

# Lösungen zu den Aufgaben

## 1. Aufgabe

Laden Sie den Datensatz `affairs`:

```
affairs_path <- "https://vincentarelbundock.github.io/Rdatasets/csv/AER/Affairs.csv"
affairs <- read_csv(affairs_path)
```

Lesen Sie das *Data Dictionary* [hier](#).

Wir definieren als “Halodrie” eine Person mit mindestens einer Affäre (laut Datensatz).

Bearbeiten Sie folgende Aufgaben:

1. Filtern Sie mal nach Halodries!
2. Sortieren Sie (absteigend) nach Anzahl der Affären!
3. Wählen Sie die Spalten zu Anzahl der Affären, ob es Kinder in der Ehe gibt und die Zufriedenheit mit der Ehe. Dann sortieren Sie dann nach Anzahl der Kinder und *danach* nach der Anzahl der Affären.
4. Berechnen Sie die mittlere Anzahl der Affären!
5. Berechnen Sie die mittlere Anzahl der Affären pro Geschlecht und aufgeteilt auf Partnerschaften mit bzw. ohne Kinder.
6. Geben Sie für jede Person die höhere der zwei Zahlen von Religiosität und Ehezufriedenheit aus!
7. Berechnen Sie jeweils das Heiratsalter!

## Lösung

Ad 1.

```
affairs %>%
  filter(affairs > 0) %>%
  head(10)
```

...1	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
6	3	male	27	1.500	no	3	18	4	4
12	3	female	27	4.000	yes	3	17	1	5
43	7	male	37	15.000	yes	5	18	6	2
53	12	female	32	10.000	yes	3	17	5	2
67	1	male	22	0.125	no	4	16	5	5
79	1	female	22	1.500	yes	2	14	1	5
122	12	male	37	15.000	yes	4	14	5	2
126	7	female	22	1.500	no	2	14	3	4
133	2	male	37	15.000	yes	2	18	6	4
138	3	female	32	15.000	yes	4	12	3	2

Hinweis: `head(10)` begrenzt die Ausgabe auf 10 Zeilen, einfach um den Bildschirm nicht vollzumüllen.

Ad 2.

```
affairs %>%
  arrange(-affairs) %>%
  head(10)
```

...1	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
------	---------	--------	-----	--------------	----------	---------------	-----------	------------	--------

...	1	affairs	gender	age	years	married	children	religiousness	education	occupation	rating
53	12	female	32	10.0	yes	3	17	5	2		
122	12	male	37	15.0	yes	4	14	5	2		
174	12	female	42	15.0	yes	5	9	4	1		
176	12	male	37	10.0	yes	2	20	6	2		
181	12	female	32	15.0	yes	3	14	1	2		
252	12	male	27	1.5	yes	3	17	5	4		
253	12	female	27	7.0	yes	4	14	6	2		
392	12	female	32	10.0	yes	2	16	5	5		
513	12	female	22	4.0	no	3	12	3	4		
516	12	male	27	7.0	yes	1	18	6	2		

Ad 3.

```
affairs %>%
  select(affairs, rating, children) %>%
  arrange(children, affairs) %>%
  head(10)
```

**affairs rating children**

0	4 no
0	4 no
0	3 no
0	5 no
0	3 no
0	5 no
0	4 no
0	4 no
0	4 no
0	5 no

Ad 4.

```
affairs %>%
  summarise(affairs_mean = mean(affairs)) %>%
  head(10)
```

**affairs\_mean**

1.455907

Ad 5.

```
affairs %>%
  group_by(gender, children) %>%
  summarise(affairs_mean = mean(affairs)) %>%
  head(10)

## # A tibble: 4 × 3
## # Groups:   gender [2]
##   gender children affairs_mean
##   <chr>   <chr>         <dbl>
## 1 female no           0.838
## 2 female yes          1.69
## 3 male   no           1.01
## 4 male   yes          1.66
```

Ad 6.

```
affairs %>%
  group_by(...1) %>%
  summarise(max(c(religiousness, rating))) %>%
  head(10)
```

**...1 max(c(religiousness, rating))**

4	4
5	4
6	4
11	4
12	5
16	5
23	3
29	5
43	5
44	3

Ad 7.

```
affairs %>%
  mutate(heiratsalter = age - yearsmarried) %>%
  head(10)
```

**affairs gender age yearsmarried children religiousness education occupation rating heiratsalter**

4	0 male	37	10.00	no	3	18	7	4	27.00
5	0 female	27	4.00	no	4	14	6	4	23.00
11	0 female	32	15.00	yes	1	12	1	4	17.00
16	0 male	57	15.00	yes	5	18	6	5	42.00
23	0 male	22	0.75	no	2	17	6	3	21.25
29	0 female	32	1.50	no	2	17	5	5	30.50
44	0 female	22	0.75	no	2	12	1	3	21.25
45	0 male	57	15.00	yes	2	14	4	4	42.00
47	0 female	32	15.00	yes	4	16	1	2	17.00
49	0 male	22	1.50	no	4	14	4	5	20.50

## 2. Aufgabe

Importieren Sie den folgenden Datensatz in R:

```
mtcars <- read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv")
```

Übersetzen Sie dann die folgende R-Sequenz ins Deutsche:

```
mtcars %>%
  drop_na() %>%
  select(mpg, hp, cyl) %>%
  filter(hp > 100, cyl >= 6) %>%
  group_by(cyl) %>%
  summarise(mpg_mean = mean(mpg))
```

**cyl mpg\_mean**

6	19.74286
8	15.10000

## Lösung

Hey R:

1. Nimm den Datensatz `mtcars` UND DANN
2. hau alle Zeilen raus, in denen es fehlende Werte gibt UND DANN
3. wähle (selektiere) die folgenden Spalten: Spritverbrauch, PS, Zylinder UND DANN
4. filter Autos mit mehr als 100 PS und mit mindestens 6 Zylindern UND DANN
5. gruppieren nach der Zahl der Zylinder UND DANN
6. fasse den Verbrauch zum Mittelwert zusammen.

### 3. Aufgabe

Welcher Kennwert ist robust (gegenüber Extremwerten)?

- a. Standardabweichung
- b. Mittelwert
- c. Korrelation
- d. Median
- e. Maximalwert

## Lösung

Der Median ist robust. Mittelwertsbasierte Kennzahlen hingegen nicht.

- a. Falsch
- b. Falsch
- c. Falsch
- d. Wahr
- e. Falsch

### 4. Aufgabe

Welcher Kennwert ist robust (gegenüber Extremwerten)?

- a. Schiefe
- b. Regressionsgewicht
- c. Summe
- d. Korrelation
- e. Interquartilsabstand

## Lösung

Der Interquartilsabstand ist robust. Mittelwertsbasierte Kennzahlen hingegen nicht.

- a. Falsch
- b. Falsch
- c. Falsch
- d. Falsch
- e. Wahr

## 5. Aufgabe

Berechnen Sie den *Median* der folgenden Datenreihe!

*Hinweis:* Runden Sie auf zwei Dezimalstellen. Beachten Sie, dass das Dezimalzeichen (Punkt oder Komma) je nach Ihrem System unterschiedlich sein kann.

```
## [1] 2.77 0.01 5.11 0.14 0.65
```

### Lösung

Die Antwort lautet 0.65.

## 6. Aufgabe

Berechnen Sie den *Mittelwert* der folgenden Datenreihe!

*Hinweis:* Runden Sie auf zwei Dezimalstellen. Beachten Sie, dass das Dezimalzeichen (Punkt oder Komma) je nach Ihrem System unterschiedlich sein kann.

```
## [1] 7.10 2.46 3.90 0.91 9.62
```

### Lösung

Die Antwort lautet 4.8.

In R kann man den Mittelwert z.B. so berechnen:

```
mean(x)
## [1] 4.798
```

## 7. Aufgabe

Berechnen Sie den Mittelwert folgender Zahlenreihe; ignorieren sie etwaige fehlende Werte. Runden Sie auf zwei Dezimalstellen.

```
## [1] -1.02 -0.08 -0.23 -0.82 0.77 NA
```

### Lösung

Der Mittelwert liegt bei -0.28.

Die Antwort lautet -0.28.

In R kann man den Mittelwert z.B. so berechnen:

```
mean(x, na.rm = TRUE)
```

```
## [1] 5.016
```

Das Argument `na.rm = TRUE` sorgt dafür, dass R *auch bei Vorhandensein fehlender Werte* ein Ergebnis ausgibt. Ohne dieses Argument würde R ein sprödes `NA` zurückgeben, falls fehlende Werte vorliegen. Dieses Verhalten von R ist recht defensiv, getreu dem Motto: Wenn es ein Problem gibt, sollte man so früh wie möglich darüber deutlich informiert werden (und nicht erst, wenn die Marsrakete gestartet ist...).

## 8. Aufgabe

Betrachten Sie die Histogramme.

Wählen Sie das Histogramm, welches am deutlichsten die Eigenschaft “symmetrisch” aufweist!

Diagramm A

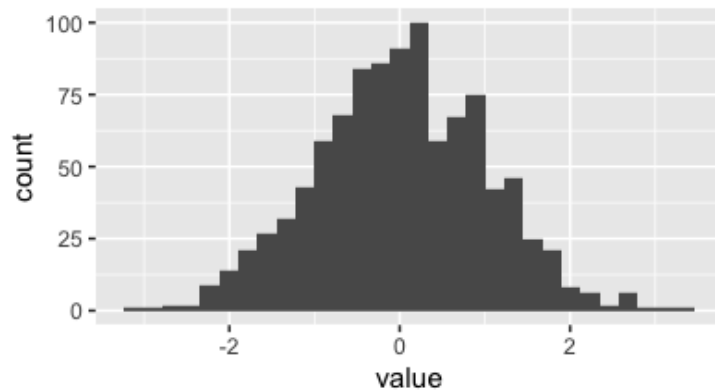


Diagramm B

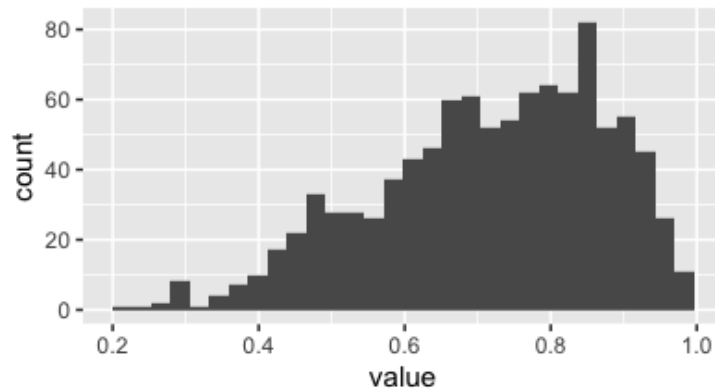
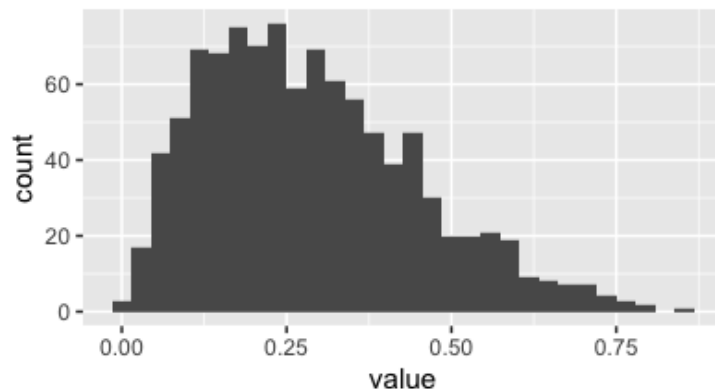
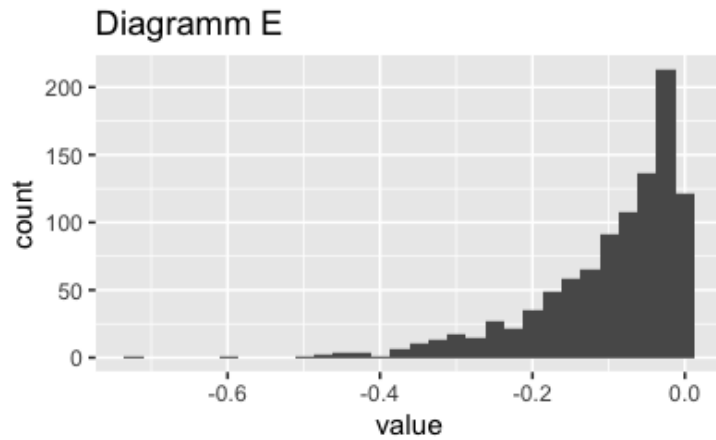
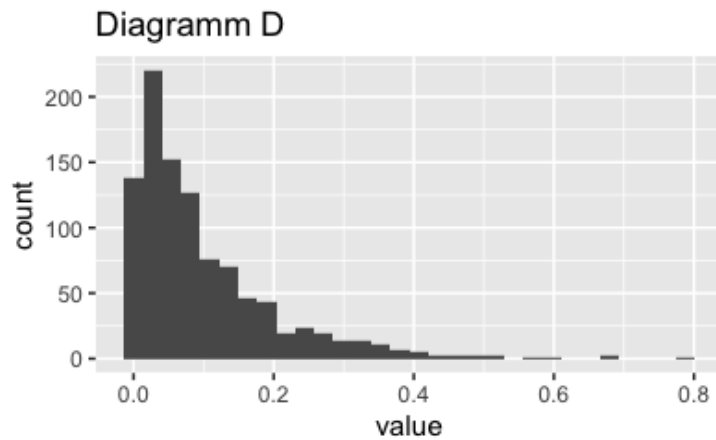


Diagramm C





- a. A
- b. B
- c. C
- d. D
- e. E

### Lösung

Das Histogramm A zeigt die Eigenschaft *symmetrisch* am deutlichsten.

- a. Wahr
- b. Falsch
- c. Falsch
- d. Falsch
- e. Falsch

### 9. Aufgabe

Betrachten Sie die Histogramme.

Wählen Sie das Histogramm, welches am deutlichsten folgende Eigenschaft aufweist:

$$MW < Md$$

*Hinweis: MW steht für Mittelwert und Md steht für Median.*

Diagramm A

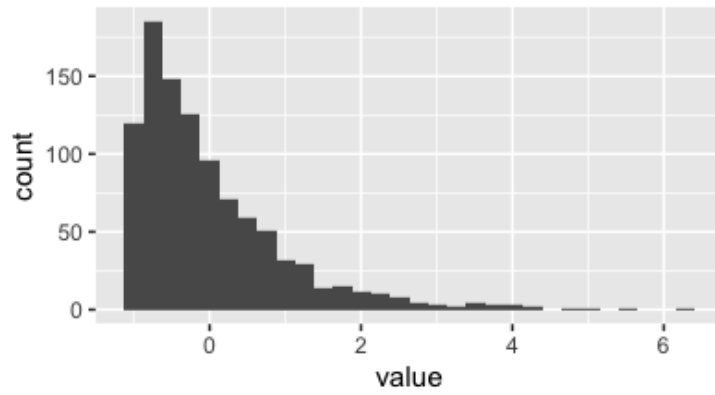


Diagramm B

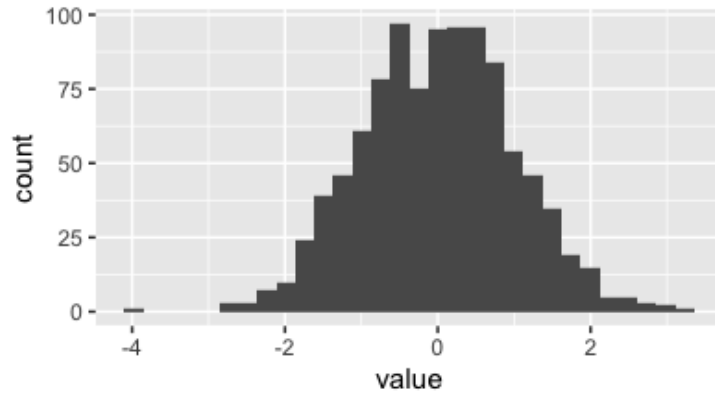


Diagramm C

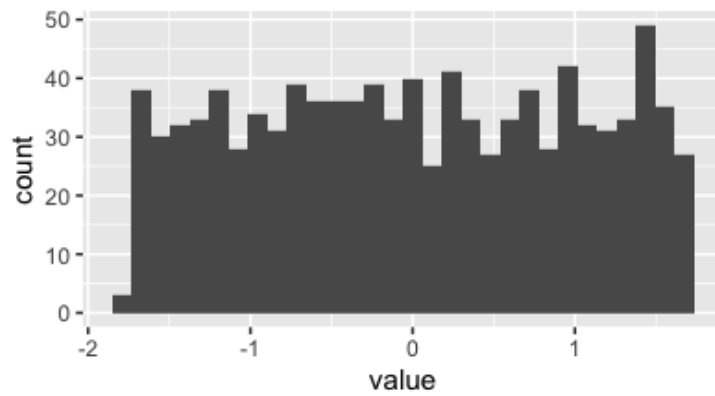
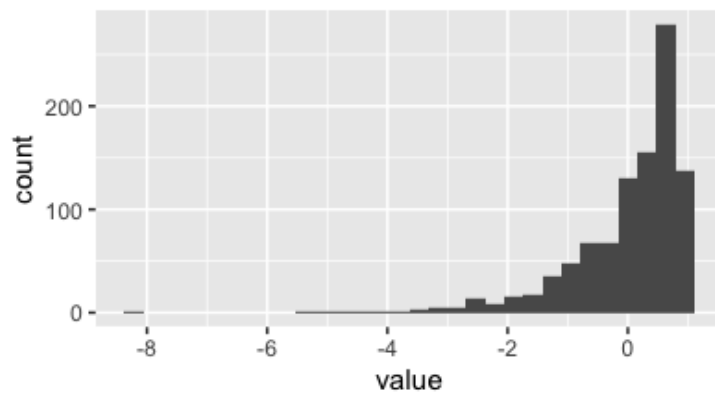


Diagramm D



- a. A
- b. B



- c. C
- d. D

### Lösung

Das Histogramm <sub>D</sub> zeigt die Eigenschaft  $MW < Md$  am deutlichsten.

- a. Falsch
- b. Falsch
- c. Falsch
- d. Wahr

### 10. Aufgabe

Welche Form der Verteilung liegt wohl (am ehesten) für die Variable Geburten je Tag im Monat vor?

- a. linksschief
- b. normalverteilt
- c. rechtsschief
- d. gleichverteilt

### Lösung

Die Variable Geburten je Tag im Monat lässt sich am ehesten beschreiben mit der Verteilungsform gleichverteilt.

- a. Falsch
- b. Falsch
- c. Falsch
- d. Richtig

### 11. Aufgabe

Sei  $X \sim \mathcal{N}(42, 7)$  und  $x_1 = 28$ .

Berechnen Sie den z-Wert für  $x_1$ !

Hinweis:

- Runden Sie ggf. auf die nächste ganze Zahl.

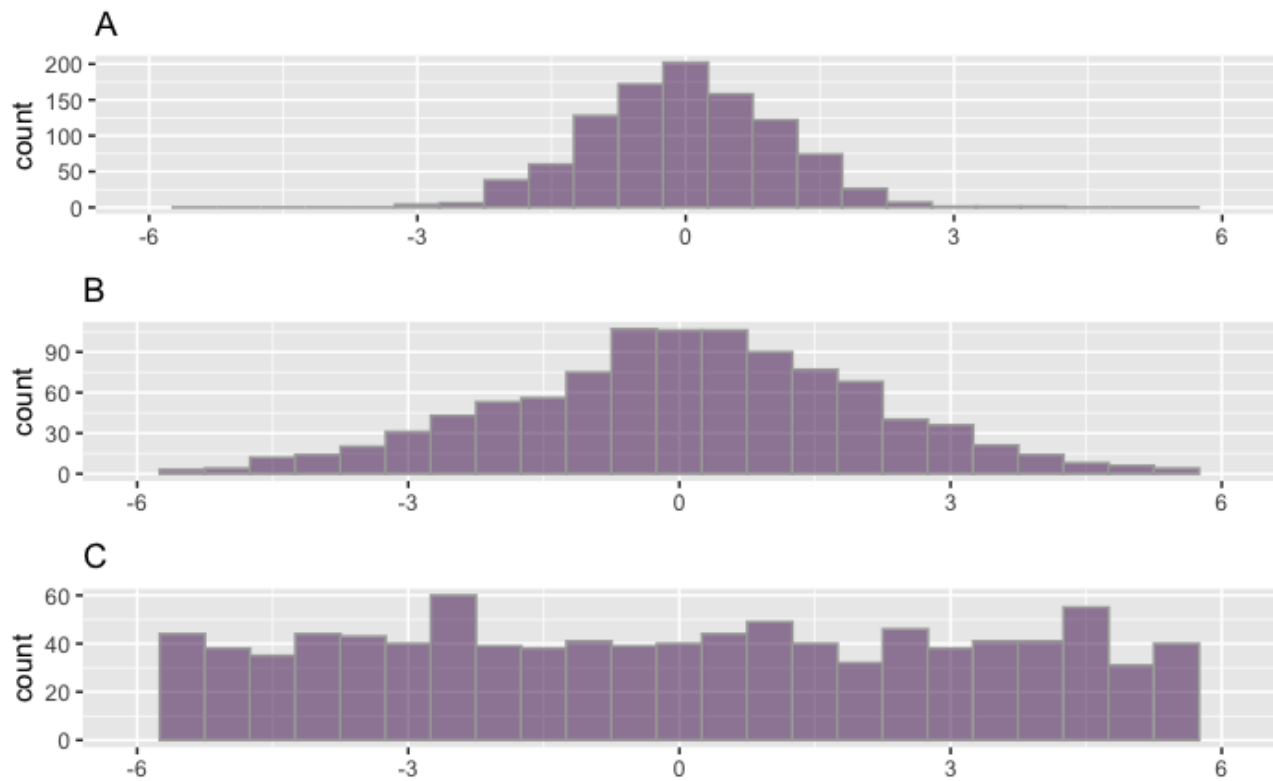
### Lösung

$$x1\_z = (x1 - x\_mw) / x\_sd$$

-2

### 12. Aufgabe

Welches der folgenden Diagramm hat die größte Streuung, gemessen in Standardabweichung?



- a. A
- b. B
- c. C
- d. alle gleich
- e. keine Antwort möglich

### Lösung

Die SD ist am größten in Diagramm C

- a. Falsch. Dieses Diagramm hat die kleinste Streuung
- b. Falsch
- c. Wahr
- d. Falsch. Die Streuungen sind unterschiedlich.
- e. Falsch

### 13. Aufgabe

Wählen Sie das Diagramm, in dem der vertikale gestrichelte Linie am genauesten die Position des Medians ( $Md$ ) widerspiegelt.

Diagramm A

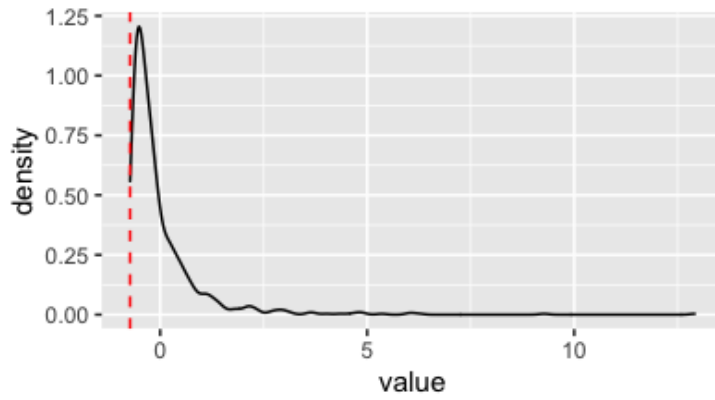


Diagramm B

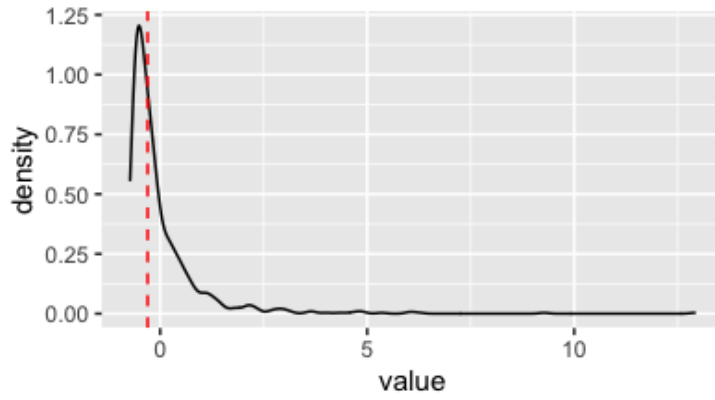


Diagramm C

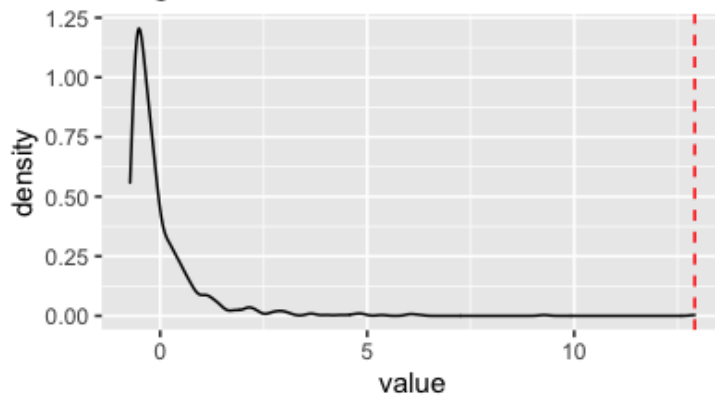
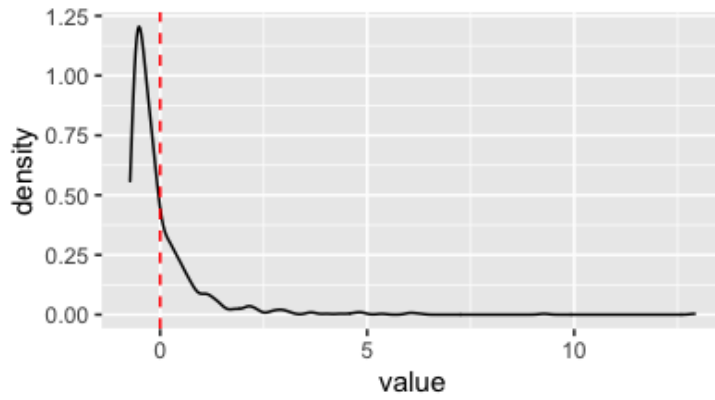


Diagramm D



- a. A
- b. B

- c. C
- d. D

## Lösung

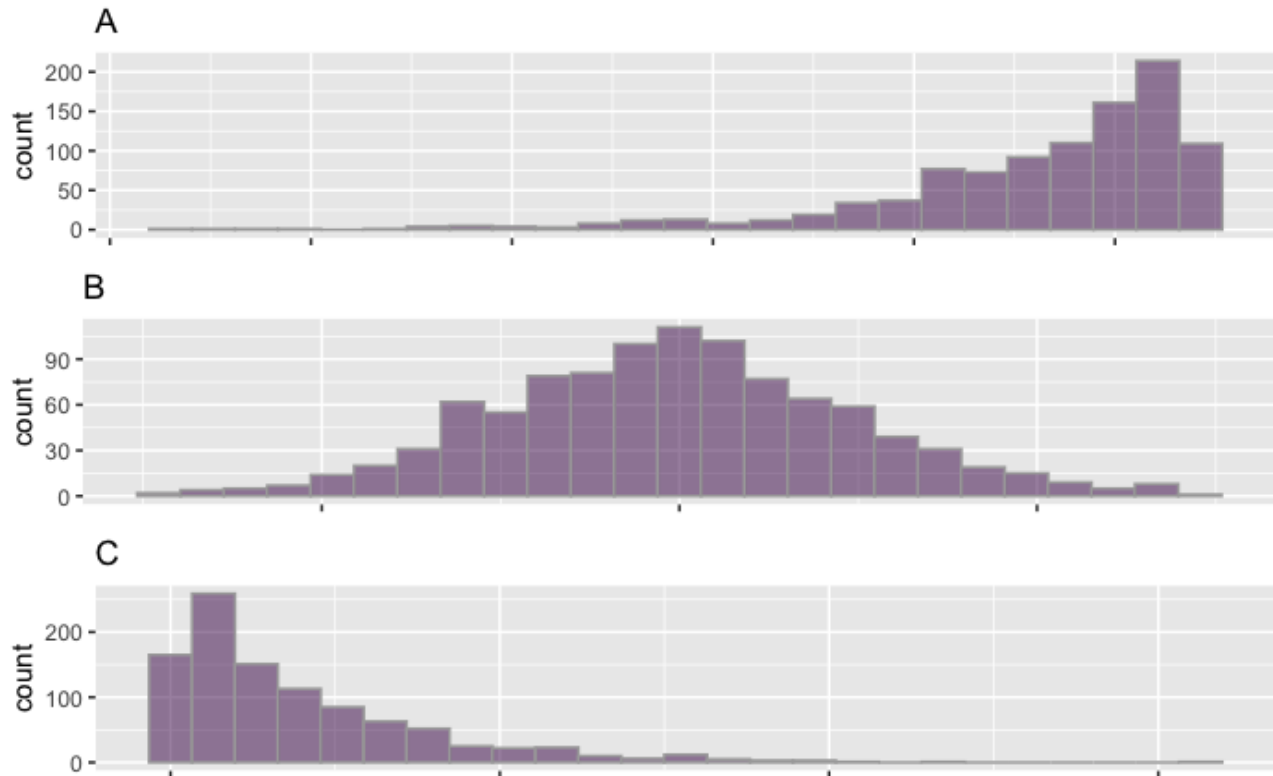
Das Diagramm **B** zeigt den Median am genauesten.

- a. Falsch
- b. Wahr
- c. Falsch
- d. Falsch

## 14. Aufgabe

Für welche Abbildung gilt, dass der Median kleiner ist als der (zugehörige) arithmetischer Mittelwert?

Anders gesagt, gesucht ist  $md < \bar{x}$



- a. A
- b. B
- c. C
- d. keine Antwort möglich

## Lösung

Der Mittelwert ist i. d. R. in Richtung “des langen Endes” einer Verteilung verschoben, daher **C**.

Faustregel:

**linksschief**

Mittelwert < Median

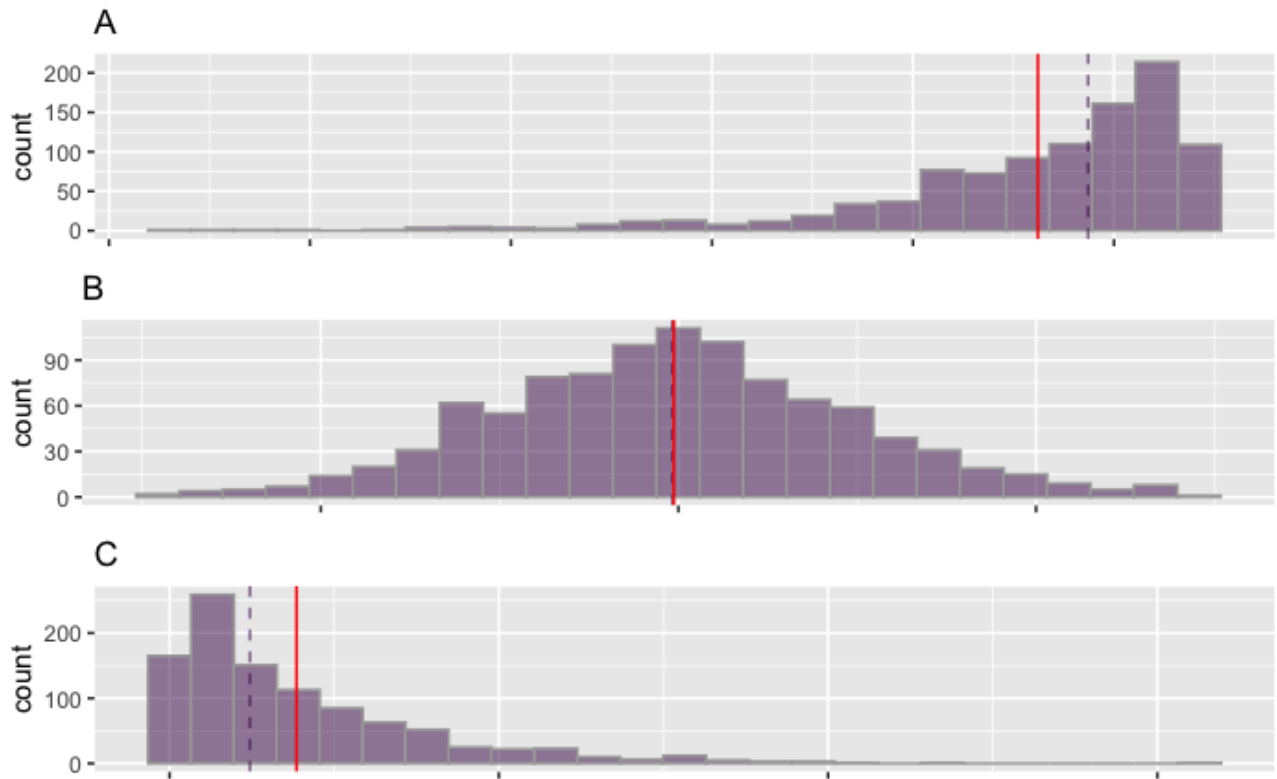
**symmetrisch**

Mittelwert  $\approx$  Median

**rechtsschief**

Mittelwert > Median

Bei (sehr) schiefen Daten beschreibt der Median (blau, gestrichelt) den Schwerpunkt der Beobachtungen besser als der arithmetische Mittelwert (rot).



- a. Falsch
- b. Falsch
- c. Wahr
- d. Falsch