.2307/2683468.

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.