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

(8) Verteilungsfunktionen und Quantile

Empirische Verteilungsfunktionen

Quantile und Boxplots

Programmierübungen und Selbstkontrollfragen

Empirische Verteilungsfunktionen

Quantile und Boxplots

Programmierübungen und Selbstkontrollfragen

Kumulative Häufigkeitsverteilungen

Definition (Kumulative absolute und relative Häufigkeitsverteilungen)

$x = (x_1, ..., x_n)$ sei ein Datensatz, $A := \{a_1, ..., a_k\}$ mit $k \le n$ die im Datensatz vorkommenden verschiedenen Zahlenwerte und h und r die absoluten und relativen Häufigkeitsverteilungen von x, respektive. Dann heißt die Funktion

$$H : A \to \mathbb{N}, a \mapsto H(a) := \sum_{a' \le a} h(a') \tag{1}$$

die *kumulative absolute Häufigkeitsverteilung* von x und die Funktion

$$R : A \to [0, 1], a \mapsto R(a) := \sum_{a' \le a} r(a') \tag{2}$$

die *kumulative relative Häufigkeitsverteilung* der Zahlwerte von x.

Bemerkung

- Mit den Definitionen der absoluten und relativen Häufigkeitsverteilungen gilt also

$$H(a) = \text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \le a \tag{3}$$

und

$$R(a) = \text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \le a \text{ geteilt durch } n. \tag{4}$$

Evaluation kumulativer absoluter und relativer Häufigkeitsverteilungen

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

Evaluation am Beispiel der Pre.BDI-Werte:

```r
# Einlesen des Beispieldatensatzes und Abbildungsverzeichnisdefinition
psy_data_path <- file.path(data_dir_path, "psychotherapie_datensatz.csv")
D             <- read.table(psy_data_path, sep = ",", header = T)

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

```
    a  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
```

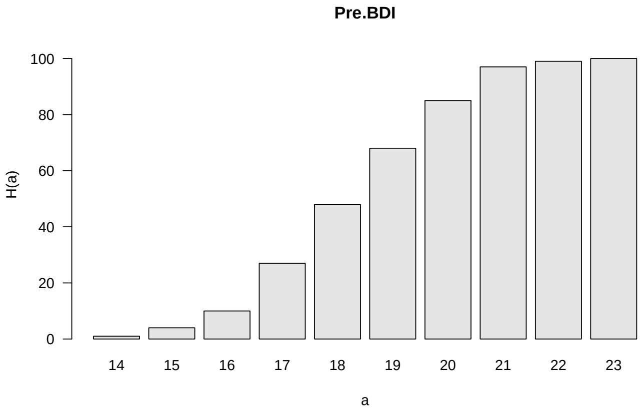
Visualisierung kumulativer Häufigkeiten

Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte

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

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

# PDF Speicherung
fig_dir_path <- file.path( # Pfad zum Output-Ordner für Abbildungen
  dirname(getwd()),          # Auf gleicher Ebene wie der Skriptordner
  "VL_Abbildungen"
)
dev.copy2pdf(
  file   = file.path(fig_dir_path, "pds_8_kh.pdf"),
  width  = 8,
  height = 5
)
```

Visualisierung kumulativer Häufigkeiten

Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte



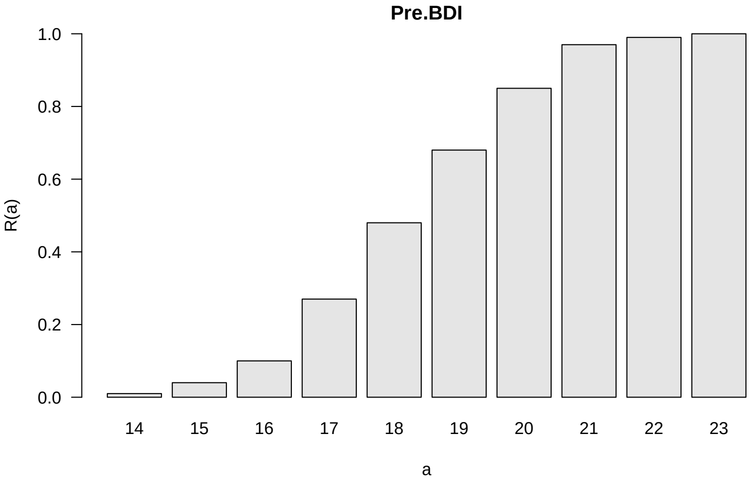
Visualisierung kumulativer Häufigkeiten

Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte

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

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

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

Visualisierung kumulativer Häufigkeiten

Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte



Empirische Verteilungsfunktionen

Definition (Empirische Verteilungsfunktion)

$x = (x_1, ..., x_n)$ sei ein Datensatz. Dann heißt die Funktion

$$F : \mathbb{R} \to [0,1], \xi \mapsto F(\xi) := \frac{\text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \le \xi}{n} \tag{5}$$

die empirische Verteilungsfunktion (EVF) von x.

Bemerkungen

- Die empirische Verteilungsfunktion wird auch *empirische kumulative Verteilungsfunktion* genannt.
- Die Definitionsmenge der EVF ist im Gegensatz zu Häufigkeitsverteilungen $\mathbb{R}$ und nicht A
- 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

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

# Evaluation der Funktionswerte der EVF
p_bis_17 <- evf(17)                                  # Anteil der Werte <= 18
p_bis_19 <- evf(19)                                  # Anteil der Werte <= 20

# 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
```

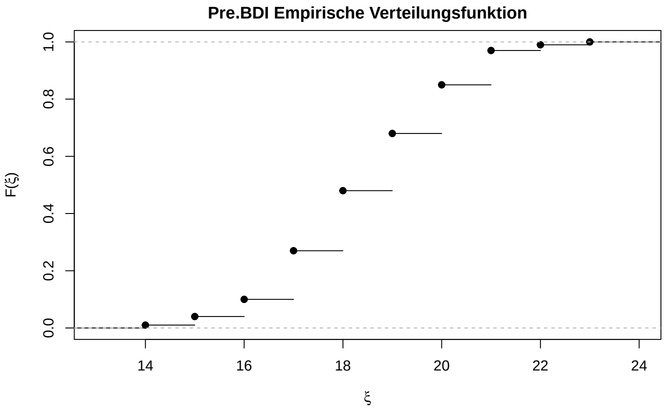
Evaluation und Visualisierung der EVF mit `ecdf()`

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

Empirische Verteilungsfunktion am Beispiel der Pre.BDI Werte

```r
library(latex2exp)                                 # LaTeX Formatierungstool, TeX()
# Visualisierung der kumulativen relativen Häufigkeitsverteilung
plot(
  evf,                                             # ecdf Objekt
  xlab = TeX("$\\xi$"),                            # x Achsenbeschriftung
  ylab = TeX("$F(\\xi)$"),                         # y Achsenbeschriftung
  main = "Pre.BDI Empirische Verteilungsfunktion" # Titel
)

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

Empirische Verteilungsfunktion der Pre.BDI Werte



Empirische Verteilungsfunktionen

Quantile und Boxplots

Programmierübungen und Selbstkontrollfragen

Definition (p-Quantil)

$x = (x_1, ..., x_n)$ sei ein Datensatz und

$$x_s = \left(x_{(1)}, x_{(2)}, ..., x_{(n)}\right) \text{ mit } \min_{1 \le i \le n} x_i = x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} = \max_{1 \le i \le n} x_i \tag{6}$$

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

$$x_p := \begin{cases} x_{(\lfloor np+1 \rfloor)} & \text{falls } np \ne \mathbb{N} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & \text{falls } np \in \mathbb{N} \end{cases} \tag{7}$$

das p-*Quantil* von x.

Bemerkungen

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

Beispiel p-Quantil

Datensatz und sortierter Datensatz

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

$$x_{0.25} = x_{(\lfloor 2.5+1 \rfloor)} = x_{(3)} = 3.0 \tag{8}$$

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

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

(Henze (2018), Kapitel 5)

Quantilsbestimmung in R

“Manuelle” Quantilbestimmung anhand obiger Definition

```r
x    <- 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(x)                                            # Anzahl Datenwerte
x_s  <- sort(x)                                              # sortierter Datensatz
p    <- 0.25                                                 # Quantilsparameter
print(n * p)                                                 # np ist hier keine natürliche Zahl
```

```
[1] 2.5
```

```r
x_p  <- x_s[floor(n * p + 1)]                                # 0.25 Quantil
print(x_p)                                                   # Ausgabe
```

```
[1] 3
```

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

```
[1] 8
```

```r
x_p  <- (1 / 2) * (x_s[n * p] + x_s[n * p + 1])              # 0.80 Quantil
print(x_p)                                                   # Ausgabe
```

```
[1] 8.75
```

Quantilsbestimmung in R

Quantilsbestimmung mithilfe vordefinierter R-Funktionen

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

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

```
x_p <- quantile(x, 0.80, type = 1)                          # 0.80 Quantil, Definition 1
print(x_p)
```

```
80%
8.5
```

```
x_p <- quantile(x, 0.80, type = 2)                          # 0.80 Quantil, Definition 2
print(x_p)
```

```
 80%
8.75
```

Visuelle Bestimmung des Quantils

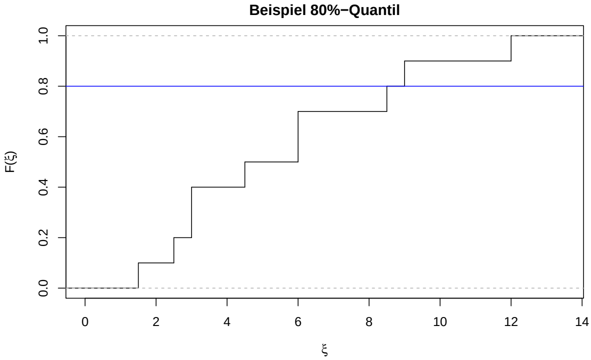
Kombination von `ecdf()`und `abline()` erlaubt prinzipiell die visuelle Bestimmung von Quantilen

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

```
library(latex2exp)                                                # LaTeX Formatierungstool, TeX()

# Daten vorbereiten
x    <- 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(x)                                                    # Evaluation der EVF

# Visualisierung der empirischen Verteilungsfunktion
plot(                                                             # plot() weiß mit ecdf object umzugehen
  evf,                                                            # ecdf Objekt
  xlab      = TeX("$\\xi$"),                                      # x Achsenbeschriftung
  ylab      = TeX("$F(\\xi)$"),                                   # 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(fig_dir_path, "pds_8_ecdf_abline_x.pdf"),
  width  = 8,
  height = 5
)
```

Visuelle Bestimmung des Quantils

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




```
x_p_1 <- quantile(x, 0.80, type = 1)  # 0.80 Quantil, Definition 1
x_p_2 <- quantile(x, 0.80, type = 2)  # 0.80 Quantil, Definition 2
cat("0.80 Quantil mit type=1 bestimmt: ",x_p_1,
    "\n0.80 Quantil mit type=2 bestimmt: ", x_p_2)
```

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

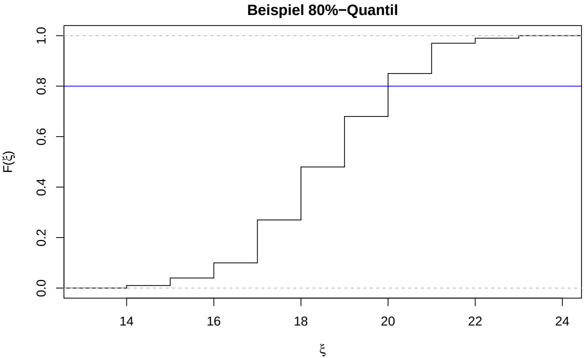
Visuelle Bestimmung des Quantils

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

```
library(latex2exp)                          # LaTeX Formatierungstool, TeX()

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

# Visualisierung der empirischen Verteilungsfunktion
plot(                                       # plot() kann mit ecdf object umgehen
  evf,                                      # ecdf Objekt
  xlab      = TeX("$\\xi$"),                # x Achsenbeschriftung
  ylab      = TeX("$F(\\xi)$"),             # 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(fig_dir_path, "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




```
x_p_1 <- quantile(D$Pre.BDI, 0.80, type = 1)  # 0.80 Quantil, Definition 1
x_p_2 <- quantile(D$Pre.BDI, 0.80, type = 2)  # 0.80 Quantil, Definition 2
cat("0.80 Quantil mit type=1 bestimmt: ",x_p_1,
    "\n0.80 Quantil mit type=2 bestimmt: ", x_p_2)
```

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

Ein Boxplot visualisiert eine Quantil-basierte Zusammenfassung eines Datensatzes.

Typischerweise werden $\min x, x_{0.25}, x_{0.50}, x_{0.75} und \max x$ visualisiert.

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

$d_Q := x_{0.75} - x_{0.25}$ 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

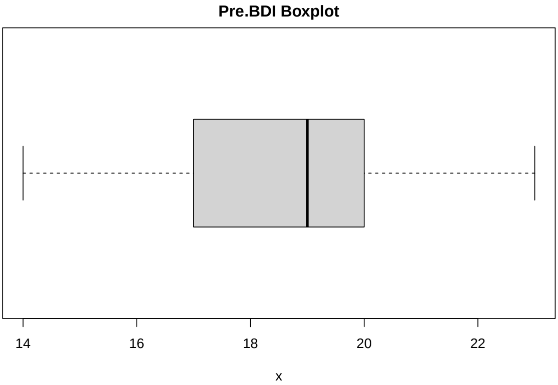
`boxplot()` erstellt einen Boxplot

Boxplot der Pre.BDI-Werte

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

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

Boxplots

Boxplot der Pre.BDI-Werte



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

Empirische Verteilungsfunktionen

Quantile und Boxplots

Programmierübungen und Selbstkontrollfragen

Programmierübungen

1. Erzeuge die kumulativen absoluten und relativen Häufigkeitsverteilungen der Post.BDI-Daten des Beispieldatensatzes *psychotherapie_datensatz.csv* und visualisiere diese.
2. Erzeuge und visualisiere die empirische Verteilungsfunktion der Post.BDI Daten.
3. Berechne das obere Quartil des Beispieldatensatzes auf Folie 17.
4. Berechne das untere Quartil, den Median und das obere Quartil der Post.BDI Daten und vergleiche deine Ergebnisse mit denen der `summary()`-Funktion.
5. Erstelle einen Boxplot der Post.BDI-Daten und kommentiere die Ergebnisse.

Selbstkontrollfragen

1. Definiere die Begriffe der kumulativen absoluten und relativen Häufigkeitsverteilungen.
2. Definiere den Begriff der empirischen Verteilungsfunktion.
3. Erläutere den Begriff des sortierten Datensatzes. Gibt dazu ein einfaches Beispiel mit drei Datenpunkten an.
4. Definiere den Begriff des p-Quantils.
5. Definiere die Begriffe unteres Quartil, Median und oberes Quartil unter Verwendung des p-Quantils.

References

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.