Computergestützte Datenanalyse

BSc Psychologie SoSe 2021

Prof. Dr. Dirk Ostwald

(6) Deskriptive Statistiken

Literaturhinweise

Fahrmeir et al. (2016, Kapitel 2) und Henze (2018, Kapitel 5) geben einen Überblick.

Deskriptive Statistiken

- Verteilungsdarstellung
- Maße der zentralen Tendenz
- Maße der Datenvariabilität
- Bivariate Deskriptivstatistik
- Übungen und Selbstkontrollfragen

Deskriptive Statistiken

- **Verteilungsdarstellung**
- Maße der zentralen Tendenz
- Maße der Datenvariabilität
- Bivariate Deskriptivstatistik
- Übungen und Selbstkontrollfragen

Der Affect Beispieldatensatz

- Daten zweier Studien ("maps" und "flat") des Personality, Motivation, and Cognition Laboratory mit $n = 330$ Versuchspersonen, Teil des psychTools R Pakets
- Fragebogenbasierte Messung von Affektvariablen vor und nach Anschauen eines Filmclips
- Filmclipbedingungen
  (1) Frontline, Dokumentation über die Befreiung des Konzentrationslagers Bergen-Belsen
  (2) Halloween, ein Horrorfilm
  (3) National Geographic, eine Naturdokumentation über die Serengeti
  (4) Parenthood, eine Komödie
- Messung von Tense Arousal (TA), Energetic Arousal (EA), Positive Affect (PA) und Negative Affect (NA) vor (1) und nach (2) Anschauen des Filmclips
- Erhebung von Daten mit fünf Skalen des Eysenck Persönlichkeitsinventar sowie State- und Trait-Anxiety Items. In der "maps" Studie außerdem Erhebung des BDI
- Für Details, siehe Rafaeli and Revelle (2006)

Der Affect Beispieldatensatz

```
install.package("psychTools")          # einmalige Installation des psychTools Pakets
library(psychTools)                    # Laden des psychTools Pakets
?affect                                # Erlaeuterung des affect Datensatzes
data(affect)                           # Laden des affect Dataframes
View(affect)                           # Inspektion des affect Dataframes
```

	Study	Film	ext	neur	imp	soc	lie	traitanx	state1	EA1	TA1	PA1	NA1	EA2	TA2	PA2	NA2	state2	MEQ	BDI
1	maps	3	18.0	9.0	7.0	10.0	3.0	24.0	22.0	24.0	14.0	26.0	2.0	6.0	5.0	7.0	4.0	NA	NA	0.04761905
2	maps	3	16.0	12.0	5.0	8.0	1.0	41.0	40.0	9.0	13.0	10.0	4.0	4.0	14.0	5.0	5.0	NA	NA	0.33333333
3	maps	3	6.0	5.0	3.0	1.0	2.0	37.0	44.0	1.0	14.0	4.0	2.0	2.0	15.0	3.0	1.0	NA	NA	0.19047619
4	maps	3	12.0	15.0	4.0	6.0	3.0	54.0	40.0	5.0	15.0	1.0	0.0	4.0	15.0	0.0	2.0	NA	NA	0.38461538
5	maps	3	14.0	2.0	5.0	6.0	3.0	39.0	67.0	12.0	20.0	7.0	13.0	14.0	15.0	16.0	13.0	NA	NA	0.38095238
6	maps	1	6.0	15.0	2.0	4.0	5.0	51.0	38.0	9.0	14.0	5.0	1.0	7.0	12.0	2.0	2.0	NA	NA	0.23809524
7	maps	1	15.0	12.0	4.0	9.0	3.0	40.0	32.0	1.0	5.0	7.0	0.0	13.0	14.0	8.0	8.0	NA	NA	0.30769231
8	maps	2	18.0	10.0	7.0	9.0	2.0	32.0	41.0	17.0	11.0	10.0	1.0	19.0	15.0	16.0	0.0	NA	NA	0.00000000
9	maps	2	15.0	1.0	3.0	11.0	3.0	22.0	26.0	19.0	5.0	14.0	0.0	19.0	6.0	14.0	0.0	NA	NA	0.00000000
10	maps	2	8.0	10.0	2.0	5.0	2.0	35.0	31.0	15.0	8.0	7.0	0.0	28.0	19.0	11.0	2.0	NA	NA	0.33333333
11	maps	1	13.0	9.0	3.0	9.0	3.0	43.0	39.0	14.0	13.0	10.0	2.0	9.0	21.0	3.0	7.0	NA	NA	0.38095238
12	maps	4	14.0	1.0	3.0	12.0	6.0	33.0	25.0	24.0	15.0	23.0	2.0	27.0	11.0	29.0	0.0	NA	NA	0.14285714
13	maps	4	15.0	2.0	4.0	10.0	5.0	23.0	32.0	7.0	14.0	1.0	0.0	11.0	17.0	4.0	0.0	NA	NA	0.00000000
14	maps	4	19.0	3.0	7.0	11.0	0.0	23.0	23.0	21.0	13.0	20.0	1.0	23.0	18.0	21.0	2.0	NA	NA	0.00000000
15	maps	1	15.0	7.0	4.0	10.0	2.0	27.0	28.0	22.0	15.0	16.0	1.0	16.0	18.0	12.0	4.0	NA	NA	0.04761905
16	maps	1	11.0	13.0	6.0	5.0	7.0	45.0	28.0	2.0	0.0	5.0	0.0	23.0	26.0	21.0	9.0	NA	NA	0.19047619
17	maps	1	16.0	18.0	5.0	10.0	0.0	58.0	56.0	3.0	11.0	3.0	7.0	8.0	17.0	4.0	18.0	NA	NA	0.66666667
18	maps	1	17.0	11.0	6.0	11.0	4.0	39.0	44.0	19.0	18.0	23.0	1.0	21.0	20.0	21.0	6.0	NA	NA	0.33333333
19	maps	1	7.0	10.0	2.0	4.0	1.0	43.0	56.0	14.0	21.0	17.0	8.0	18.0	22.0	12.0	12.0	NA	NA	0.42857143
20	maps	1	13.0	12.0	4.0	8.0	4.0	38.0	35.0	9.0	6.0	1.0	0.0	11.0	10.0	8.0	1.0	NA	NA	0.00000000
21	maps	2	14.0	3.0	5.0	7.0	3.0	35.0	28.0	18.0	8.0	12.0	0.0	16.0	17.0	7.0	2.0	NA	NA	0.04761905
22	maps	2	11.0	10.0	3.0	7.0	1.0	37.0	47.0	11.0	12.0	5.0	1.0	16.0	22.0	12.0	4.0	NA	NA	0.52380952
23	maps	2	20.0	10.0	7.0	10.0	2.0	43.0	43.0	5.0	10.0	2.0	1.0	10.0	25.0	2.0	6.0	NA	NA	0.04761905

...

Der Affect Beispieldatensatz

"Tense Arousal" und " Energetic Arousal"

- *Activation-Deactivation Adjective Check List* nach Thayer (1986).
- Circa 25 Adjektive mit Vierpunkteskala
  - Wie gut beschreibt folgendes [Adjektiv] Ihre momentane Gefühlslage?
  - "I (1) do not feel, (2) cannot decide, (3) feel slightly, (4) definitely feel [Adjektiv]"
- Hohe TA Werte entsprechen hohen Selbsteinschätzungen bei Adjektiven wie
  - tense, clutched-up, fearful, jittery, intense
- Niedrige TA Werte entsprechen hohen Selbsteinschätzungen bei Adjektiven wie
  - still, at-rest, calm, quiet, placid
- Hohe EA Werte entsprechen hohen Selbsteinschätzungen bei Adjektiven wie
  - energetic, lively, active, vigorous, full-of-pep, wakeful, wide-awake
- Niedrige EA Werte entsprechen hohen Selbsteinschätzungen bei Adjektiven wie
  - sleepy, drowsy, tired

Häufigkeitsverteilungen

Definition (Absolute und relative Häufigkeitsverteilungen)

$x = (x_1, ..., x_n)$ sei ein *Datensatz* ("Urliste") und $A := \{a_1, ..., a_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden verschiedenen Zahlenwerte ("Merkmalsausprägungen"). Dann heißt die Funktion

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

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

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

die *relative Häufigkeitsverteilung* der Zahlwerte von x.

Häufigkeitsverteilungen

table() zum Erzeugen einer absoluten Häufigkeitsverteilung

Division durch n zum Berechnen der relativen Häufigkeitsverteilung

```
x        = affect$TA1                    # double vector der TA1 Werte
n        = length(x)                     # Anzahl der Datenwerte (330)
H        = as.data.frame(table(x))       # absolute Haeufigkeitsverteilung (dataframe)
names(H)= c("a", "h")                    # Spaltenbenennung
H$r      = H$h/n                         # relative Haeufigkeitsverteilung
print(H, digits = 1)                     #  Ausgabe
      a  h     r
1     0  1 0.003
2   2.5  1 0.003
3     3  2 0.006
4     4  2 0.006
5     5  8 0.024
6     6  7 0.021
7     7  9 0.027
8     8 19 0.058
9     9 13 0.039
10   10 40 0.121
11   11 28 0.085
12   12 27 0.082
```

Häufigkeitsverteilungen

barplot() zur Visualisierung der absoluten Häufigkeitsverteilung

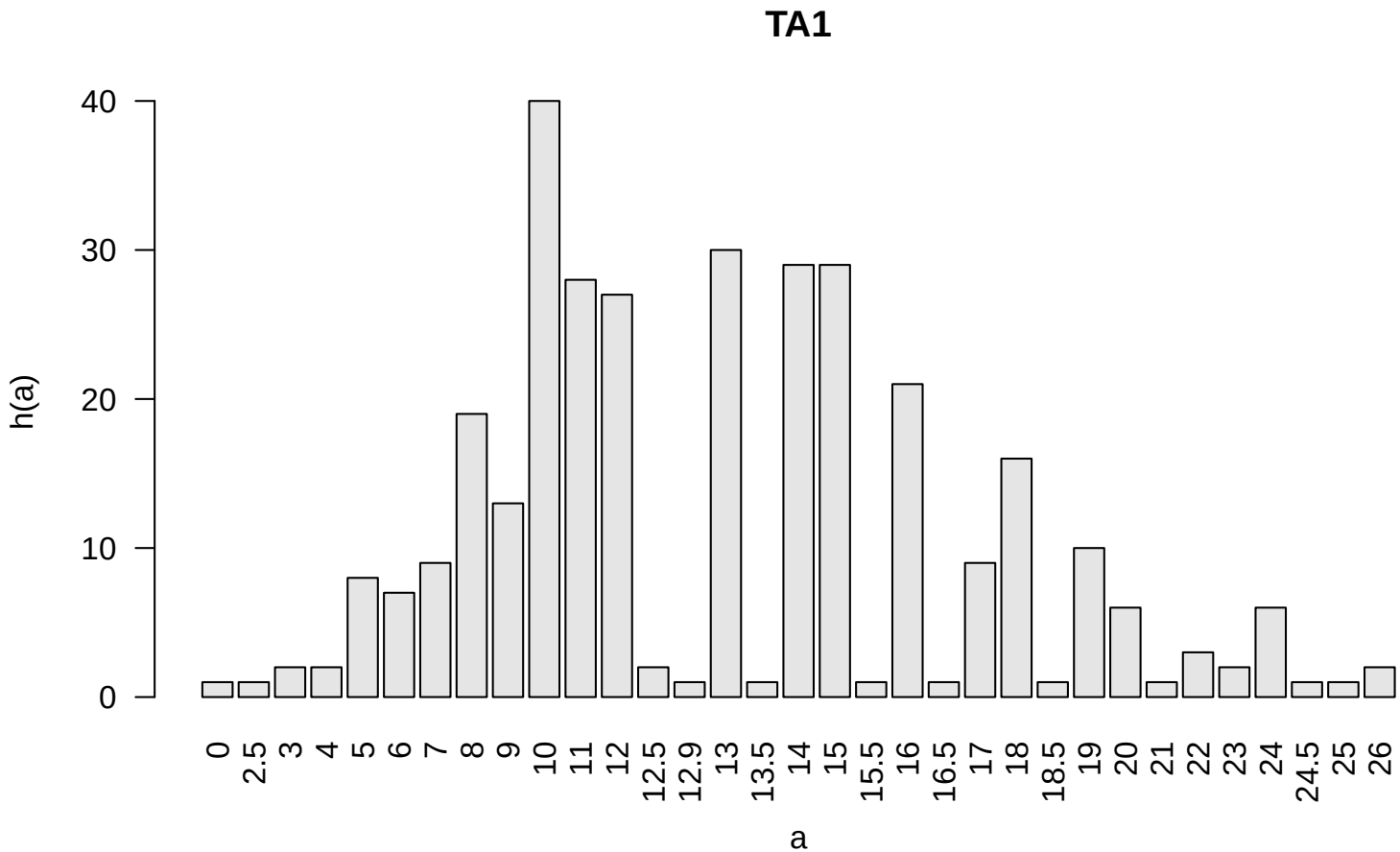
```
h          = H$h                    # h(a) Werte
names(h)   = H$a                    # barplot braucht a Werte als names
dev.new()                           # Abbildungsinitialisierung
barplot(                            # Balkendiagramm
h,                                  # absolute Haeufigkeiten
col        = "gray90",              # Balkenfarbe
xlab       = "a",                   # x Achsenbeschriftung
ylab       = "h(a)",                # y Achsenbeschriftung
las        = 2,                     # Anzeigen aller x Werte
main       = "TA1"                  # Titel
)
```

dev.copy2pdf() zum Speichern der Abbildung als .pdf Datei

```
dev.copy2pdf(                       # PDF Kopiefunktion
file       = "...  .pdf",           # Dateiname
width      = 8,                     # Breite (inch)
height     = 5                      # Hoehe (inch)
)
```

Häufigkeitsverteilungen

barplot() zur Visualisierung der relativen Häufigkeitsverteilung



Häufigkeitsverteilungen

barplot() zur Visualisierung der relativen Häufigkeitsverteilung

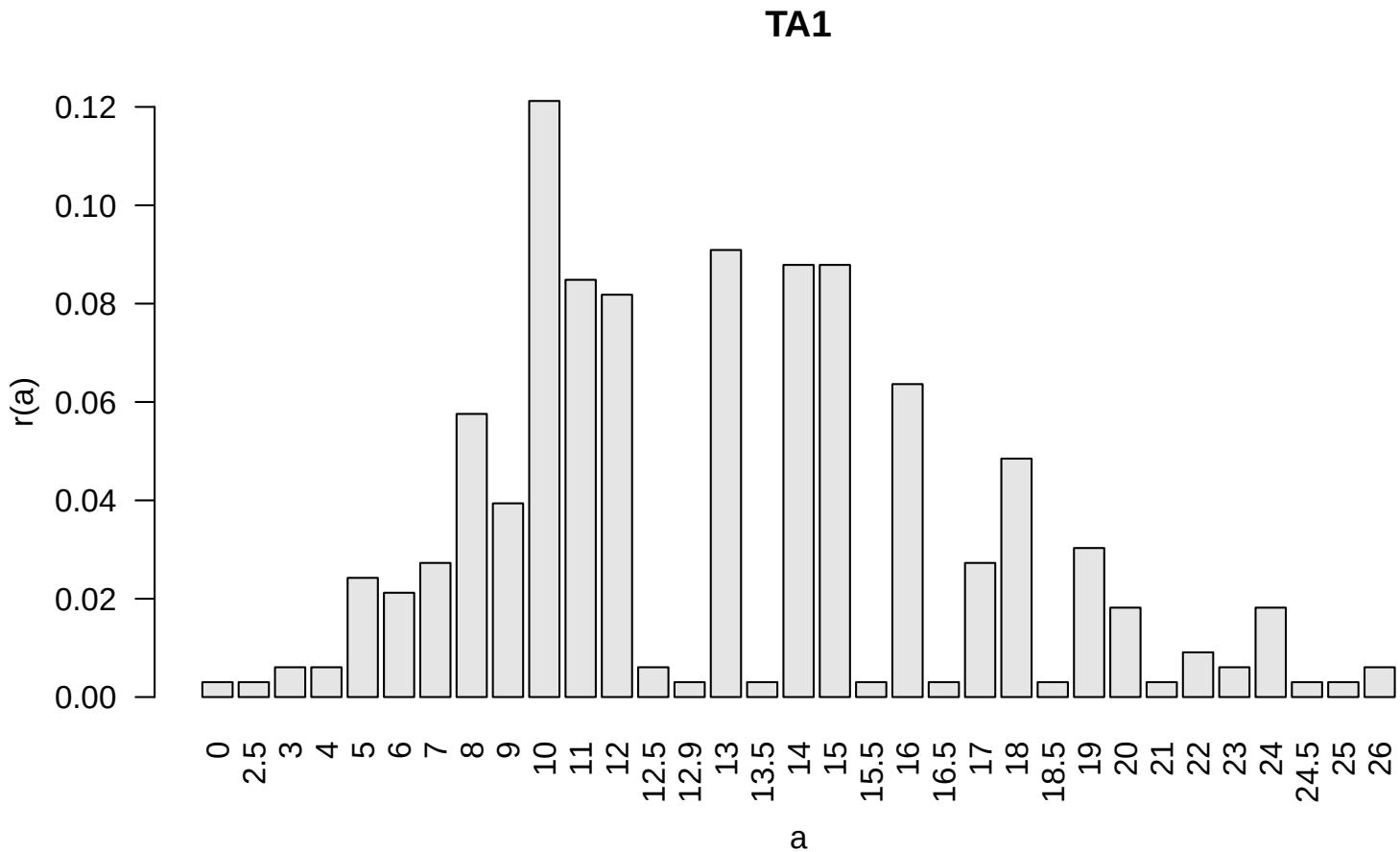
```
r          = H$r                    # r(a) Werte
names(h)   = H$r                    # barplot braucht a Werte als names
dev.new()                           # Abbildungsinitialisierung
barplot(                            # Balkendiagramm
h,                                  # relative Haeufigkeiten
col        = "gray90",              # Balkenfarbe
xlab       = "a",                   # x Achsenbeschriftung
ylab       = "r(a)",                # y Achsenbeschriftung
las        = 2,                     # Anzeigen aller x Werte
main       = "TA1"                  # Titel
)
```

dev.copy2pdf() zum Speichern der Abbildung als .pdf Datei

```
dev.copy2pdf(                       # PDF Kopiefunktion
file       = "...  .pdf",           # Dateiname
width      = 8,                     # Breite (inch)
height     = 5                      # Hoehe (inch)
)
```

Häufigkeitsverteilung

barplot() zur Visualisierung der relativen Häufigkeitsverteilung



Histogramme

Definition (Histogramm)

Ein *Histogramm* ist ein Diagramm, in dem zu einem Datensatz $x = (x_1, ..., x_n)$ mit verschiedenen Zahlwerten $A := \{a_1, ..., a_m\}, m \leq n$ über benachbarten Intervallen $[b_{j-1}, b_j[$, welche *Klassen* oder *Bins* genannt werden, für $j = 1, ..., k$ 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)
$h := 3.49S/\sqrt[3]{n}$	Min. MSE Dichteschätzung bei Normalverteilung (Scott, 1979)

Histogramme

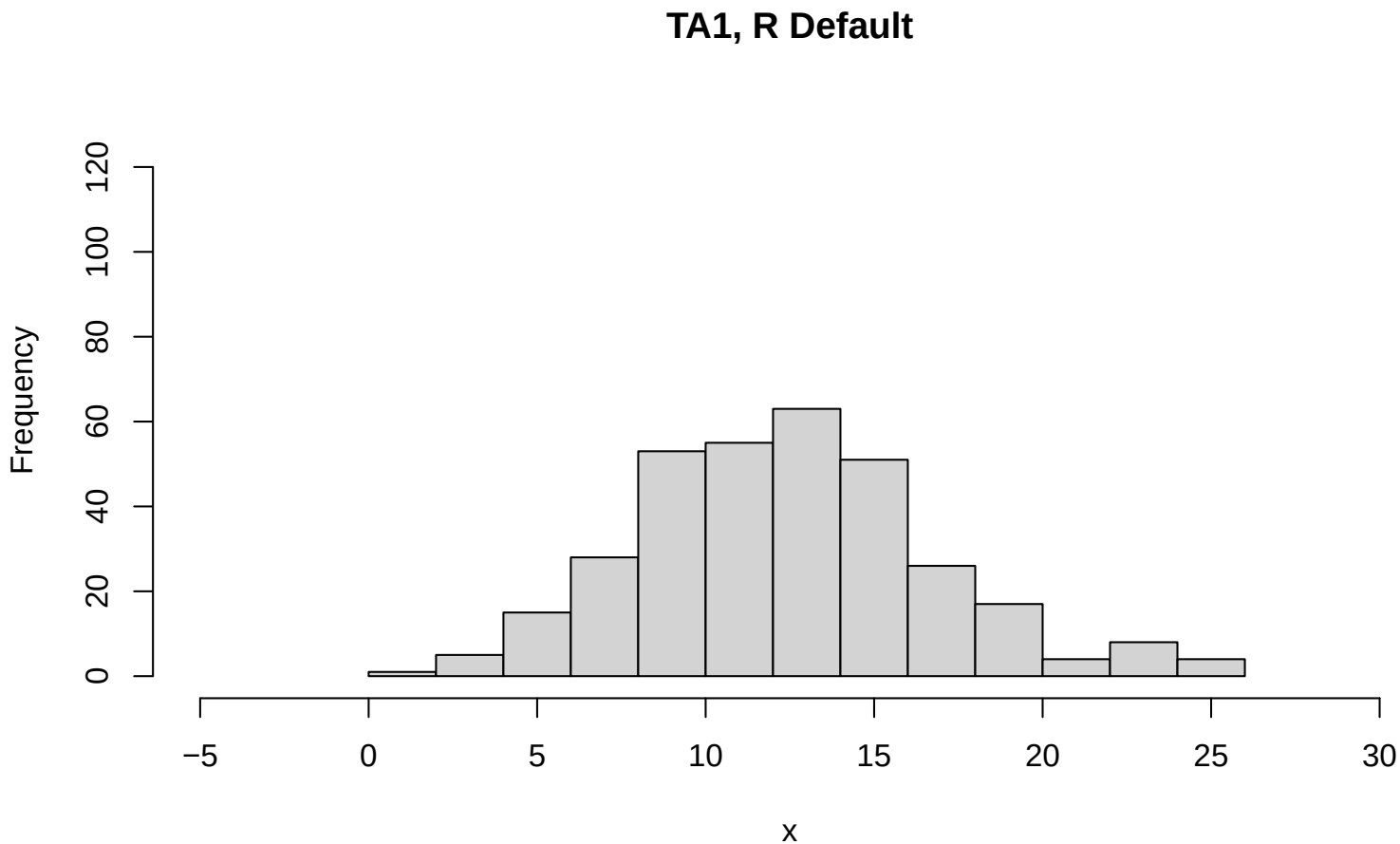
Die Funktion **hist()** berechnet und visualisiert Histogramme

- Die Klassen $[b_{j-1}, b_j[, j = 1, ..., k$ werden als Argument *breaks* festgelegt
- *breaks* ist der atomic 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

```
# Default Histogramm
x       = affect$TA1                    # Datensatz
x_min   = -5                            # x Achsengrenze (unten)
x_max   = 30                            # x Achsengrenze (oben)
y_min   = 0                             # y Achsengrenze (oben)
y_max   = 130                           # y Achsengrenze (unten)
hist(                                   # Histogramm
x,                                      # Datensatz
xlim    = c(x_min, x_max),              # x Achsengrenzen
ylim    = c(y_min, y_max),              # y Achsengrenzen
main    =  "TA1, R Default"             # Titel
)
```

Histogramme

Default Histogramm

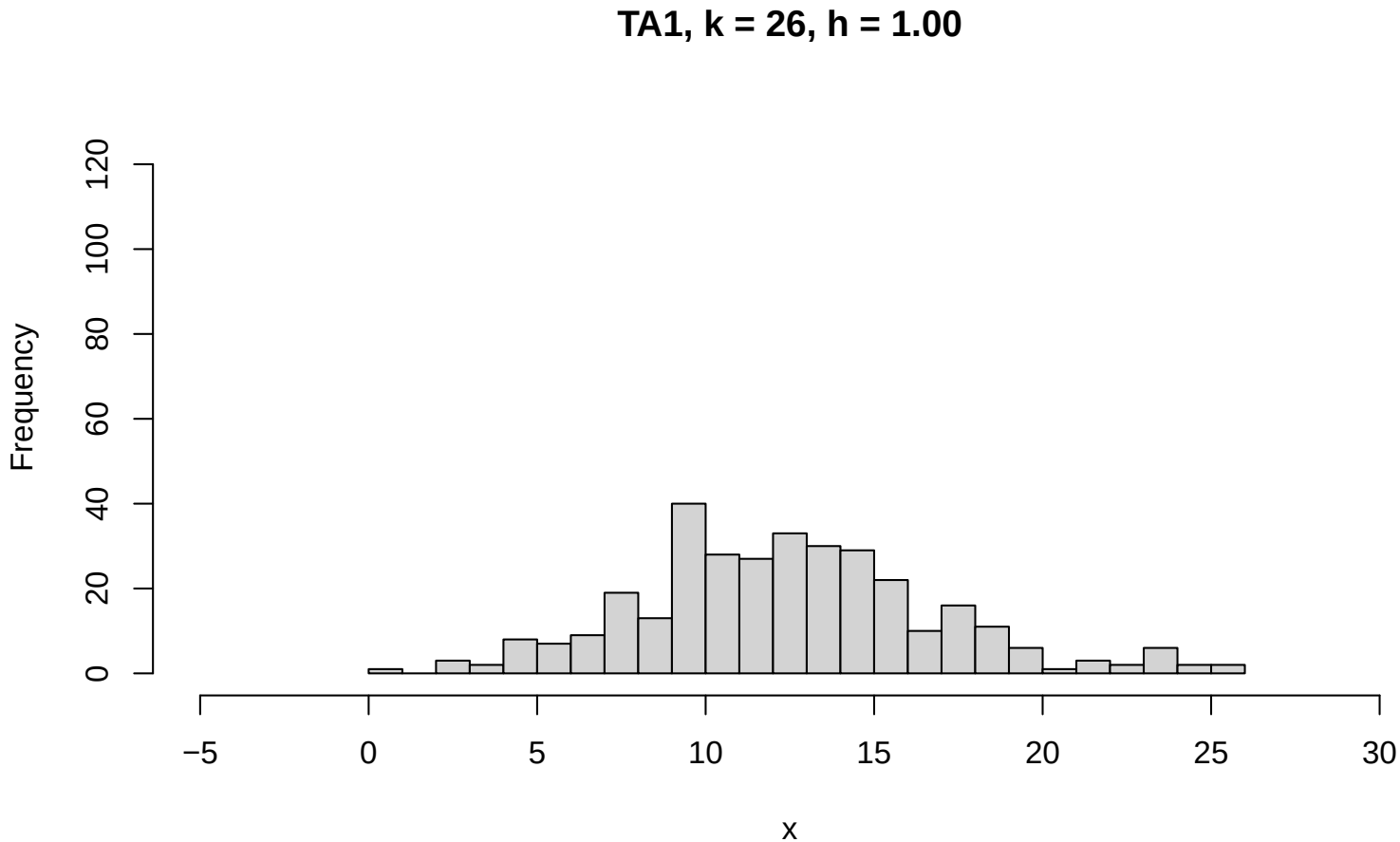


Histogramme

```
# Histogramm mit gewuenschter Klassenbreite
h   = 1                                  # gewuenschte 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[
```

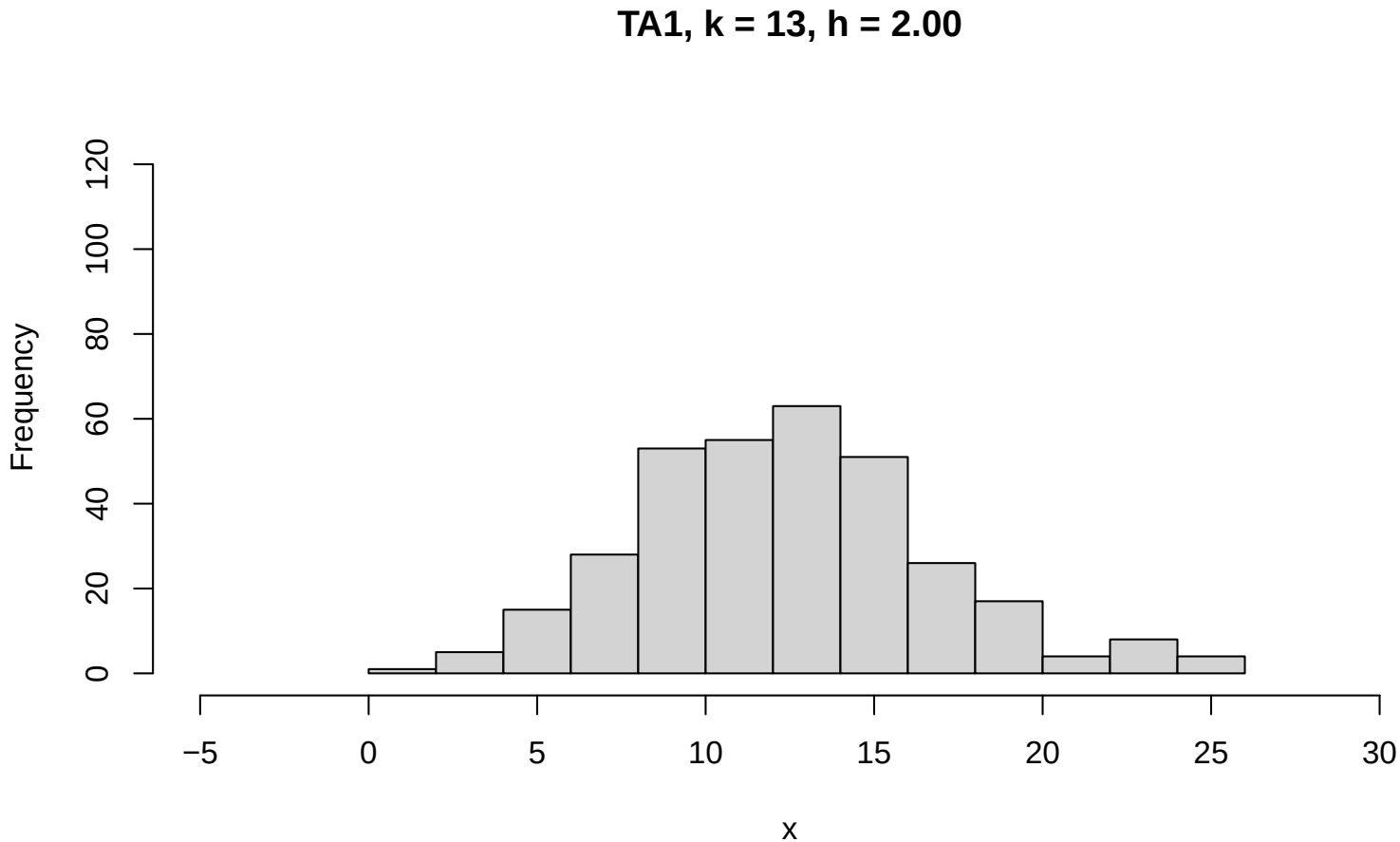
Histogramme

Gewünschte Klassenbreite $h := 1$



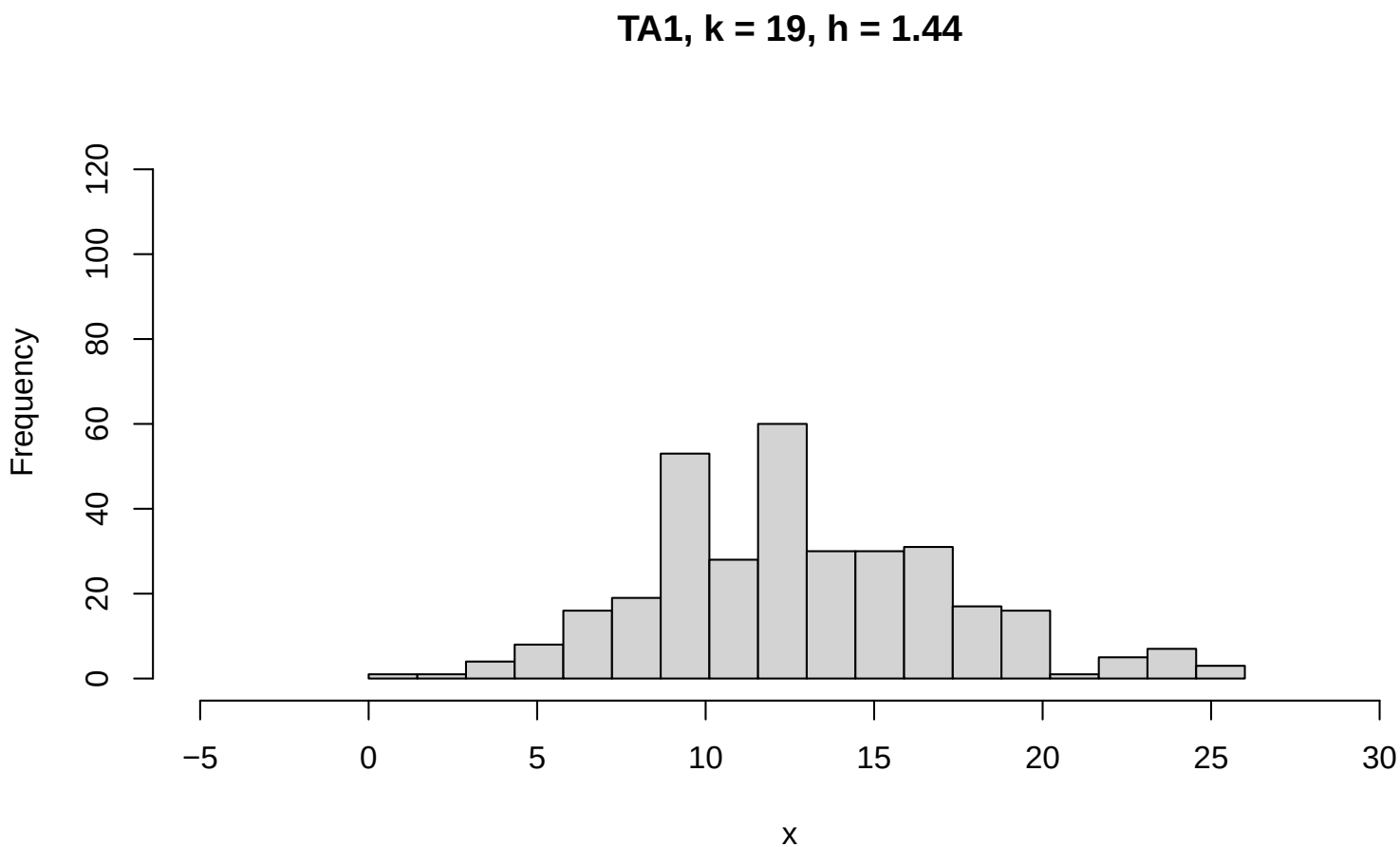
Histogramme

Gewünschte Klassenbreite $h := 2$



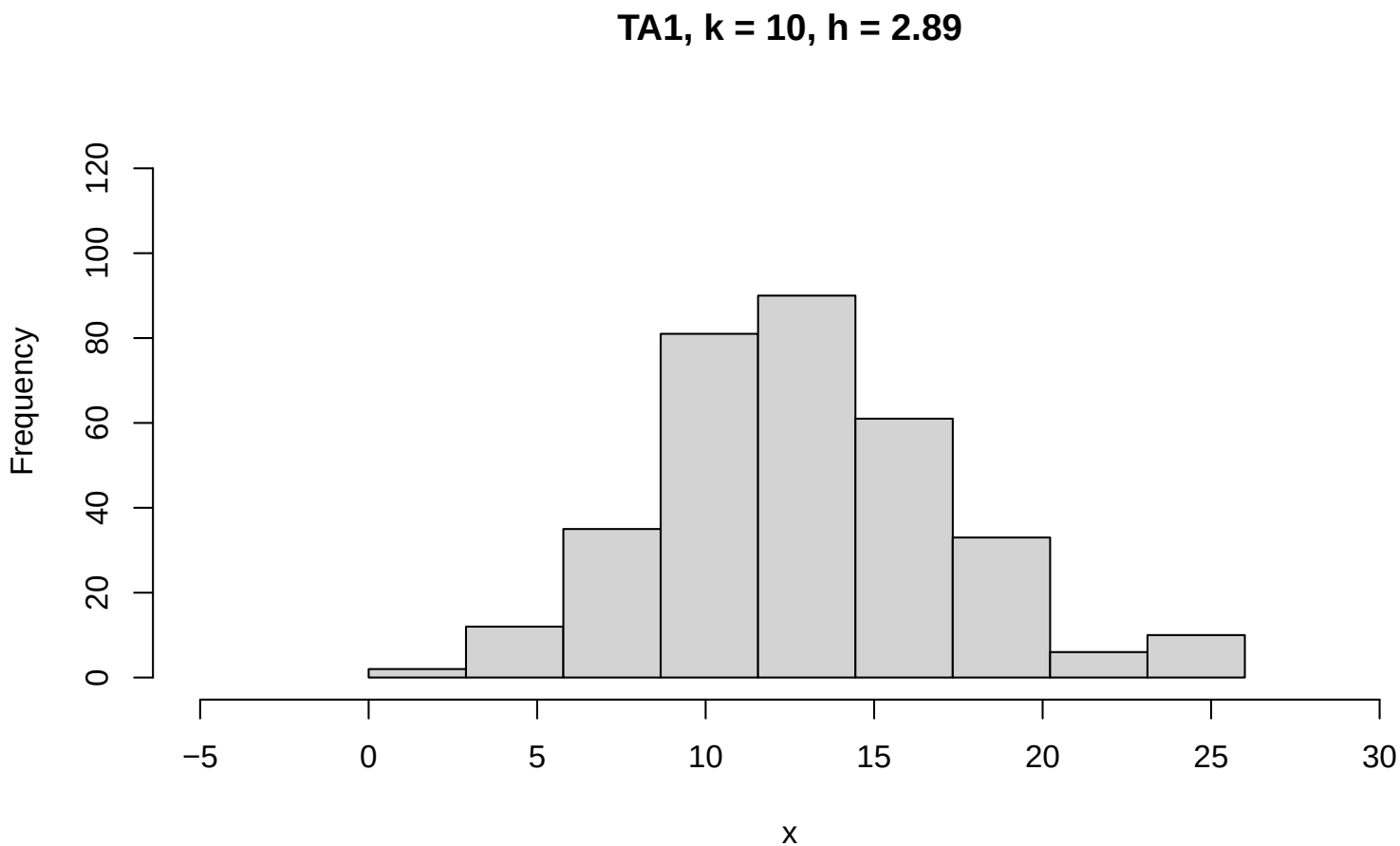
Histogramme

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



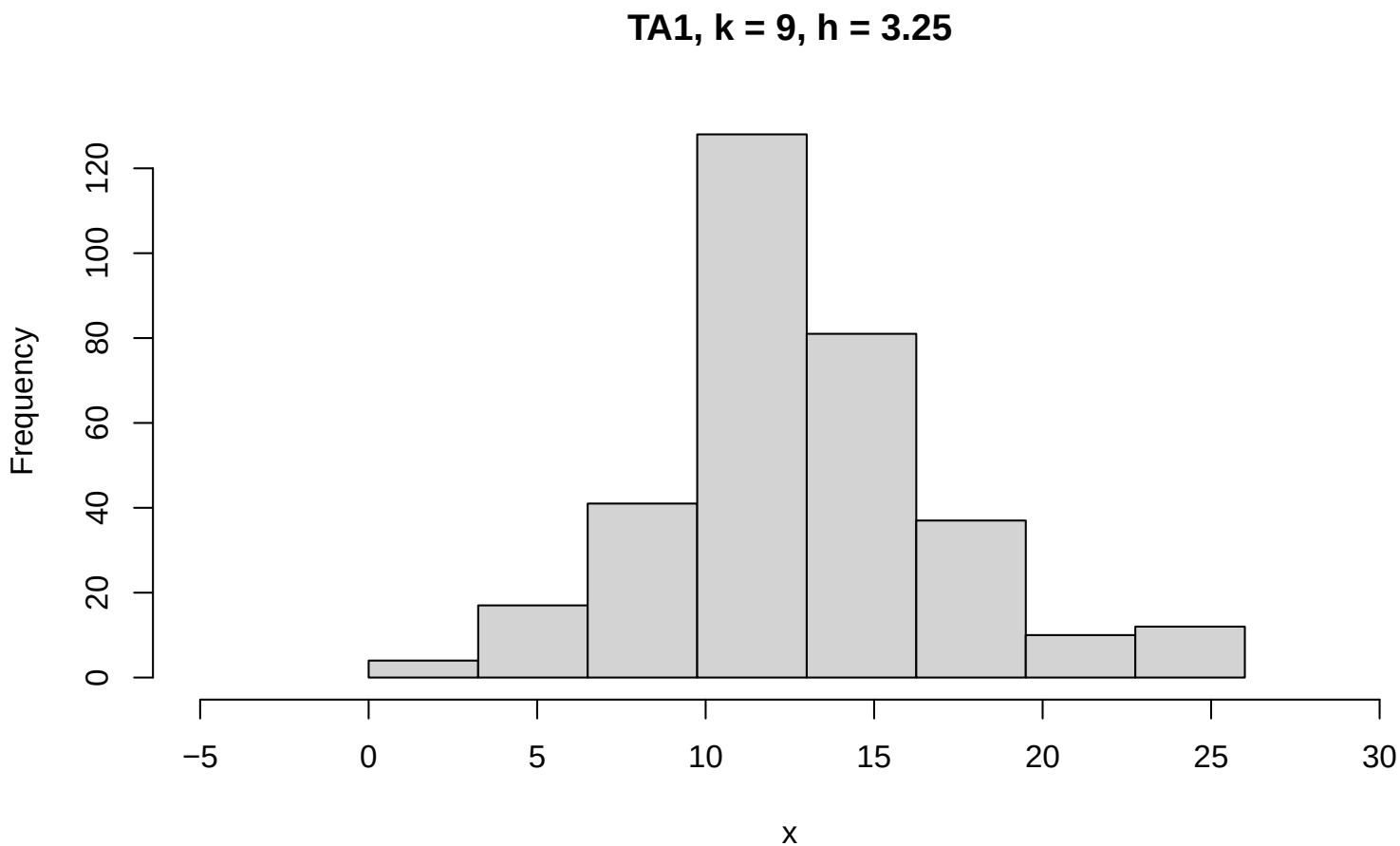
Histogramme

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



Histogramme

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



Empirische Verteilungsfunktion

Definition (Kumulative absolute und relative Häufigkeitsverteilungen)

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

$$H : A \to \mathbb{N}, a \mapsto H(a) := \sum_{a' \leq a} h(a') \tag{3}$$

die *kumulative absolute Häufigkeitsverteilung* von x und die Funktion

$$R : A \to [0, 1], a \mapsto R(a) := \sum_{a' \leq a} r(a') \tag{4}$$

die *kumulative relative Häufigkeitsverteilung* der Zahlwerte von x.

Mit den Definitionen der absoluten und relativen Häufigkeitsverteilungen gilt also

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

und

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

Empirische Verteilungsfunktion

cumsum() erlaubt die Berechnung kumulativer Summen von atomic vector Elementen

```
x         = affect$TA1                  # double vector der TA1 Werte
n         = length(x)                   # Anzahl der Datenwerte
H         = as.data.frame(table(x))     # absolute Hauefigkeitsverteilung dataframe
names(H)= c("a", "h")                   # Spaltenbenennung
H$H       = cumsum(H$h)                 # kumulative absolute Haeufigkeitsverteilung
H$r       = H$h/n                       # relative Haeufigkeitsverteilung
H$R       = cumsum(H$r)                 # kumulative relative Haeufigkeitsverteilung
```

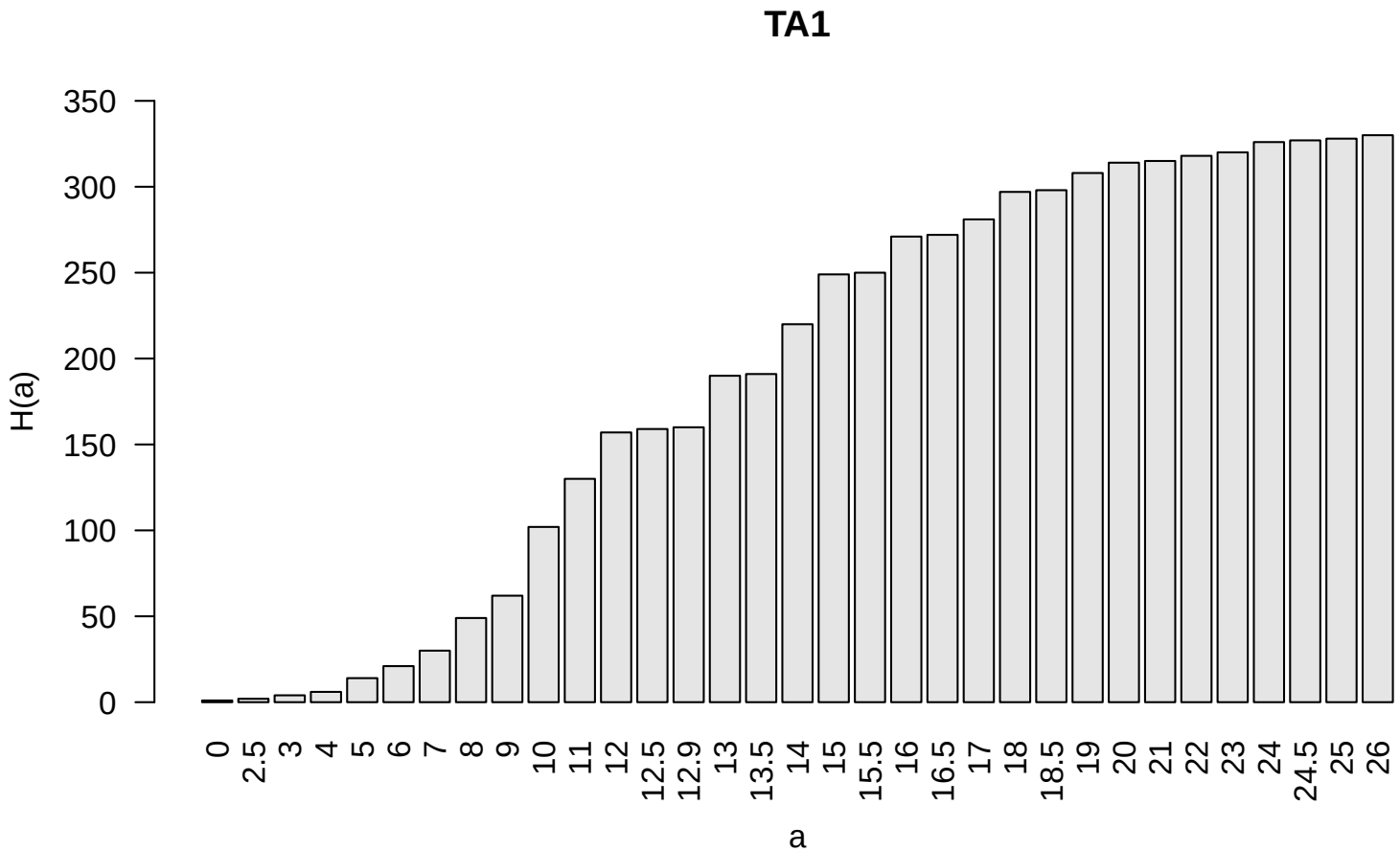
	a	h	H	r	R
1	0	1	1	0.003	0.003
2	2.5	1	2	0.003	0.006
3	3	2	4	0.006	0.012
4	4	2	6	0.006	0.018
5	5	8	14	0.024	0.042
6	6	7	21	0.021	0.064
7	7	9	30	0.027	0.091
8	8	19	49	0.058	0.148
9	9	13	62	0.039	0.188
10	10	40	102	0.121	0.309
11	11	28	130	0.085	0.394
12	12	27	157	0.082	0.476
13	12.5	2	159	0.006	0.482
14	12.9	1	160	0.003	0.485

Empirische Verteilungsfunktion

```
# Visualisierung der kumulativen absoluten Haeufigkeitsverteilung
Ha          = H$H                   # H(a) Werte
names(Ha)   = H$a                   # barplot braucht a Werte als names
dev.new()                           # Abbildungsinitialisierung
barplot(                            # Balkendiagramm
Ha,                                 # H(a) Werte
col         = "gray90",             # Balkenfarbe
xlab        = "a",                  # x Achsenbeschriftung
ylab        = "H(a)",               # y Achsenbeschriftung
las         = 2,                    # Anzeigen aller x Werte
ylim        = c(0,350),             # y Achsenlimits
main        = "TA1")                # Titel

# Visualisierung der kumulativen relativen Haeufigkeitsverteilung
R           = H$R                   # R(a) Werte
names(R)    = H$a                   # barplot braucht a Werte als names
dev.new()                           # Abbildungsinitialisierung
barplot(                            #  Balkendiagramm
R,                                  # R(a) Werte
col         = "gray90",             # Balkenfarbe
xlab        = "a",                  # x Achsenbeschriftung
ylab        = "R(a)",               # y Achsenbeschriftung
las         = 2,                    # Anzeigen aller x Werte
ylim        = c(0,1),               # y Achsenlimits
main        = "TA1")                # Titel
```

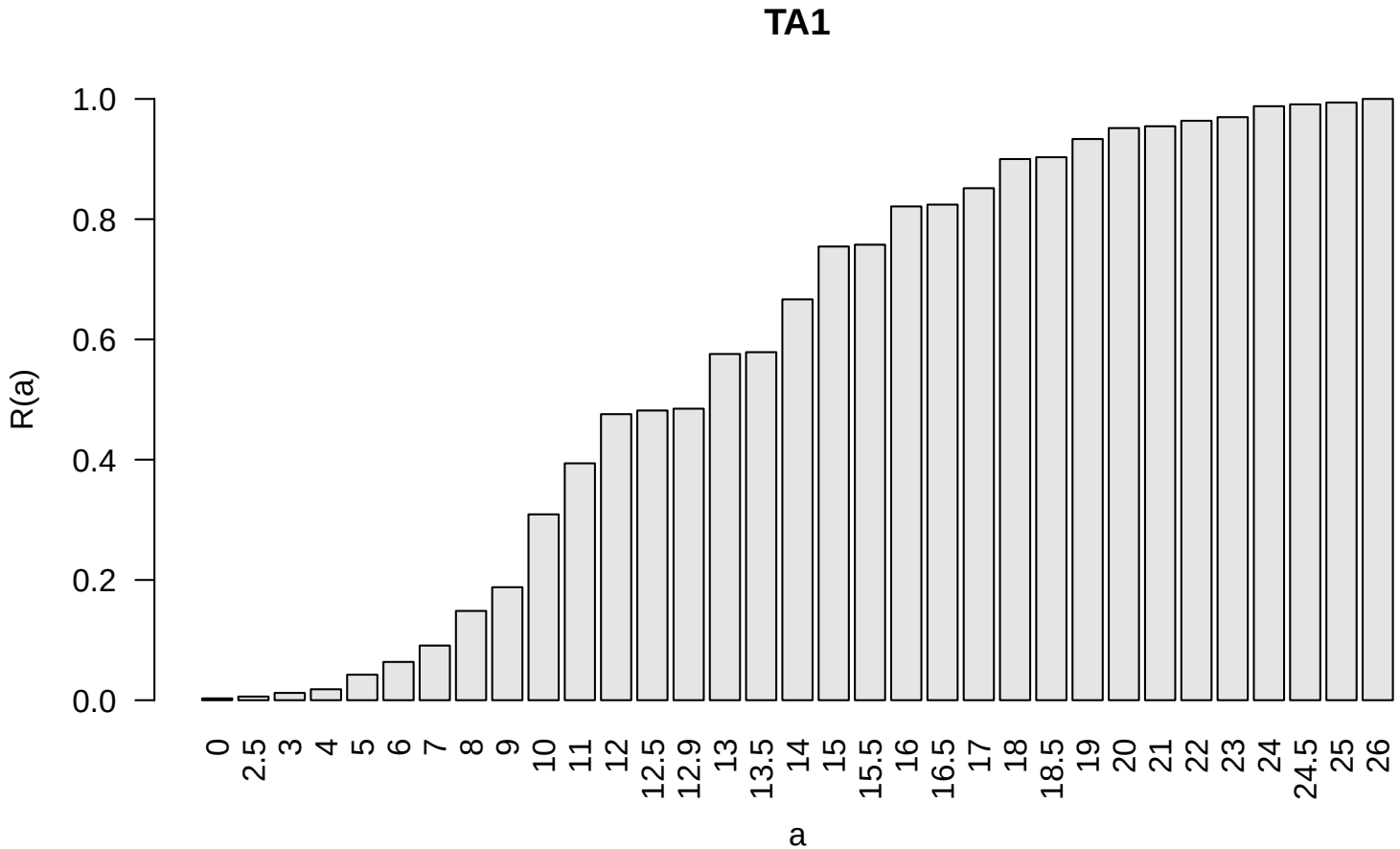
Empirische Verteilungsfunktion

barplot() zur Visualisierung der kumulativen absoluten Häufigkeitsverteilung



Empirische Verteilungsfunktion

barplot() zur Visualisierung der kumulativen relativen Häufigkeitsverteilung



Empirische Verteilungsfunktion

Definition (Empirische Verteilungsfunktion)

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

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

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.

Empirische Verteilungsfunktion

ecdf() erlaubt die Evaluation der empirischen Verteilungsfunktion

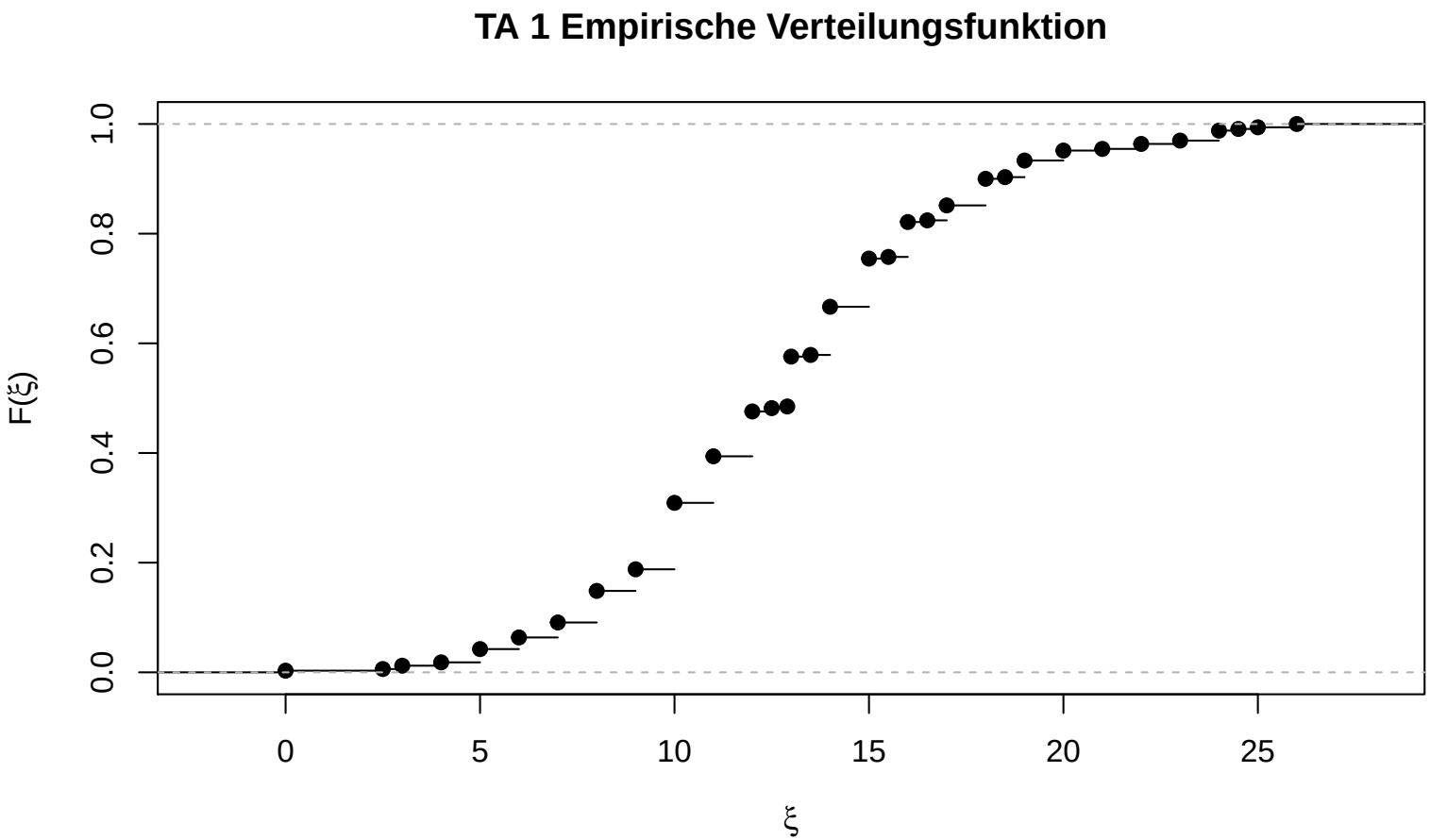
```
x    = affect$TA1                                  # double vector der TA1 Werte
evf  = ecdf(x)                                     # Evaluation der EVF
plot(                                              # plot weiss mit ecdf object umzugehen
evf,                                               # ecdf Objekt
xlab = TeX("$\\xi$"),                              # x Achsenbeschriftung
ylab = TeX("$F(\\xi)$") ,                          # y Achsenbeschriftung
main = "TA1 Empirische Verteilungsfunktion")       # Titel
```

Kombination mit **abline()** erlaubt das Ablesen von *Quantilen*

```
plot(                                              # plot weiss mit ecdf Objekt umzugehen
evf,                                               # ecdf Objekt
verticals    = TRUE,                               # vertikale Linien
do.points    = FALSE,                              # keine Punkte
xlab         = TeX("$\\xi$"),                      # x Achsenbeschriftung
ylab         = TeX("$F(\\xi)$") ,                  # y Achsenbeschriftung
main  = "TA 1 Empirische Verteilungsfunktion")     # Titel
abline(                                            # horizontale Linie
h            = 0.25,                               # y Ordinate der Linie
lty          = 3,                                  # fein gestrichelt
col          = "blue")                             # blau
```

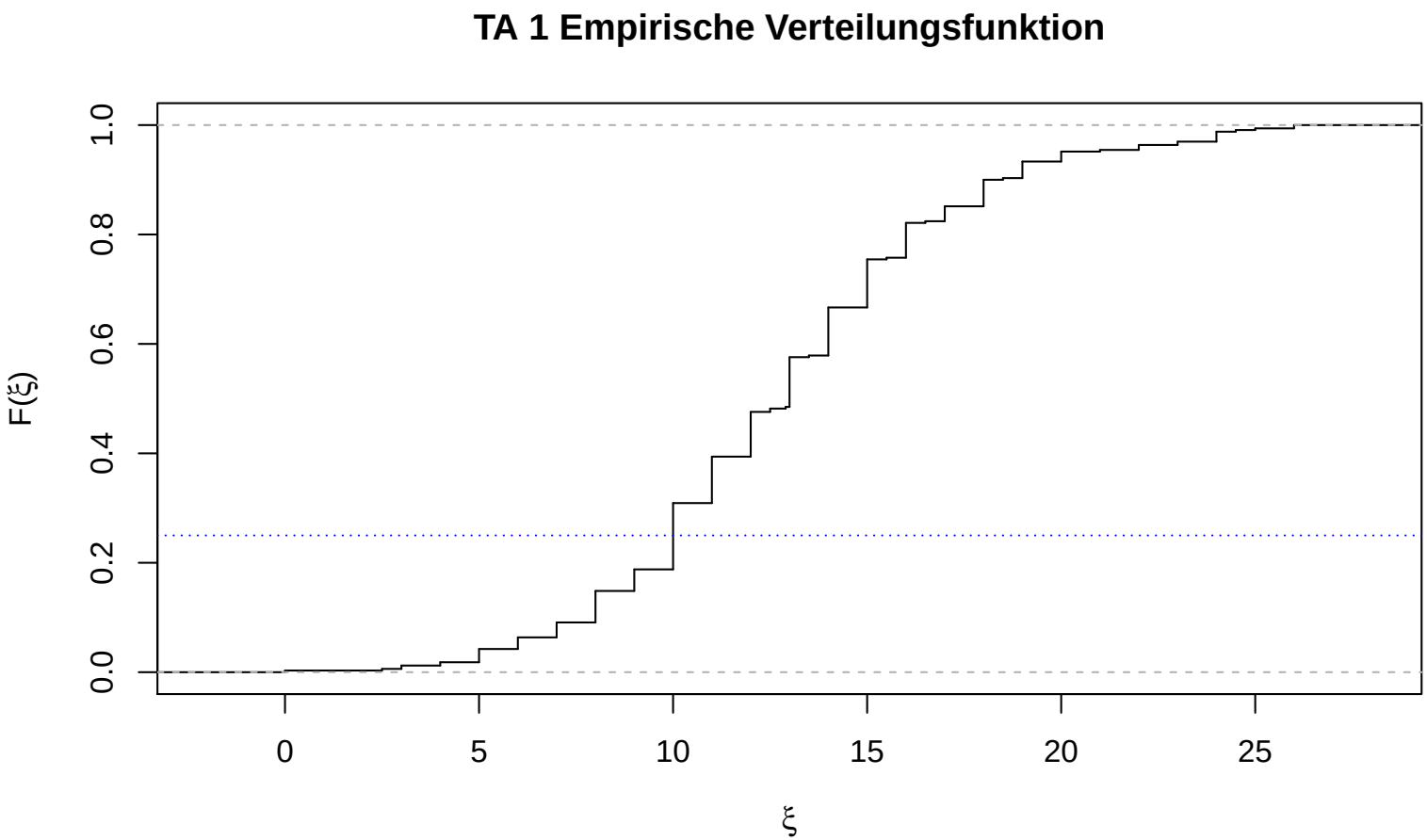
Empirische Verteilungsfunktion

Die Punkt-Liniendarstellung betont die rechtsseitige Stetigkeit.



Empirische Verteilungsfunktion

Kombination mit **abline()** zum Ablesen von *Quantilen*



Quantile und Boxplots

Definition (p-Quantil)

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

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

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

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

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*.

Quantile und Boxplots

Beispiel (p-Quantil, Henze (2018, Kapitel 5))

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

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

Quantile und Boxplots

Beispiel (p-Quantil, Henze (2018, Kapitel 5))

- "Manuelle" Quantilbestimmung anhand obiger Definition

```
x   = c(8.5,1.5,75,4.5,6.0,3.0,3.0,2.5,6.0,9.0)   # Beispieldaten
n   = length(x)                                    # Anzahl Datenwerte
x_s = sort(x)                                      # sortierter Datensatz
p   = 0.25                                         # np \notin \mathbb{N}
x_p = x_s[floor(n*p + 1)]                          # 0.25 Quantil
[1] 3.0
p   = 0.80                                         # np \in \mathbb{N}
x_p = (1/2)*(x_s[n*p] + x_s[n*p + 1])              # 0.80 Quantil
[1] 8.75
```

- **quantile()** wertet Quantile anhand der Quantildefinition *type* = aus
- Es gibt mindestens neun verschiedene Quantildefinitionen (Hyndman and Fan, 1996)

```
x_p = quantile(x, 0.25, type = 2)                  # 0.25 Quantil, Definition 2
[1] 3.0
x_p = quantile(x, 0.80, type = 2)                  # 0.80 Quantil, Definition 2
[1] 8.75
x_p = quantile(x, 0.80, type = 1)                  # 0.80 Quantil, Definition 1
[1] 8.5
```

Quantile und Boxplots

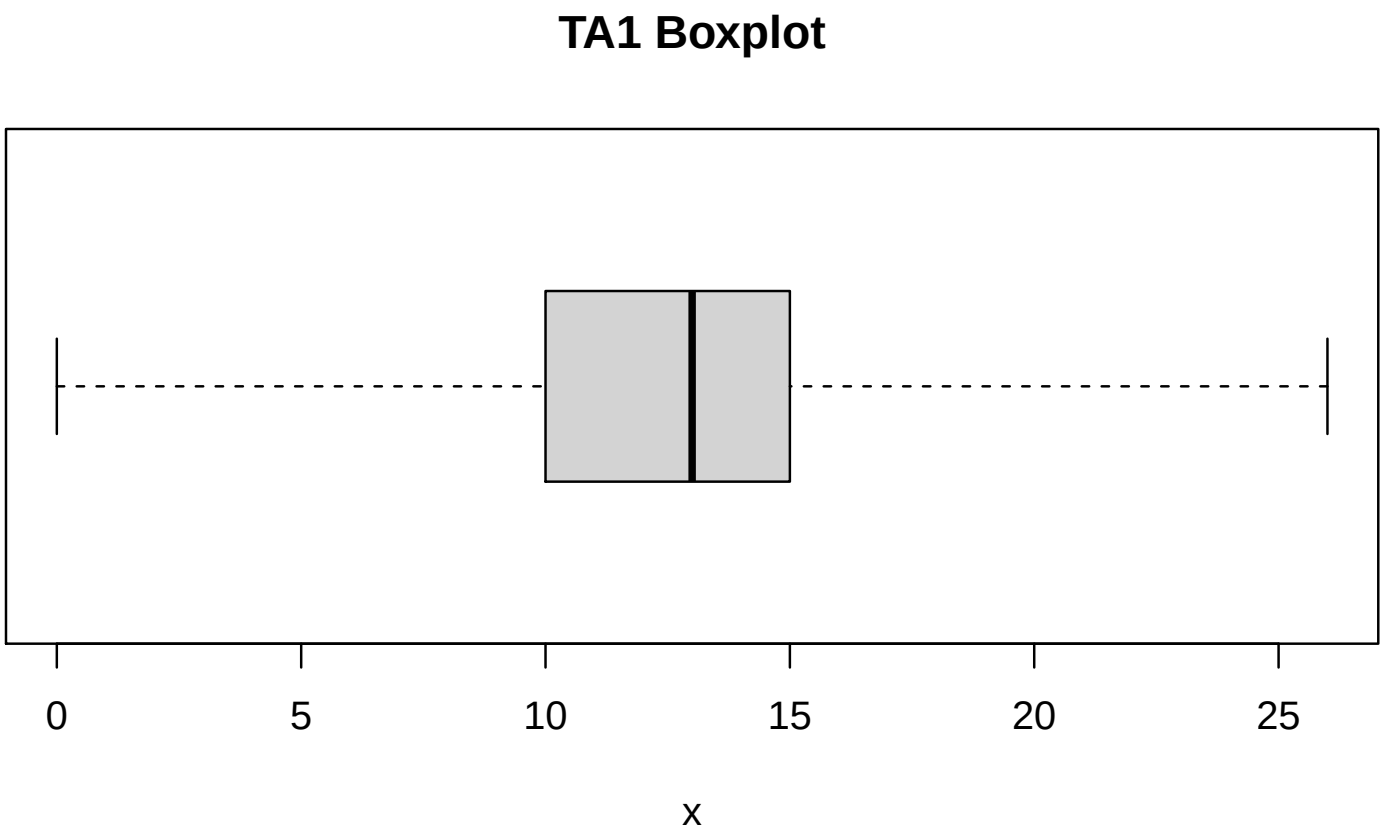
Boxplot

- Ein Boxplot visualisiert eine Quantil-basierte Zusammenfassung eines Datensatzes
- Typischerweise werden $\min x, x_{0.25}, x_{0.50}, x_{0.75}, \max x$ 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
- **boxplot()** erstellt einen Boxplot, **summary()** liefert wesentliche Kennzahlen

```
# Sechswertezusammenfassung
summary(affect$TA1)
Min.   1st Qu.    Median    Mean   3rd Qu.  Max.
0.00   10.00      13.00     12.91   15.00    26.00

# Boxplot
boxplot(                              # Boxplot
affect$TA1,                           # Datensatz
horizontal  = T,                      # horizontale Darstellung
range       = 0,                      # Whiskers bis zu min x und max x
xlab        = "x",                    # x Achsenbeschriftung
main        = "TA1 Boxplot")          # Titel
```

Empirische Verteilungsfunktion

Boxplot



Es gibt viele Boxplotvariationen(z.B. McGill et al., 1978), Erläuterung ist immer nötig!

Univariate Deskriptivstatistiken

- Verteilungsdarstellung
- **Maße der zentralen Tendenz**
- Maße der Datenvariabilität
- Übungen und Selbstkontrollfragen

Mittelwert

Definition (Mittelwert)

$x = (x_1, ..., x_n)$ sei ein Datensatz. Dann heißt $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$ der Mittelwert von x.

mean() zur Berechnung des Mittelwerts

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
n       = length(x)                     # Anzahl der Werte
x_bar   = (1/n)*sum(x)                  # "manuelle" Mittelwertsberechnung
x_bar   = mean(x)                       # "automatische" Mittelwertsberechnung
```

Die Summe der Abweichungen vom Mittelwert ist Null

```
s    = sum(x - mean(x))                 # Summe der Abweichungen vom Mittelwert
[1] -8.704149e-14                       # Rundungsfehler
```

Die absoluten Summen positiver und negativer Abweichungen vom Mittelwert sind gleich.

```
s_1 = sum(x[x <= mean(x)] - mean(x))    # Summe aller negativer Abweichungen
s_1
[1] -565.4909
s_2 = sum(x[x > mean(x)] - mean(x))     # Summe aller positiver Abweichungen
s_2
[1] 565.4909
```

Mittelwert

Der Mittelwert der Summe zweier Datenreihen entspricht der Summe ihrer Mittelwerte:

$$\overline{x+y} := \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i =: \bar{x} + \bar{y}$$

```
x     = affect$TA1                    # double Vektor der TA1 Werte
x_bar = mean(x)                       # Mittelwert der TA1 Werte
y     = affect$EA1                    # double Vektor der EA1 Werte
y_bar = mean(y)                       # Mittelwert der EA1 Wert
z     = x + y                         # double Vektor der TA und EA Werte
z_bar = mean(z)                       # Mittelwert der Summe der TA und EA Werte
[1] 22.10576
xy_bar  = x_bar + y_bar               # Summe der Mittelwerte der TA und EA Werte
[1] 22.10576
```

Mittelwert

Linear-affine Transformation der Daten transformiert den Mittelwert linear-affin:

$$\overline{ax + b} = a\bar{x} + b$$

Beweis

$$\begin{aligned}\overline{ax + b} &:= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\ &= \sum_{i=1}^{n}\left(\frac{1}{n}ax_i + \frac{1}{n}b\right) = \sum_{i=1}^{n}\left(\frac{1}{n}ax_i\right) + \sum_{i=1}^{n}\left(\frac{1}{n}b\right) = a\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}b = a\bar{x} + b\end{aligned}$$

□

```
x     = affect$TA1                    # double Vektor der TA1 Werte
x_bar = mean(x)                       # Mittelwert der TA1 Werte
a     = 2                             # Multiplikationskonstante
b     = 5                             # Additionskonstante
y     = a*x + b                       # linear-affine Transformation der TA1 Werte
y_bar = mean(y)                       # Mittelwert der transfomierten TA1 Werte
[1] 30.82364
y_bar   = a*x_bar + b                 # Transformation des TA1 Mittelwerts
[1] 30.82364
```

Mittelwert

Beim *getrimmten Mittelwert* wird ein relativer Anteil an Extremwerten ausgeschlossen.

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
n       = length(x)                     # Anzahl der Werte
t       = 0.3                           # getrimmter relativer Anteil
n_t     = round(0.3*n)                  # Anzahl zu trimmender Werte
x_sort    = sort(x)                     # aufsteigend sortierter Vektor
x_trim  = x_sort[(n_t+1):(n - n_t)]     # getrimmter Datensatz
x_tbar  = mean(x_trim)                  # getrimmter Mittelwert
[1] 12.68485
x_tbar  = mean(x, trim = 0.3)           # "automatische" Berechnung
[1] 12.68485
```

Median

Definition (Median)

$x = (x_1, ..., x_n)$ sei ein Datensatz und $x_s = (x_{(1)}, ..., x_{(n)})$ der zugehörige aufsteigend sortierte Datensatz. Dann ist der Median von x definiert als

$$\tilde{x} := \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{x_{(n/2)} + x_{((n+1)/2)}}{2} & \text{falls } n \text{ gerade} \end{cases} \tag{10}$$

median() zur Berechnung des Medians

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
n       = length(x)                     # Anzahl der Werte
x_s     = sort(x)                       # aufsteigend sortierter Vektor
if(n %% 2 == 1)                         # n ungerade, n mod 2 == 1
{
    x_tilde = x_s[(n+1)/2]
} else                                  # n gerade, n mod 2 == 0
{
    x_tilde = (x_s[n/2] + x_s[(n+1)/2])/2
}
x_tilde   = median(x)                   # "automatische" Berechnung
```

Median

Die Summe der Abweichungsbeträge vom Median ist minimal

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
x_bar   = mean(x)                       # Mittelwert der TA1 Werte
x_tilde = median(x)                     # Median der TA1 Werte
s_1     = sum(abs(x - x_bar))           # Summe Abweichungsbetraege vom Mittelwert
[1] 1130.982
s_2     = sum(abs(x - x_tilde))         # Summe Abweichungsbetraege vom Median
[1] 1130.1
```

Der Median ist weniger anfällig für Ausreißer als der Mittelwert

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
x_bar   = mean(x)                       # Mittelwert der TA1 Werte
[1] 12.91182
x_tilde = median(x)                     # Median der TA1 Werte
[1] 13
y       = x                             # neuer Datensatz mit
y[1]    = 10000                         # ... einem Extremwert
y_bar   = mean(y)                       # Mittelwert der des neuen Datensatzes
[1] 43.17242
y_tilde = median(y)                     # Median bleibt unveraendert
[1] 13
```

Deskriptive Statistiken

- Verteilungsdarstellung
- Maße der zentralen Tendenz
- **Maße der Datenvariabilität**
- Bivariate Deskriptivstatistik
- Übungen und Selbstkontrollfragen

Spannbreite

Definition

$x = (x_1, ..., x_n)$ sei ein Datensatz. Dann ist die *Spannbreite* von $x_1, ..., x_n$ definiert als

$$S := \max(x_1, ..., x_n) - \min(x_1, ..., x_n). \tag{11}$$

range() zur Berechnung der Spannbreite

```
x       = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
x_max   = max(x)                        # Maximum der TA1 Werte
x_min   = min(x)                        # Mininum der TA1 Werte
S       = x_max - x_min                 # Spannbreite
[1] 26
MinMax  = range(x)                      # "automatische" Berechnung von min(x), max(x)
[1] 0 26
S       = MinMax[2] - MinMax[1]         # Spannbreite
```

Stichprobenvarianz

Definition (Stichprobenvarianz, empirische Stichprobenvarianz)

$x = (x_1, ..., x_n)$ sei ein Datensatz. Die *Stichprobenvarianz* von x ist definiert als u

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

und die *empirische Stichprobenvarianz* von x ist definiert als

$$\tilde{S}^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Inferenzstatistische Anmerkungen

- S^2 ist ein unverzerrter Schätzer von $\mathbb{V}(X)$, $\tilde{S}^2$ ist ein verzerrter Schätzer $\mathbb{V}(X)$.
- Für $n \to \infty$ gilt $\frac{1}{n} \approx \frac{1}{n-1}$, $\tilde{S}^2$ ist ein asymptotisch unverzerrter Schätzer von $\mathbb{V}(X)$.
- $\tilde{S}^2$ ist der ML Schätzer, S^2 ist der ReML Schätzer von σ^2 bei $X_1, ..., X_n \sim N(\mu, \sigma^2)$.
- Es gelten $\tilde{S}^2 = \frac{n-1}{n} S^2$, $S^2 = \frac{n}{n-1} \tilde{S}^2$ und $0 \leq \tilde{S}^2 \leq S^2$.

Stichprobenvarianz

var() zum Berechnen der Stichprobenvarianz

```
x         = affect$TA1                          # double Vektor der Tense Arousal 1 Werte
n         = length(x)                           # Anzahl der Werte
S2        = (1/(n-1))*sum((x - mean(x))^2)      # Stichprobenvarianz
[1] 19.53934
S2        = var(x)                              # "automatische" Stichprobenvarianz
[1] 19.53934
S2_tilde  = (1/n)*sum((x - mean(x))^2)          # Empirische Stichprobenvarianz
[1] 19.48013
S2_tilde  = ((n-1)/n)*var(x)                    # "automatische" emp. Stichprobenvarianz
[1] 19.48013
```

Stichprobenvarianz

Für einen Datensatz $x_1, ..., x_n$, den linear-affin transformierten Datentsatz $y_1 := ax_1 + b, ..., y_n := ax_n + b$ und mit Stichprobenvarianzen S_x^2 und S_y^2 der jeweiligen Datensätze gilt

$$S_y^2 = a^2 S_x^2.$$

Beweis

$$\begin{aligned}
S_y^2 &:= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (ax_i + b - (a\bar{x} + b))^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (ax_i + b - a\bar{x} - b)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (a(x_i - \bar{x}))^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} a^2 (x_i - \bar{x})^2 \\
&= a^2 \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= a^2 S_x^2
\end{aligned}$$

□

Stichprobenvarianz

Für einen Datensatz $x_1, ..., x_n$, den linear-affin transformierten Datentsatz $y_1 := ax_1 + b, ..., y_n := ax_n + b$ und mit Stichprobenvarianzen S_x^2 und S_y^2 der jeweiligen Datensätze gilt

$$S_y^2 = a^2 S_x^2.$$

```
x     = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
S2x   = var(x)                        # Stichprobenvarianz von x_1,...,x_n
a   = 2                               # Multiplikationskonstante
b     = 5                             # Additionskonstante
y     = a*x + b                       # y_i = ax_i + b
S2y   = var(y)                        # Stichprobenvarianz y_1,...,y_n
[1] 78.15737
S2y   = a^2 * S2x                     # Stichprobenvarianz y_1,...,y_n
[1] 78.15737
```

Stichprobenvarianz

Für einen Datensatz x und sein elementweises Quadrat x^2 gilt

$$\tilde{S}^2 = \overline{x^2} - \bar{x}^2 \tag{12}$$

Beweis

$$\begin{aligned}
\tilde{S}^2 &:= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\
&= \overline{x^2} - 2\bar{x}\bar{x} + \frac{1}{n}n\bar{x}^2 \\
&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \overline{x^2} - \bar{x}^2
\end{aligned}$$

□

Stichprobenvarianz

Für einen Datensatz x und sein elementweises Quadrat x^2 gilt

$$\tilde{S}^2 = \overline{x^2} - \bar{x}^2 \tag{13}$$

```
x          = affect$TA1                    # double Vektor der Tense Arousal 1 Werte
x_bar      = mean(x)                       # Stichprobenmittel
S2_tilde   = ((n-1)/n)*var(x)              # empirische Stichprobenvarianz
[1] 19.48013
S2_tilde   = mean(x^2) - (mean(x))^2       # \bar{x^2} - \bar{x}^2
[1] 19.48013
S2         = var(x)                        # S^2 \neq \bar{x^2} - \bar{x}^2
[1] 19.53934
```

Stichprobenstandardabweichung

Definition (Stichprobenstandardabweichung, empirische)

$x = (x_1, ..., x_n)$ sei ein Datensatz. Die *Stichprobenstandardabweichung* von x ist definiert als

$$S := \sqrt{S^2}$$

und die *empirische Stichprobenstandardabweichung* von x ist definiert als

$$\tilde{S} := \sqrt{\tilde{S}^2}.$$

Inferenzstatistische Anmerkungen

- S ist ein verzerrter Schätzer von $\mathbb{S}(X)$.
- S^2 misst Variabilität in quadrierten Einheiten, zum Beispiel Quadratmeter (m^2).
- S misst Variabilität in unquadrierten Einheiten, zum Beispiel Meter (m).
- Es gilt $\tilde{S} = \sqrt{(n-1)/n}S$.

Stichprobenstandardabweichung

sd() zur Berechnung der Stichprobenstandardabweichung

```
x      = affect$TA1                                  # double Vektor der TA 1 Werte
n      = length(x)                                   # Anzahl der Werte
S      = sqrt((1/(n-1))*sum((x - mean(x))^2))        # Standardabweichung
[1] 4.420333
S      = sd(x)                                       # "automatische" Berechnung
[1] 4.420333
S_tilde = sqrt((1/(n))*sum((x - mean(x))^2))         # empirische Standardabweichung
[1] 4.41363
S_tilde   = sqrt((n-1)/n)*sd(x)                      # empirische Standardabweichung
[1] 4.41363
```

Stichprobenstandardabweichung

Für einen Datensatz $x_1, ..., x_n$, den linear-affin transformierten Datentsatz $y_1 := ax_1 + b, ..., y_n := ax_n + b$ und mit Stichprobenstandardabweichung S_x und S_y der jeweiligen Datensätze gilt

$$S_y = |a| S_x.$$

Beweis

$$\begin{aligned}
S_y &:= \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2} \\
&= \left(\frac{1}{n-1} \sum_{i=1}^{n} (ax_i + b - (a\bar{x} + b))^2 \right)^{1/2} \\
&= \left(\frac{1}{n-1} \sum_{i=1}^{n} (a(x_i - \bar{x}))^2 \right)^{1/2} \\
&= \left(\frac{1}{n-1} \sum_{i=1}^{n} a^2 (x_i - \bar{x})^2 \right)^{1/2} \\
&= \left(a^2\right)^{1/2} \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2}
\end{aligned}$$

Also gilt $S_y = aS_x$, wenn $a \geq 0$ und $S_y = -aS_x$, wenn $a < 0$. Dies aber entspricht $S_y = |a| S_x$. □

Stichprobenvarianz

Für einen Datensatz $x_1, ..., x_n$, den linear-affin transformierten Datentsatz $y_1 := ax_1 + b, ..., y_n := ax_n + b$ und mit Stichprobenstandardabweichung S_x und S_y der jeweiligen Datensätze gilt

$$S_y = |a| S_x.$$

```
# a >= 0
x   = affect$TA1                    # double Vektor der TA1 Werte
Sx  = sd(x)                         # Stichprobenvarianz von x
a = 2                               # Multiplikationskonstante
b   = 5                             # Additionskonstante
y   = a*x + b                       # y_i = ax_i + b
Sy  = sd(y)                         # Stichprobenvarianz von y
Sy  = a*Sx                          # Stichprobenvarianz von y

# a < 0
x   = affect$TA1                    # double Vektor der TA1 Werte
Sx  = sd(x)                         # Stichprobenvarianz von x
a = -3                              # Multiplikationskonstante
b   = 10                            # Additionskonstante
y   = a*x + b                       # y_i = ax_i + b
Sy  = sd(y)                         # Stichprobenvarianz von y
Sy  = (-a)*Sx                       # Stichprobenvarianz von y
```

Deskriptive Statistiken

- Verteilungsdarstellung
- Maße der zentralen Tendenz
- Maße der Datenvariabilität
- **Bivariate Deskriptivstatistik**
- Übungen und Selbstkontrollfragen

Bivariate Datensätze

- Wir bezeichnen einen bivariaten Datensatz mit

$$x = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \,. \tag{14}$$

- x_i und y_i bezeichnen das erste und zweite Merkmal der $i$ten Einheit, respektive.
- n ist die Anzahl an bivariaten Datenpunkten (x_i, y_i).
- Untenstehende Tabelle zeigt ein Beispiel.

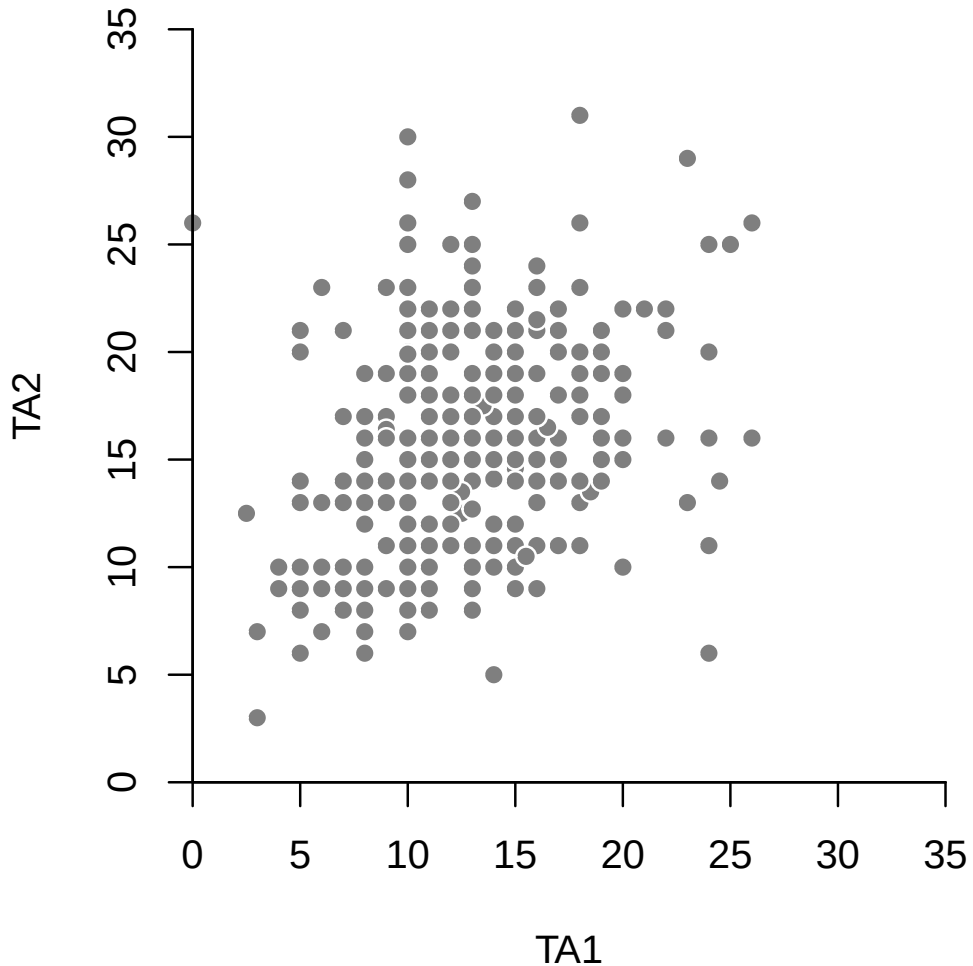
i	1	2	3	4	5	6	7	8	9	10
x_i	3.4	1.5	2.7	4.5	6.1	3.8	2.0	2.5	6.2	9.1
y_i	5.5	7.3	1.1	1.9	4.5	2.3	8.4	8.6	3.9	1.6

Streudiagramme

Die Darstellung der $(x_i, y_i), i = 1, ..., n$ in einem Koordinatensystem heißt *Streudiagramm*.

```
x       = affect$TA1                    # atomic vector x_1,...,x_n
y       = affect$TA2                    # atomic vector y_1,...,y_n
dev.new()                               # Abbildungsinitialisierung
par(                                    # Abbildungsparameter
pty     = "s",                          # Square Abbildung
bty     = "l",                          # Plot Box L
xaxs    = "i",                          # internal (tight) x Achsenstil
yaxs    = "i")                          # internal (tight) y Achsenstil
plot(                                   # Visualisierung
x,                                      # x
y,                                      # y
type    = "p",                          # Punkte
pch     = 21,                           # Punkttyp (?pch)
col     = "white",                      # Punktlinienfarbe
bg      = "gray50",                     # Punktfuellfarbe
xpd     = TRUE,                         # Punkte vor Achsen
cex     = 1.1,                          # Punktgroessenfaktor
xlab    = "TA1",                        # x Achsenbeschriftung
ylab    = "TA2",                        # y Achsenbeschriftung
xlim    = c(0,35),                      # x Achsengrenzen
ylim    = c(0,35))                      # y Achsengrenzen
```

Streudiagramme



Bivariate Histogramme

Definition (Bivariates Histogramm)

Ein *bivariates Histogramm* ist ein Diagramm, in dem zu einem Datensatz $x = ((x_i, y_i))$ mit $i = 1, ..., n$ über den Rechteckklassen

$$[b^x_{j-1}, b^x_j[\times [b^y_{k-1}, b^y_k[\text{ für } j = 1, ..., m_x, k = 1, ..., m_y \tag{15}$$

Blöcke mit Grundkante $[b^x_{j-1}, b^x_j[$ in der x-Ordinate und Grundkante $[b^y_{k-1}, b^y_k[$ in der y-Ordinate und Höhe

$$h_{jk} := \text{Anzahl der } (x_i, y_i) \text{ in } x \text{ mit } (x_i, y_i) \in [b^x_{j-1}, b^x_j[\times [b^y_{k-1}, b^y_k[\tag{16}$$

bzw.

$$r_{jk} := \frac{h_{jk}}{n} \tag{17}$$

abgebildet sind.

Bemerkungen

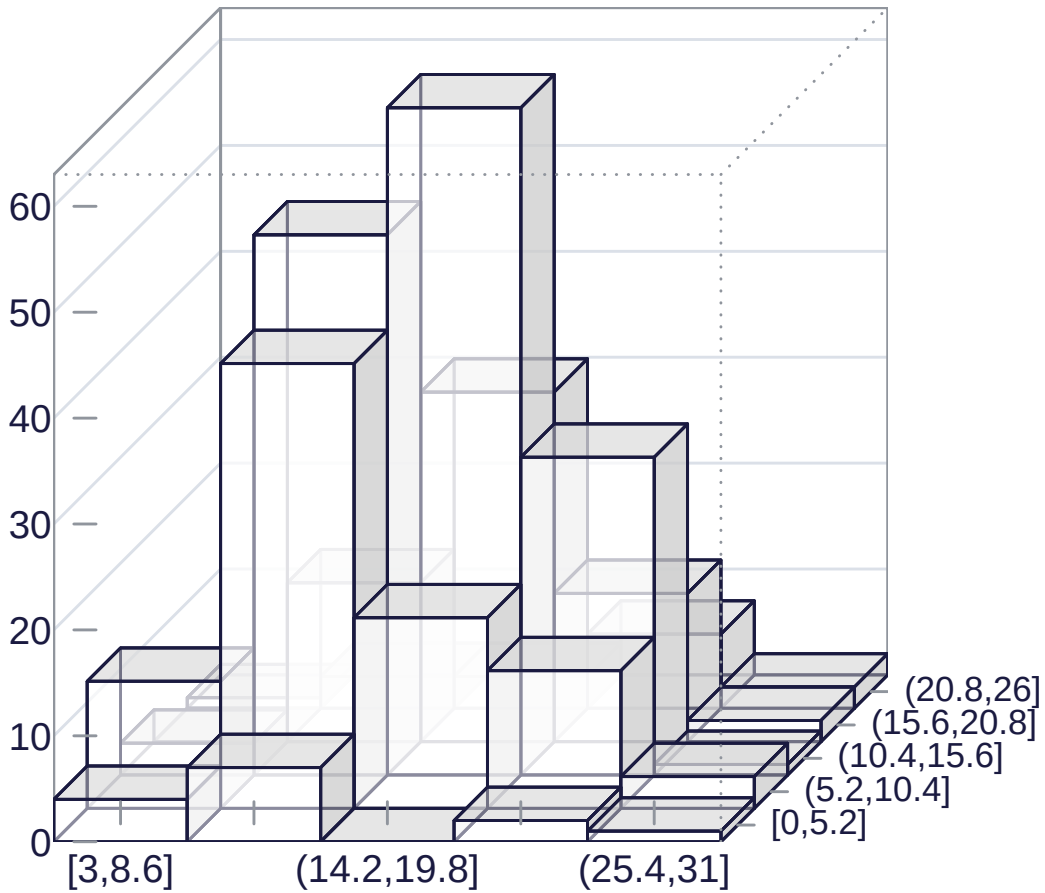
- Ein bivariates Histogramm ist die Generalisierung eines Histogramms für zwei Dimensionen.
- Das Aussehen eines bivariaten Histogramms hängt stark von der Wahl der Rechteckklassen ab.

Bivariate Histogramme

Erzeugung eines einfachen bivariaten Histogramms mithilfe der Pakete **gplots** und **epade**

```
library(gplots)                                  # hist2d Funktionalitaet
library(epade)                                   # bar3d Funktionalitaet
x      = affect$TA1                              # x_1,...,x_n
y      = affect$TA2                              # y_1, ..., y_n
h      = hist2d(x,y,nbins = c(5,5),show = F)     # Bivariate Haeufigkeitenbestimmung
dev.new()                                        # Abbildungsinitialisierung
par(                                             # Abbildungsparameter
pty    = "s",                                    # Square Abbildung
bty    = "l",                                    # Plot Box L
xaxs   = "i",                                    # "internal" (tight) x Achsenstil
yaxs   = "i")                                    # "internal" (tight) y Achsenstil
bar3d.ade(                                       # 3D Blockvisualisierung
h$counts,                                        # Tabelle der h_{ij} Werte
col    = "white",                                # Blockfarbe
wall   = 2,                                      # Dekorationsstil
xw     = 1)                                      # x Blockbreite
```

Bivariate Histogramme



Stichprobenkovarianz

Definition (Empirische Stichprobenkovarianz)

$x = ((x_1, y_1), ..., (x_n, y_n))$ sei ein bivariater Datensatz. Dann heißt die Zahl

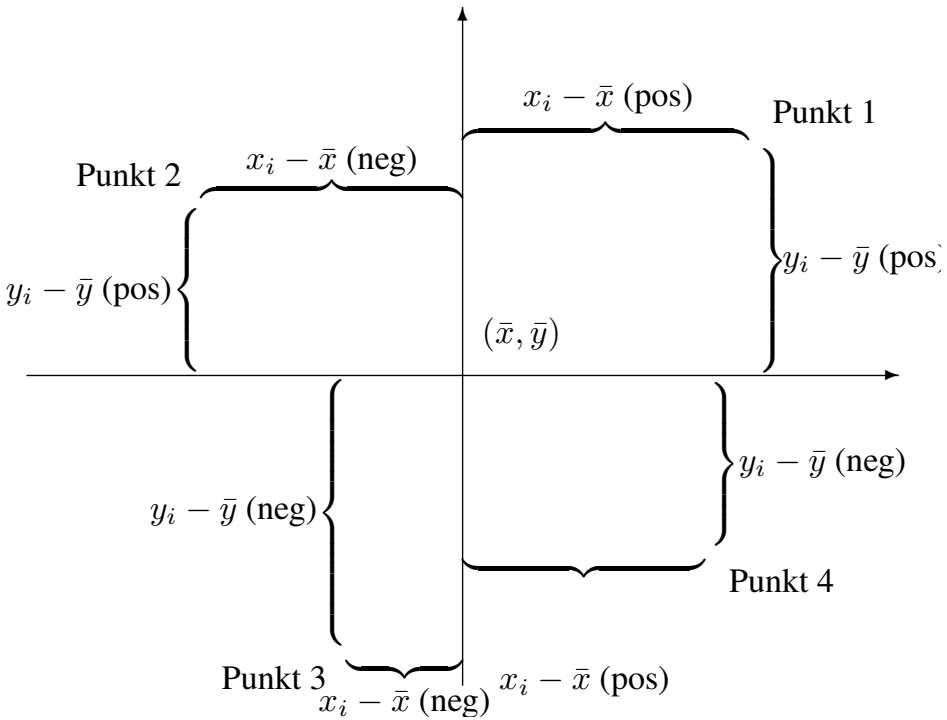
$$C := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{18}$$

empirische Stichprobenkovarianz von $x_1, ..., x_n$ und $y_1, ..., y_n$.

Bemerkungen

- $\bar{x}$ und $\bar{y}$ bezeichnen die Mittelwerte der $x_1, ..., x_n$ und $y_1, ..., y_n$, respektive.
- Der Faktor $1/n$ normiert für die Stichprobengröße.
- Eine Intuition vermittelt untenstehende Abbildung (Fahrmeir et al., 2016, Kapitel 3.4)

Stichprobenkovarianz



	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
Punkt 1 (1.Quadrant)	positiv	positiv	positiv
Punkt 2 (2.Quadrant)	negativ	positiv	negativ
Punkt 3 (3.Quadrant)	negativ	negativ	positiv
Punkt 4 (4.Quadrant)	positiv	negativ	negativ

Abbildung 3.10 und Tabelle 3.9 aus Fahrmeir et al. (2016, Kapitel 3.4)

Pearson's Stichprobenkorrelationskoeffizient

Definition (Pearson's Stichprobenkorrelationskoeffizient)

$x = ((x_1, y_1), ..., (x_n, y_n))$ sei ein bivariater Datensatz, C sei die empirische Stichprobenvarianz von $x_1, ..., x_n$ und $y_1, ..., y_n$ und S_x bzw. S_y seien die empirischen Stichprobenstandardabweichungen von $x_1, ..., x_n$ bwz. $y_1, ..., y_n$. Dann heißt die Zahl

$$r := \frac{C}{S_x S_y} \tag{19}$$

Pearson's Stichprobenkorrelationskoeffizient oder *empirischer Korrelationskoeffizient*.

Bemerkungen

- Es gilt $-1 \leq r \leq 1$
- r misst die Stärke des positive oder negativen linearen Zusammenhangs von x und y.
- Eine Intuition vermittelt untenstehende Abbildung (Fahrmeir et al., 2016, Kapitel 3.4)
- Korrelationstärken werden in etwa nach

$$|r| < 0.5,\ 0.5 \leq |r| \leq 0.8,\ |r| > 0.8 \tag{20}$$

als "schwache", "mittelere",oder "starke" Korrelation bezeichnet

Pearson's Stichprobenkorrelationskoeffizient

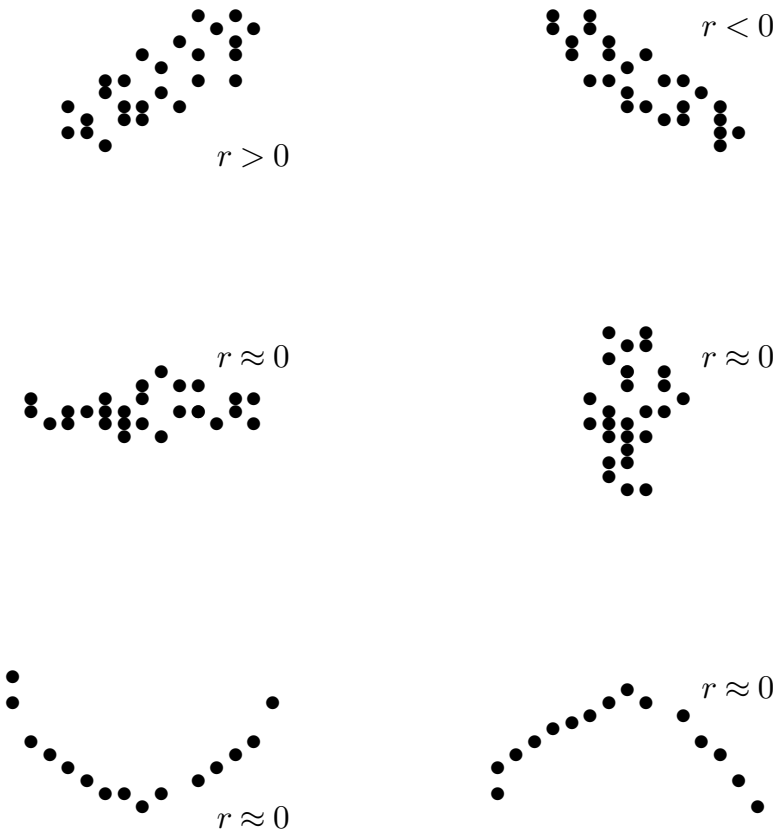


Abbildung 3.11 aus Fahrmeir et al. (2016, Kapitel 3.4)

Pearson's Stichprobenkorrelationskoeffizient

```
# Bivariater Datensatz
x      = affect$TA1                                    # x_1,..., x_n
y      = affect$TA2                                    # y_1,..., y_n

"manuelle" Berechnung
n      = length(x)                                      # Anzahl Datenpunkte
x_bar  = mean(x)                                        # \bar{x}
y_bar  = mean(y)                                        # \bar{y}
c_xy   = (1/n)*sum((x - mean(x))*(y - mean(y)))         # c
s_x    = sqrt((1/n)*sum((x - mean(x))^2))               # s_x
s_y    = sqrt((1/n)*sum((y - mean(y))^2))               # s_y
r_m    = c_xy/(s_x*s_y)                                 # "manuelles" r
[1] 0.3117369

# "automatische" Berechnung
r_a    = cor(x,y, method = "pearson")                   # "automatisches" r
[1] 0.3117369
```

Univariate Deskriptivstatistiken

- Verteilungsdarstellung
- Maße der zentralen Tendenz
- Maße der Datenvariabilität
- **Übungen und Selbstkontrollfragen**

Übungen und Selbstkontrollfragen

1. Dokumentieren Sie die in dieser Einheit eingeführten R Befehle in einem R Skript.

2. x sei ein als double vector vorliegender univariater Datensatz, z.B. x = affect$TA2.

   - Berechnen Sie Minimum, Maximum, Median, und Interquartilsabstand von x.
   - Erzeugen Sie einen Boxplot von x
   - Berechnen Sie Mittelwert, Median, Varianz, und Standardabweichung von x.
   - Erzeugen Sie ein Histogramm von x.
   - Visualisieren Sie die empirische Verteilungsfunktion von x.

3. x und y seien zwei gleich große Datensätze, z.B. x = affect$ext und y = affect$neur.

   - Stellen Sie x und y in einem Streudiagramm dar.
   - Berechnen Sie die empirische Kovarianz von x und y.
   - Berechnen Sie Pearson's Stichprobenkorrelationskoeffizienten zu x und y.

Literatur

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., and Tutz, G. (2016). *Statistik*. Springer-Lehrbuch. Springer Berlin Heidelberg, Berlin, Heidelberg.

Henze, N. (2018). *Stochastik für Einsteiger*. Springer Fachmedien Wiesbaden, Wiesbaden.

Hyndman, R. J. and Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, 50(4):361.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, 32(1):12.

Rafaeli, E. and Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion*, 30(1):1–12.

Scott, D. W. (1979). On optimal and data-based histograms. page 6.

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153):65–66.

Thayer, R. E. (1986). Activation-Deactivation Adjective Check List: Current Overview and Structural Analysis. *Psychological Reports*, 58(2):607–614.