



Programmieren für Psychologie

MSc Umweltpsychologie/Mensch-Technik-Interaktion

MSc Klinische Psychologie und Psychotherapie

MSc Psychologie

Joram Soch | Wintersemester 2026/2027

(8) Berechnung von Deskriptivstatistiken

Übersicht

Stichprobenmittel

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (1)$$

Stichprobenvarianz

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2)$$

Stichprobenkovarianz

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (3)$$

Stichprobenkorrelation

$$r_{xy} = \frac{c_{xy}}{\sqrt{s_x^2} \sqrt{s_y^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

Beispiel

Beispieldatensatz: (x_i, y_i) für $i = 1, \dots, 20$ von $n = 40$

```
# Einlesen des Datensatzes
D = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n = nrow(D)                            # Anzahl Datenpunkte

# Ausgabe des Dataframes
library(knitr)                          # R-Paket laden
knitr::kable(D[1:20,], digits = 2)     # Tabellenoutput
```

DUR	dBDI
1.37	9
2.18	8
1.16	11
3.60	0
2.33	8
1.18	7
2.49	3
2.74	7
2.58	3
1.69	11
3.51	9
2.39	7
1.38	15
-0.21	11
3.12	4
1.96	6
1.98	8
2.94	6
2.82	5
2.59	13

Stichprobenmittel

Stichprobenvarianz

Stichprobenkovarianz

Stichprobenkorrelation

Stichprobenmittel

Stichprobenvarianz

Stichprobenkovarianz

Stichprobenkorrelation

Definition (Stichprobenmittel)

$y = (y_1, \dots, y_n)$ mit $y_i \in \mathbb{R}$ für $i = 1, \dots, n$ sei ein univariater Datensatz. Dann ist das Stichprobenmittel von y definiert als

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i . \quad (5)$$

Bemerkung

- Das Stichprobenmittel wird auch als *arithmetisches Mittel* oder *Durchschnitt* bezeichnet.

Berechnung mit R base

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

x_i    = D[,1] # = D$DUR                    # x_i-Werte
y_i    = D[,2] # = D$dBDI                  # y_i-Werte
x_bar  = 1/n * sum(x_i)                    # Stichprobenmittel
y_bar  = 1/n * sum(y_i)                    # Stichprobenmittel

cat( "Stichprobenmittel der x_i : ", x_bar, # Ergebnisausgabe
     "\nStichprobenmittel der y_i : ", y_bar)
```

Stichprobenmittel der x_i : 3.092026

Stichprobenmittel der y_i : 3.975

Berechnung mit R-Funktionen

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

x_bar = mean(D[,1])                         # Stichprobenmittel
y_bar = mean(D[,2])                         # Stichprobenmittel

cat( "Stichprobenmittel der x_i : ", x_bar, # Ergebnisausgabe
     "\nStichprobenmittel der y_i : ", y_bar)
```

Stichprobenmittel der x_i : 3.092026

Stichprobenmittel der y_i : 3.975

Berechnung mit dplyr

```
D      = read.csv("Daten/Daten_2.csv")    # Daten einlesen
n      = nrow(D)                          # Anzahl Datenpunkte

library(dplyr)                            # R-Paket laden
D_mean = D %>%                            # Übergabe des Dataframes
  summarise(across(everything(), mean)) %>% # Berechnung Stichprobenmittel
  as.data.frame()                          # Konvertierung in Dataframe

rownames(D_mean) = c("Mean")              # Zeile benennen
print(D_mean)                             # Ergebnisausgabe
```

```
DUR  dBDI
Mean 3.092026 3.975
```

Stichprobenmittel

Stichprobenvarianz

Stichprobenkovarianz

Stichprobenkorrelation

Definition (Stichprobenvarianz)

$y = (y_1, \dots, y_n)$ mit $y_i \in \mathbb{R}$ für $i = 1, \dots, n$ sei ein univariater Datensatz. Dann ist die Stichprobenvarianz von y definiert als

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad (6)$$

wobei $\bar{y}$ das Stichprobenmittel von y ist.

Bemerkung

- Die Stichprobenvarianz entspricht der Summe der quadrierten Abweichungen der Datenpunkte von ihrem Stichprobenmittel, geteilt durch Stichprobenumfang minus 1.

Definition (Stichprobenstandardabweichung)

$y = (y_1, \dots, y_n)$ mit $y_i \in \mathbb{R}$ für $i = 1, \dots, n$ sei ein univariater Datensatz. Dann ist die Stichprobenstandardabweichung von y definiert als

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7)$$

wobei $\bar{y}$ das Stichprobenmittel von y ist.

Bemerkung

- Die Stichprobenstandardabweichung ist die Quadratwurzel aus der Stichprobenvarianz.

Stichprobenvarianz

Berechnung mit R base

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

x_i    = D[,1] # = D$DUR                     # x_i-Werte
y_i    = D[,2] # = D$dBDI                   # y_i-Werte
x_bar  = 1/n * sum(x_i)                     # Stichprobenmittel
y_bar  = 1/n * sum(y_i)                     # Stichprobenmittel
s2_x   = 1/(n-1) * sum((x_i - x_bar)^2)     # Stichprobenvarianz
s2_y   = 1/(n-1) * sum((y_i - y_bar)^2)     # Stichprobenvarianz
s_x    = sqrt(s2_x)                         # Stichprobenstandardabweichung
s_y    = sqrt(s2_y)                         # Stichprobenstandardabweichung
                                              # Ergebnisausgabe
cat( "Stichprobenvarianz der x_i           : ", s2_x,
     "\nStichprobenvarianz der y_i           : ", s2_y,
     "\nStichprobenstandardabweichung der x_i : ", s_x,
     "\nStichprobenstandardabweichung der y_i : ", s_y)
```

```
Stichprobenvarianz der x_i           : 1.609773
Stichprobenvarianz der y_i           : 25.97372
Stichprobenstandardabweichung der x_i : 1.268768
Stichprobenstandardabweichung der y_i : 5.096442
```

Berechnung mit R-Funktionen

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

s2_x   = var(D[,1])                         # Stichprobenvarianz
s2_y   = var(D[,2])                         # Stichprobenvarianz
s_x    = sd(D[,1])                          # Stichprobenstandardabweichung
s_y    = sd(D[,2])                          # Stichprobenstandardabweichung
                                             # Ergebnisausgabe

cat( "Stichprobenvarianz der x_i           : ", s2_x,
     "\nStichprobenvarianz der y_i           : ", s2_y,
     "\nStichprobenstandardabweichung der x_i : ", s_x,
     "\nStichprobenstandardabweichung der y_i : ", s_y)
```

Stichprobenvarianz der x_i : 1.609773

Stichprobenvarianz der y_i : 25.97372

Stichprobenstandardabweichung der x_i : 1.268768

Stichprobenstandardabweichung der y_i : 5.096442

Berechnung mit dplyr

```
D      = read.csv("Daten/Daten_2.csv")    # Daten einlesen
n      = nrow(D)                          # Anzahl Datenpunkte

library(dplyr)                            # R-Paket laden
D_var  = rbind(                            # Berechnung Deskriptivstatistiken
  summarise(D, across(everything(), mean)),
  summarise(D, across(everything(), var)),
  summarise(D, across(everything(), sd))) %>%
  as.data.frame()                          # Konvertierung in Dataframe

rownames(D_var) = c("Mean", "Var", "SD")  # Zeilen benennen
print(D_var)                              # Ergebnisausgabe
```

	DUR	dBDI
Mean	3.092026	3.975000
Var	1.609773	25.973718
SD	1.268768	5.096442

Stichprobenmittel

Stichprobenvarianz

Stichprobenkovarianz

Stichprobenkorrelation

Definition (Stichprobenkovarianz)

$x = (x_1, \dots, x_n)$ und $y = (y_1, \dots, y_n)$ mit $x_i, y_i \in \mathbb{R}$ für $i = 1, \dots, n$ seien zwei univariate Datensätze. Dann ist die Stichprobenkovarianz von x und y definiert als

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) , \quad (8)$$

wobei $\bar{x}$ und $\bar{y}$ die Stichprobenmittel von x bzw. y sind.

Bemerkung

- Die Stichprobenkovarianz entspricht der Summe der multiplizierten Abweichungen der Datenpunkte von ihren Stichprobenmitteln, geteilt durch Stichprobenumfang minus 1.

Berechnung mit R base

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

x_i    = D[,1] # = D$DUR                    # x_i-Werte
y_i    = D[,2] # = D$dBDI                  # y_i-Werte
x_bar  = 1/n * sum(x_i)                    # Stichprobenmittel
y_bar  = 1/n * sum(y_i)                    # Stichprobenmittel
c_xy   = 1/(n-1) * sum((x_i - x_bar)*      # Stichprobenkovarianz
                      (y_i - y_bar))
                                             # Ergebnisausgabe
cat("Stichprobenkovarianz der x_i und y_i : ", c_xy)
```

Stichprobenkovarianz der x_i und y_i : -5.345646

Berechnung mit R-Funktionen

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

c_xy   = cov(D[,1], D[,2])                  # Stichprobenkovarianz
                                              # Ergebnisausgabe
cat("Stichprobenkovarianz der x_i und y_i : ", c_xy)
```

Stichprobenkovarianz der x_i und y_i : -5.345646

Berechnung mit dplyr

```
D      = read.csv("Daten/Daten_2.csv")    # Daten einlesen
n      = nrow(D)                          # Anzahl Datenpunkte

library(dplyr)                            # R-Paket laden
D_cov  = D %>%                             # Übergabe des Dataframes
  summarise("DUR/dBDI" = cov(DUR, dBDI)) %>% # Berechnung Stichkovarianz
  as.data.frame()                          # Konvertierung in Dataframe

rownames(D_cov) = c("Cov")                 # Zeile benennen
print(D_cov)                              # Ergebnisausgabe
```

DUR/dBDI

Cov -5.345646

Stichprobenmittel

Stichprobenvarianz

Stichprobenkovarianz

Stichprobenkorrelation

Definition (Stichprobenkorrelation)

$x = (x_1, \dots, x_n)$ und $y = (y_1, \dots, y_n)$ mit $x_i, y_i \in \mathbb{R}$ für $i = 1, \dots, n$ seien zwei univariate Datensätze. Dann ist die Stichprobenkorrelation von x und y definiert als

$$r_{xy} = \frac{c_{xy}}{\sqrt{s_x^2} \sqrt{s_y^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (9)$$

wobei c_{xy} die Stichprobenkovarianz von x und y ist, während s_x^2 und s_y^2 die Stichprobenvarianzen von x bzw. y sind.

Bemerkung

- Die Stichprobenkorrelation entspricht dem Quotienten aus Stichprobenkovarianz und dem Produkt der Stichprobenstandardabweichungen.

Stichprobenkorrelation

Berechnung mit R base

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

x_i    = D[,1] # = D$DUR                    # x_i-Werte
y_i    = D[,2] # = D$dBDI                  # y_i-Werte
x_bar  = 1/n * sum(x_i)                    # Stichprobenmittel
y_bar  = 1/n * sum(y_i)                    # Stichprobenmittel
s2_x   = 1/(n-1) * sum((x_i - x_bar)^2)     # Stichprobenvarianz
s2_y   = 1/(n-1) * sum((y_i - y_bar)^2)     # Stichprobenvarianz
c_xy   = 1/(n-1) * sum((x_i - x_bar)*      # Stichprobenkovarianz
                      (y_i - y_bar))
r_xy   = c_xy/(sqrt(s2_x) * sqrt(s2_y))     # Stichprobenkorrelation

cat( "Stichprobenvarianz der x_i           : ", s2_x,
     "\nStichprobenvarianz der y_i           : ", s2_y,
     "\nStichprobenkovarianz der x_i und y_i : ", c_xy,
     "\nStichprobenkorrelation der x_i und y_i : ", r_xy)
```

```
Stichprobenvarianz der x_i           : 1.609773
Stichprobenvarianz der y_i           : 25.97372
Stichprobenkovarianz der x_i und y_i : -5.345646
Stichprobenkorrelation der x_i und y_i : -0.8267056
```

Berechnung mit R-Funktionen

```
D      = read.csv("Daten/Daten_2.csv")      # Daten einlesen
n      = nrow(D)                            # Anzahl Datenpunkte

r_xy   = cor(D[,1], D[,2])                  # Stichprobenkorrelation
                                              # Ergebnisausgabe
cat("Stichprobenkorrelation der x_i und y_i : ", r_xy)
```

Stichprobenkorrelation der x_i und y_i : -0.8267056

Berechnung mit dplyr

```
D      = read.csv("Daten/Daten_2.csv")    # Daten einlesen
n      = nrow(D)                          # Anzahl Datenpunkte

library(dplyr)                            # R-Paket laden
D_cov  = D %>%                             # Übergabe des Dataframes
  summarise("DUR/dBDI" = cor(DUR, dBDI)) %>% # Berechnung Stichkovarianz
  as.data.frame()                          # Konvertierung in Dataframe

rownames(D_cov) = c("Corr")                # Zeile benennen
print(D_cov)                              # Ergebnisausgabe
```

DUR/dBDI

Corr -0.8267056