



Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2022/23

Belinda Fleischmann

Inhalte basieren auf Programmierung und Deskriptive Statistik von Dirk Ostwald, lizenziert unter CC BY-NC-SA 4.0

(9) Maße der Zentralen Tendenz

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Definition (Mittelwert)

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

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

der *Mittelwert* von x .

Bemerkung

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

Berechnung des Mittelwerts

```
# Einlesen des Beispieldatensatzes und Abbildungsverzeichnisdefinition
fname = file.path(data_dir, "psychotherapie_datensatz.csv")
D      = read.table(fname, 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)        # Mittelwertberechnung
print(x_bar)                 # Ausgabe
```

```
> [1] 18.6
```

mean() zur Berechnung des Mittelwerts

```
x_bar = mean(x)           # Mittelwertberechnung
print(x_bar)              # Ausgabe
```

```
> [1] 18.6
```

Theorem (Eigenschaften des Mittelwerts)

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

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

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

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

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

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

$$\overline{x + y} = \bar{x} + \bar{y} \quad (4)$$

- (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

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}) = 0 &\Leftrightarrow \sum_{j=1}^{n_j} (x_j - \bar{x}) + \sum_{k=1}^{n_k} (x_k - \bar{x}) = 0 \Leftrightarrow \sum_{k=1}^{n_k} (x_k - \bar{x}) = - \sum_{j=1}^{n_j} (x_j - \bar{x}) \\ &\Leftrightarrow \left| \sum_{j=1}^{n_j} (x_j - \bar{x}) \right| = \left| \sum_{k=1}^{n_k} (x_k - \bar{x}) \right|. \end{aligned}$$

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

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

Mittelwert

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)                    # Rundungsfehler
```

```
> [1] 5.68e-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(s_1)                    # Ausgabe
```

```
> [1] -71.3
```

```
print(s_2)                    # Ausgabe
```

```
> [1] 71.3
```

Mittelwert

Summation von Datensätzen

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

```
> [1] 31.7
```

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

```
> [1] 31.7
```

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 transformierten Pre.BDI Werte
print(y_bar)               # Ausgabe
```

```
> [1] 42.2
```

```
ax_bar_b = a*x_bar + b     # Transformation des TA1 Mittelwerts
print(ax_bar_b)            # Ausgabe
```

```
> [1] 42.2
```

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Definition (Median)

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

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

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

Beispiel für n ungerade

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



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

Beispiel für n gerade

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



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

Median

Berechnung 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

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] 18.6
```

```
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
```

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

```
> [1] 19
```

Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Definition (Modalwert)

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

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

also der am häufigsten im Datensatz vorkommende Wert.

Bemerkungen

- Modalwerte sind nur bei Datensätzen mit Datenpunktwiederholungen sinnvoll.

Bestimmung des Modalwertes

```
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

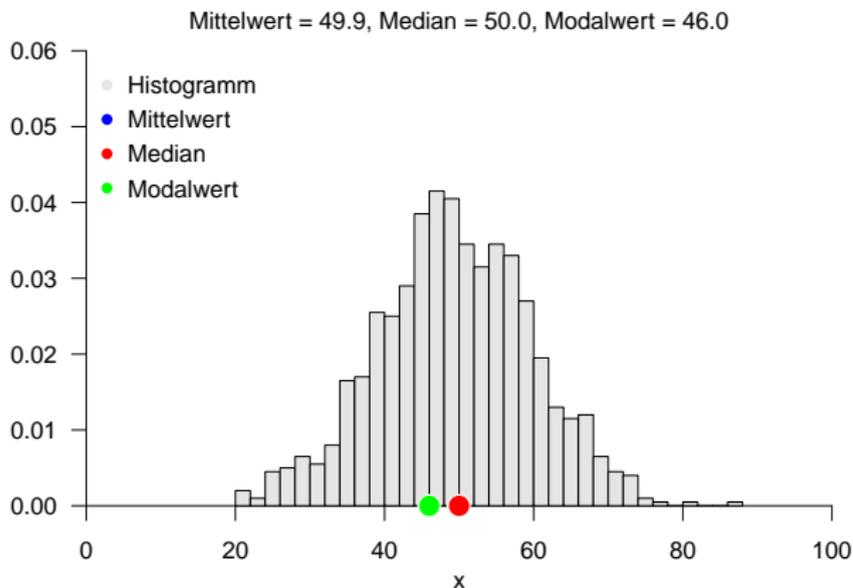
Median

Modalwert

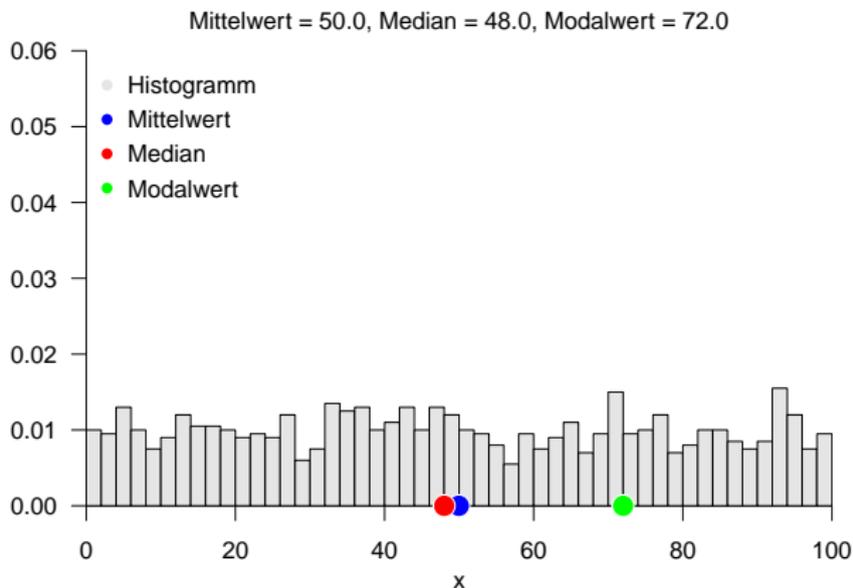
Visuelle Intuitionen

Übungen und Selbstkontrollfragen

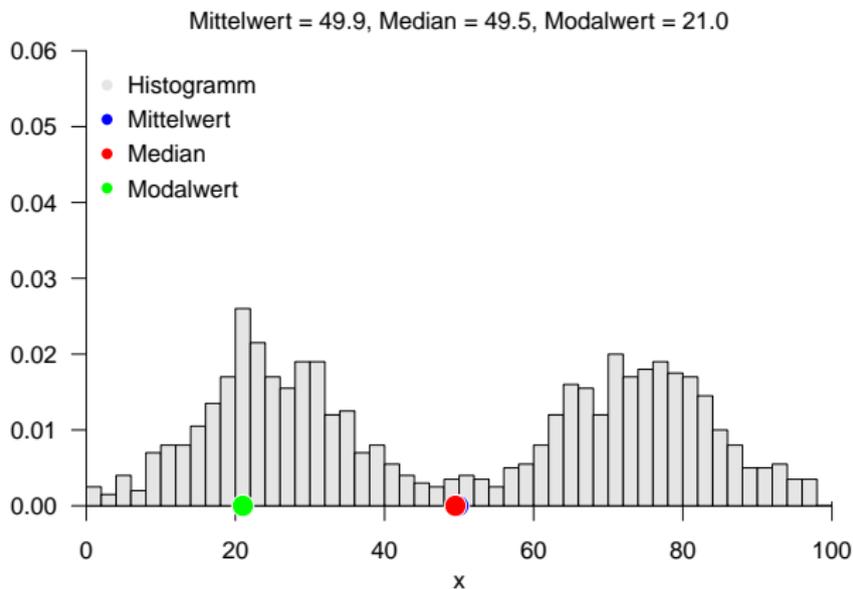
Visuelle Intuition zu Maßen zentraler Tendenz bei Normalverteilung



Visuelle Intuition zu Maßen zentraler Tendenz bei Gleichverteilung

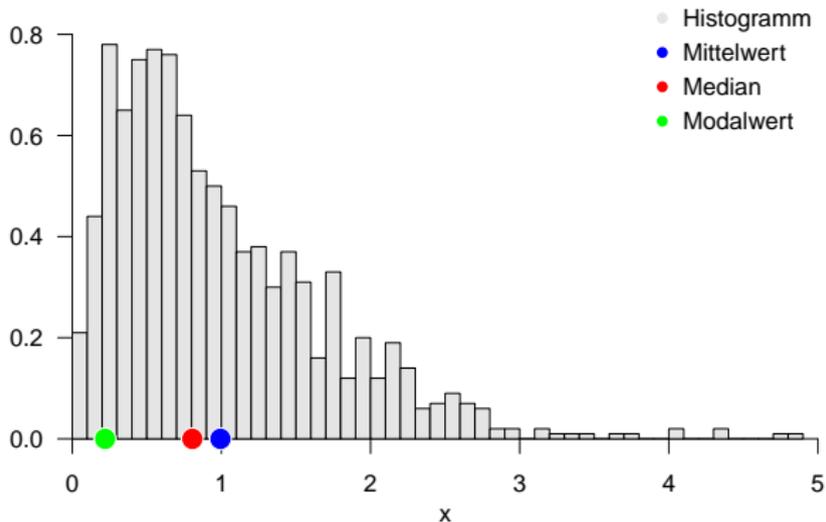


Visuelle Intuition zu Maßen zentraler Tendenz bei bimodalen Verteilungen



Visuelle Intuition zu Maßen zentraler Tendenz bei nicht-symmetrischen Verteilungen

Mittelwert = 1.0, Median = 0.8, Modalwert = 0.2



Mittelwert

Median

Modalwert

Visuelle Intuitionen

Übungen und Selbstkontrollfragen

Übungen und Selbstkontrollfragen

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.