Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2023/24

Belinda Fleischmann

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

(9) Maße der zentralen Tendenz

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Mittelwert

Definition (Mittelwert)

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

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

der *Mittelwert* von x.

Bemerkung

- Im Kontext der Inferenzstatistik heißt der Mittelwert *Stichprobenmittel*.
- Die Inferenzstatistik gibt der Mittelwertsbildung ihren Sinn.

Berechnung des Mittelwerts in R

"Manuelle" Berechnung des Mittelwerts

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

# Mittelwertberechnung
x     <- D$Pre.BDI                  # double Vektor der Pre-BDI Werte
n     <- length(x)                  # Anzahl der Werte
x_bar <- (1 / n) * sum(x)               # Mittelwertsberechnung
print(x_bar)                       # Ausgabe
```

```
[1] 18.61
```

mean() zur Berechnung des Mittelwerts

```
x_bar <- mean(x)                   # Mittelwertsberechnung
print(x_bar)                       # Ausgabe
```

```
[1] 18.61
```

Eigenschaften des Mittelwerts

Theorem (Eigenschaften des Mittelwerts)

$x = (x_1, ..., x_n)$ und sei ein Datensatz und $\bar{x}$ sei der Mittelwert von x. Dann gelten

(1) Die Summe der Abweichungen vom Mittelwert ist Null,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \tag{2}$$

(2) Die absoluten Summen negativer und positiver Abweichungen vom Mittelwert sind gleich, d.h. wenn $j = 1, ..., n_j$ die Datenpunktindizes mit $(x_j - \bar{x}) < 0$ und $k = 1, ..., n_k$ die Datenpunktindizes mit $(x_k - \bar{x}) \geq 0$ bezeichnen, dann gilt mit $n_j + n_k$

$$|\sum_{j=1}^{n_j} (x_j - \bar{x})| = |\sum_{k=1}^{n_k} (x_k - \bar{x})|. \tag{3}$$

(3) Der Mittelwert der Summe zweier gleich großer Datensätze entspricht der Summe ihrer Mittelwerte, d.h. für einen weiteren Datensatz $y = (y_1, ..., y_n)$ mit Mittelwert $\bar{y}$ gilt

$$\overline{x + y} = \bar{x} + \bar{y} \tag{4}$$

(4) Eine linear-affine Transformation eines Datensatz transformiert den Mittelwert des Datensatzes linear-affin, d.h für $a, b \in \mathbb{R}$ gilt

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

Beweis

(1) Es gilt

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \frac{n}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

(2) Seien $j = 1, \dots, n_j$ die Indizes mit $(x_j - \bar{x}) < 0$ und $k = 1, \dots, n_k$ die Indizes mit $(x_k - \bar{x}) \geq 0$, so dass $n = n_j + n_k$. Dann gilt

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \Leftrightarrow \sum_{j=1}^{n_j}(x_j - \bar{x}) + \sum_{k=1}^{n_k}(x_k - \bar{x}) = 0 \Leftrightarrow \sum_{k=1}^{n_k}(x_k - \bar{x}) = -\sum_{j=1}^{n_j}(x_j - \bar{x})$$

$$\Leftrightarrow |\sum_{j=1}^{n_j}(x_j - \bar{x})| = |\sum_{k=1}^{n_k}(x_k - \bar{x})|.$$

Eigenschaften des Mittelwerts

Beweis

(3) Es gilt

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

(4) Es gilt

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

Eigenschaften des Mittelwerts

Summe der Abweichungen

```
x <- D$Pre.BDI                          # double Vektor der Werte
s <- sum(x - mean(x))                   # Summe der Abweichungen vom Mittelwert
print(s)                                # Ausgabe (Ergebnis hat Rundungsfehler)
```

```
[1] 5.684342e-14
```

Beträge der positiven und negativen Abweichungen

```
x   <- D$Pre.BDI                        # double Vektor der  Werte
s_1 <- sum(x[x <= mean(x)] - mean(x))   # Summe aller negativer Abweichungen
s_2 <- sum(x[x > mean(x)]  - mean(x))   # Summe aller positiver Abweichungen
print(abs(s_1))                         # Ausgabe des Betrags
```

```
[1] 71.28
```

```
print(abs(s_2))                         # Ausgabe des Betrags
```

```
[1] 71.28
```

Eigenschaften des Mittelwerts

Summation von Datensätzen

```
x        <- D$Pre.BDI                  # double Vektor der Pre.BDI-Werte
x_bar    <- mean(x)                    # Mittelwert der Pre.BDI-Werte
y        <- D$Post.BDI                 # double Vektor der Post.BDI Werte
y_bar    <- mean(y)                    # Mittelwert der Post.BDI Werte
z        <- x + y                      # double Vektor der Summe der Pre und Post Werte
z_bar    <- mean(z)                    # Mittelwert der Summe der Werte
print(z_bar)                           # Ausgabe
```

```
[1] 31.68
```

```
xy_bar  <- x_bar + y_bar               # Summe der Mittelwerte der Pre- und Post.BDI Werte
print(xy_bar)                          # Ausgabe
```

```
[1] 31.68
```

Linear-affine Transformation

```
x         <- D$Pre.BDI                 # double Vektor der Pre.BDI Werte
x_bar     <- mean(x)                   # Mittelwert der Pre.BDI Werte
a         <- 2                         # Multiplikationskonstante
b         <- 5                         # Additionskonstante
y         <- a*x + b                   # linear-affine Transformation der Pre.BDI Werte
y_bar     <- mean(y)                   # Mittelwert der transfomierten Pre.BDI Werte
print(y_bar)                           # Ausgabe
```

```
[1] 42.22
```

```
ax_bar_b <-a*x_bar + b                 # Transformation des Mittelwerts
print(ax_bar_b)                        # Ausgabe
```

```
[1] 42.22
```

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

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

$$\tilde{x} := \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{falls } n \text{ gerade} \end{cases} \tag{5}$$

Bemerkungen

- Der Median ist identisch mit dem 0.5-Quantil.
- Mindestens 50% aller x_i sind kleiner oder gleich $\tilde{x}$
- Mindestens 50% aller x_i sind größer oder gleich $\tilde{x}$.
- Anstelle eines Beweises verweisen wir auf untenstehende Abbildungen.

Median - Beispiele

Beispiele

Beispiel für n ungerade

$$n \coloneqq 5 \Rightarrow \left(\frac{5+1}{2}\right) = (3) \Rightarrow \tilde{x} \coloneqq x_{(3)}$$

$$x_{(1)}, x_{(2)}, x_{(3)} \leq \tilde{x} \leq x_{(3)}, x_{(4)}, x_{(5)}$$

Beispiel für n gerade

$$n \coloneqq 6 \Rightarrow \left(\frac{6}{2}\right) = (3), \left(\frac{6}{2}+1\right) = (4) \Rightarrow \tilde{x} \coloneqq \frac{1}{2}\left(x_{(3)} + x_{(4)}\right)$$

$$x_{(1)}, x_{(2)}, x_{(3)} < \tilde{x} < x_{(4)}, x_{(5)}, x_{(6)}$$

Berechnung des Medians in R

Manuelle Bestimmung des Medians

```
x   <- D$Pre.BDI                            # double Vektor der Pre.BDI 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/2 + 1])/2
}
print(x_tilde)
```

```
[1] 19
```

Berechnung des Medians mit `median()`

```
x_tilde <- median(x)                        # Berechnung des Medians
print(x_tilde)                              # Ausgabe
```

```
[1] 19
```

Median vs. Mittelwert

Der Median ist weniger anfällig für Ausreißer als der Mittelwert

```
x       <- D$Pre.BDI                    # double Vektor der Pre.BDI Werte
x_bar   <- mean(x)                      # Mittelwert der Pre.BDI Werte
x_tilde <- median(x)                    # Median der Pre.BDI Werte
print(x_bar)                            # Ausgabe
```

```
[1] 18.61
```

```
print(x_tilde)                          # Ausgabe
```

```
[1] 19
```

```
y       <- x                            # neuer Datensatz mit ...
y[1]    <- 10000                        # ... einem Extremwert
y_bar   <- mean(y)                      # Mittelwert des neuen Datensatzes
print(y_bar)                            # Ausgabe
```

```
[1] 118.44
```

```
y_tilde <- median(y)                    # Mittelwert des neuen Datensatzes
print(y_tilde)                          # Ausgabe
```

```
[1] 19
```

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Modalwert

Definition (Modalwert)

$x := (x_1, ..., x_n)$ mit $x_i \in \mathbb{R}$ sei ein Datensatz, $A := \{a_1, ..., a_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden verschiedenen Zahlenwerte und $h : A \to \mathbb{N}$ sei die absolute Häufigkeitsverteilung der Zahlwerte von x. Dann ist der *Modalwert (oder Modus)* von x definiert als

$$\operatorname{argmax}_{a \in A} h(a), \tag{6}$$

also der am häufigsten im Datensatz vorkommende Wert.

Bemerkungen

- Modalwerte sind nur bei Datensätzen mit Datenpunktwiederholungen sinnvoll.

Bestimmung des Modalwertes in R

```
x        <- D$Pre.BDI                    # double Vektor der Pre.BDI Werte
n        <- length(x)                    # Anzahl der Datenwerte (100)
H        <- as.data.frame(table(x))      # absolute Haeufigkeitsverteilung (dataframe)
names(H) <- c("a", "h")                  # Konsistente Benennung
mod      <- H$a[which.max(H$h)]          # Modalwert
print(as.numeric(as.vector(mod)))        # Ausgabe als numeric vector, nicht factor
```

```
[1] 18
```

Mittelwert

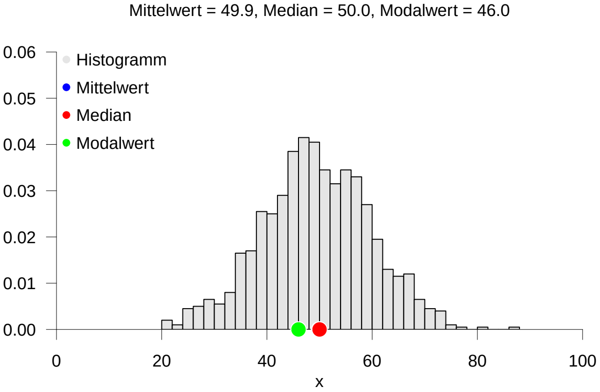
Median

Modalwert

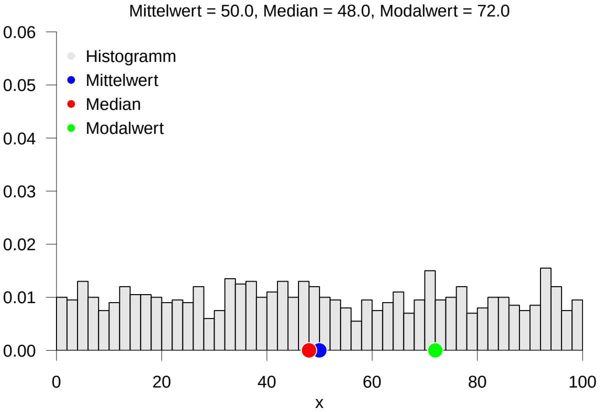
Visuelle Intuitionen

Übungen und Selbstkontrollfragen

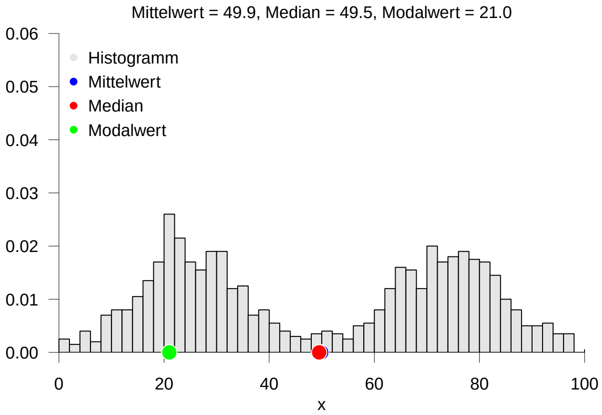
Visuelle Intuition zu Maßen zentraler Tendenz bei **Normalverteilung**



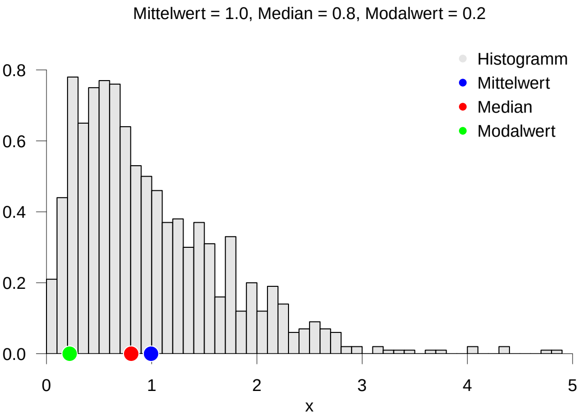
Visuelle Intuition zu Maßen zentraler Tendenz bei **Gleichverteilung**



Visuelle Intuition zu Maßen zentraler Tendenz bei **bimodalen Verteilungen**



Vis. Intuition zu Maßen zentr. T. bei **nicht-symmetrischen Verteilungen**



Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Übungen und Selbskontrollfragen

1. Geben Sie die Definition des Mittelwertes eines Datensatzes wieder.
2. Berechnen Sie den Mittelwert der Post.BDI Daten.
3. Geben Sie das Theorem zu den Eigenschaften des Mittelwerts wieder.
4. Geben Sie die Definition des Median eines Datensatzes wieder.
5. Berechnen Sie den Median der Post.BDI Daten.
6. Wie verhalten sich Mittelwert und Median in Bezug auf Datenausreißer?
7. Geben Sie die Definition des Modalwertes eines Datensatzes wieder.
8. Berechnen Sie den Modalwert des Post.BDI Datensatzes.
9. Visualisieren Sie die Häufigkeitsverteilung des Post.BDI Datensatzes und diskutieren Sie die berechneten Werte von Mittelwert, Median und Modalwert vor dem Hintergrund dieser Häufigkeitsverteilung.