Deskriptive Statistik

Vitaly Belik

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

Oct 13, 2019

Was haben wir letztes mal gelernt?

http://menti.com

Wiederholung

In der letzen Vorlesung haben wir über die Aufgaben der (Bio)statistik gersprochen.

- Deskriptive (beschreibende) und Schliessende (induktive) Statistik
- Grundgesamtheit und Stichprobe
- Skalenniveaus von Merkmalen

Deskriptive (beschreibende) Statistik

Methoden zur:

- *Auswertung*
- übersichtlichen *Darstellung*
- *Zusammenfassung* von Daten.

Deskriptive (beschreibende) Statistik(1)

Tabellen

Graphiken

Charakterischtische Maßzahlen

Tabellen

Table 1: Titanic dataset

Index	survived	pclass	sex	age	deck	fare	alone
0	0	3	male	22		7.2500	False
1	1	1	female	38	C	71.2833	False
2	1	3	female	26		7.9250	True
3	1	1	female	35	C	53.1000	False
4	0	3	male	35		8.0500	True
5	0	3	male	NA		8.4583	True
6	0	1	male	54	E	51.8625	True
7	0	3	male	2		21.0750	False
8	1	3	female	27		11.1333	False
9	1	2	female	14		30.0708	False

Urliste

Die ungeordnete Form von Messungen (Beobachtungen) einer Untersuchung, die der Reihe nach zusammengestellt ist.

Table 2: Urliste

	1	2	3	4	5	6	7	8	9	10
age	22	38	26	35	35	NA	54	2	27	14

Tabellen(1)

Primäre Tafel oder geordnete Liste

Table 3: Geordnete Liste

	8	10	1	3	9	4	5	2	7
age	2	14	22	26	27	35	35	38	54

Häufigkeitstabelle

Table 4: Häufigkeitstabelle

age	freq
0.42	1
0.67	1
0.75	2
0.83	2
0.92	1
1	7
2	10
3	6
4	10
5	4
6	3
7	3
8	4
9	8
10	2
11	4
12	1
13	2
14	6
14.5	1
15	5
16	17
17	13
18	26
19	25
20	15
20.5	1
21	24
22	27
23	15

Stamm-Blatt-Diagramm (stem and leaf diagram)

```
The decimal point is 1 digit(s) to the right of the |

0 | 011111111111112222222223333334444444444
0 | 5555666777888999999999
1 | 001111233444444
1 | 555555666666666666666677777777777778888888888888888888888888999999+7
2 | 000000000000000111111111111111111111111122222222222222222222222223+33
2 | 555555555555555555555556666666666666666667777777777777777788888888+27
3 | 000000000000000000000000011111111111111111122222222222222222333333+14
3 | 555555555555555555566666666666666666666667777777888888888899999999
4 | 000000000000111111112222222222223333344444444
4 | 555555555556666677777777888888888999999
5 | 000000000111111122222344444444
5 | 5566666778888899
6 | 000011122223344
6 | 5556
7 | 001114
7 |
8 | 0
```

Man rundet die Ergebnisse auf Blatt-Genaugkeit

Graphiken

Balkendiagramm

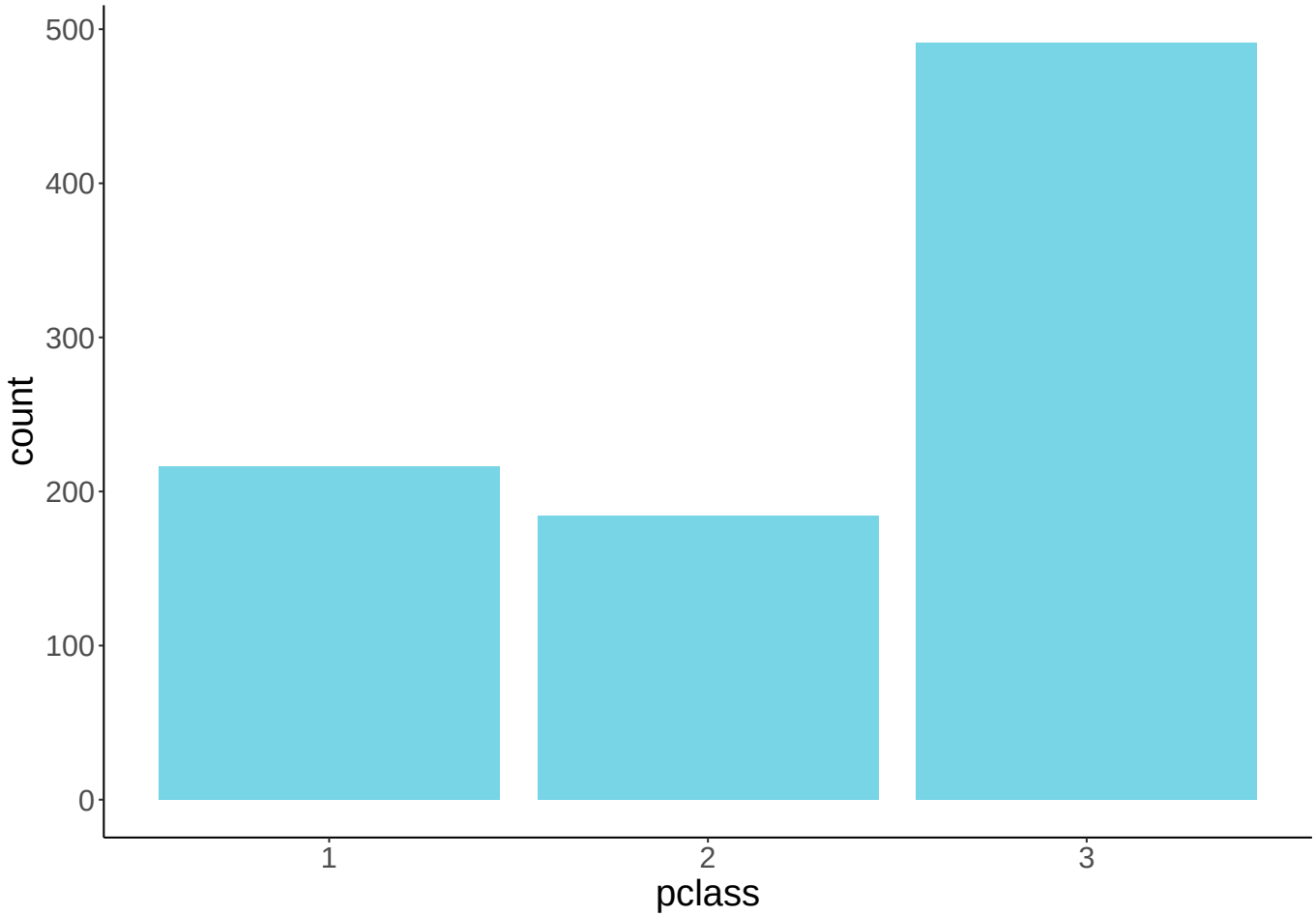


Figure 1: Balkendiagramm

Balkendiagramm (relative Häufigkeit)

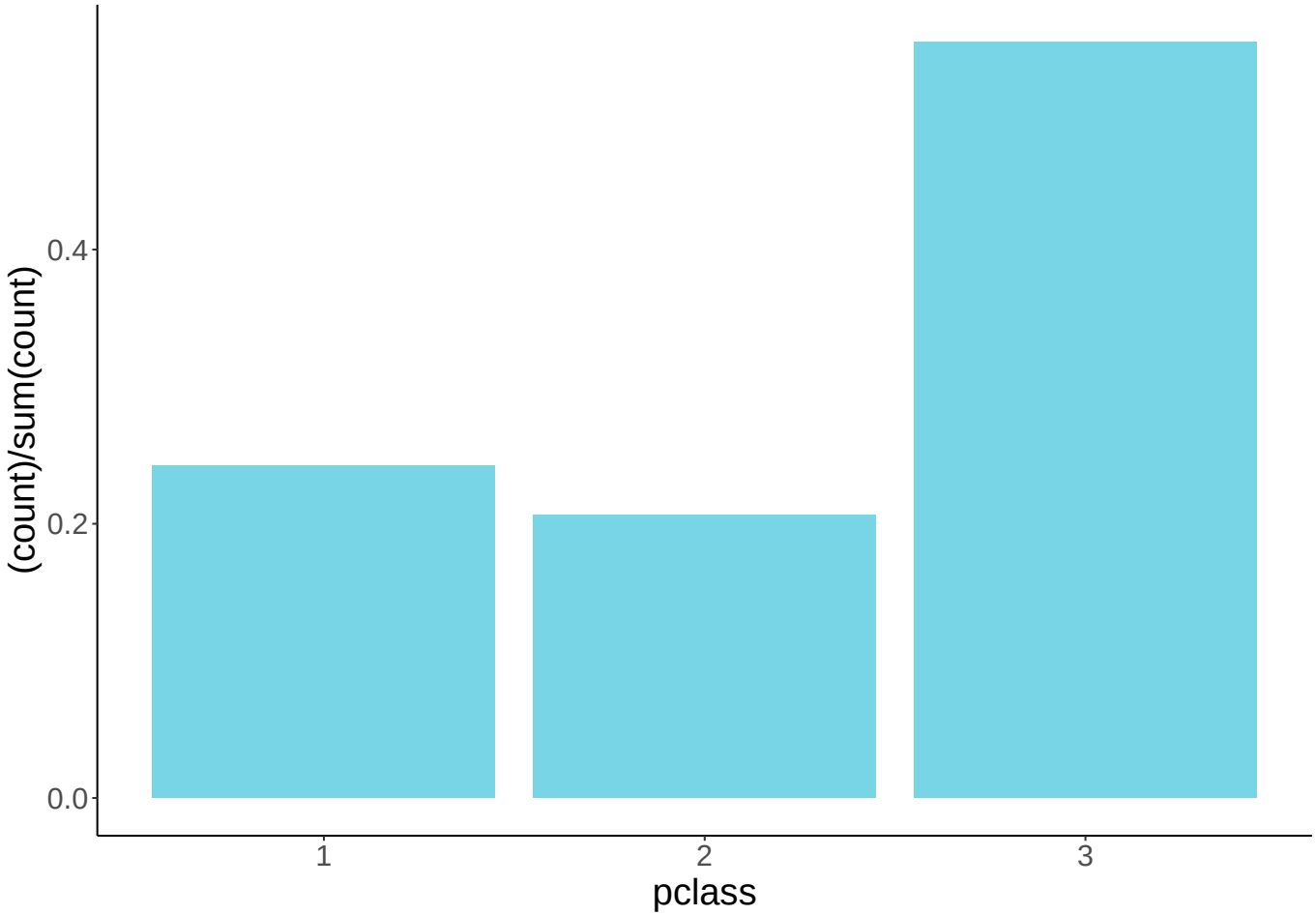


Figure 2: Balkendiagramm (relative Häufigkeit)

Komponenten-Balkendiagramm

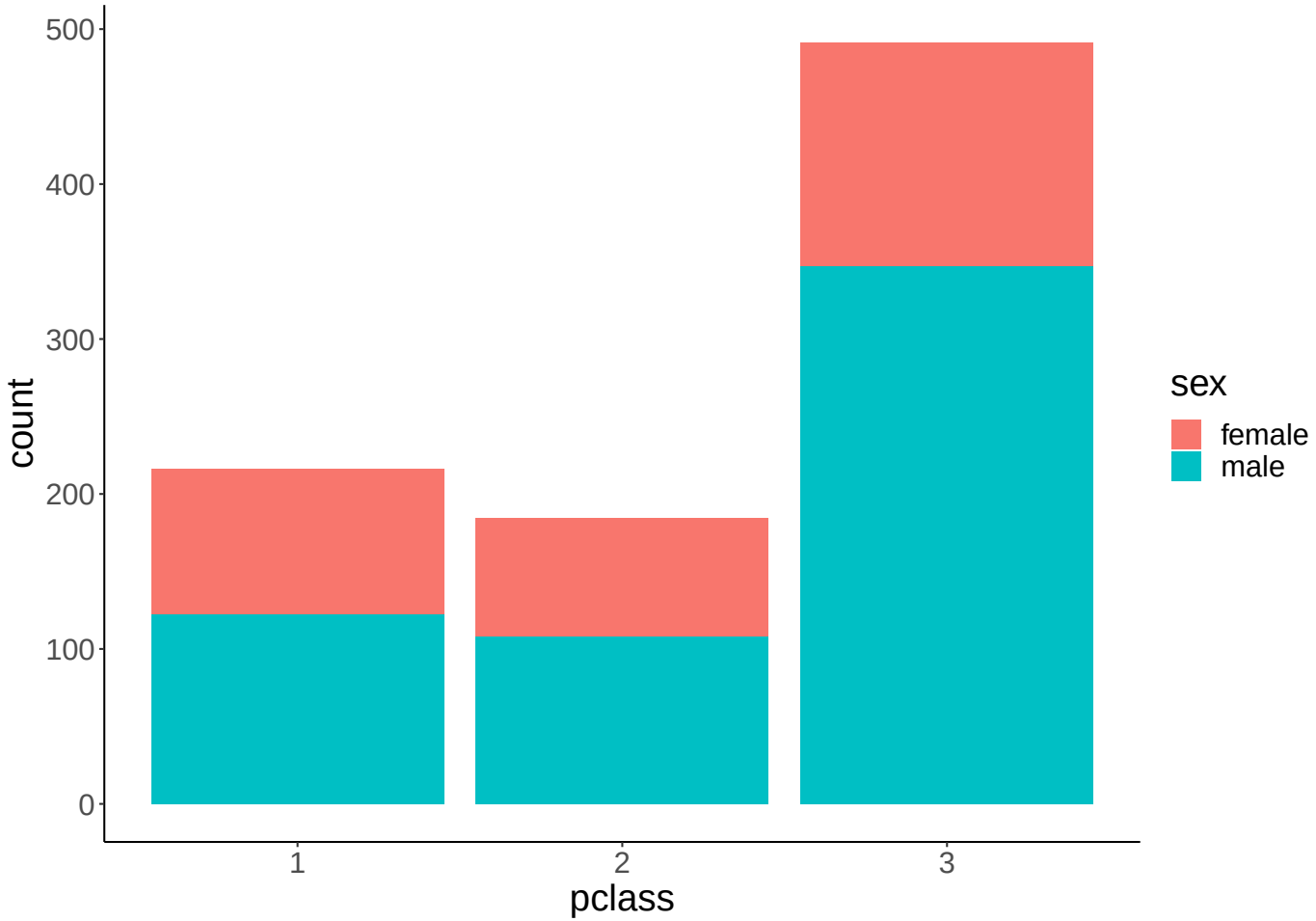


Figure 3: Komponenten-Balkendiagramm

Komponenten-Balkendiagramm (relative Häufigkeit)

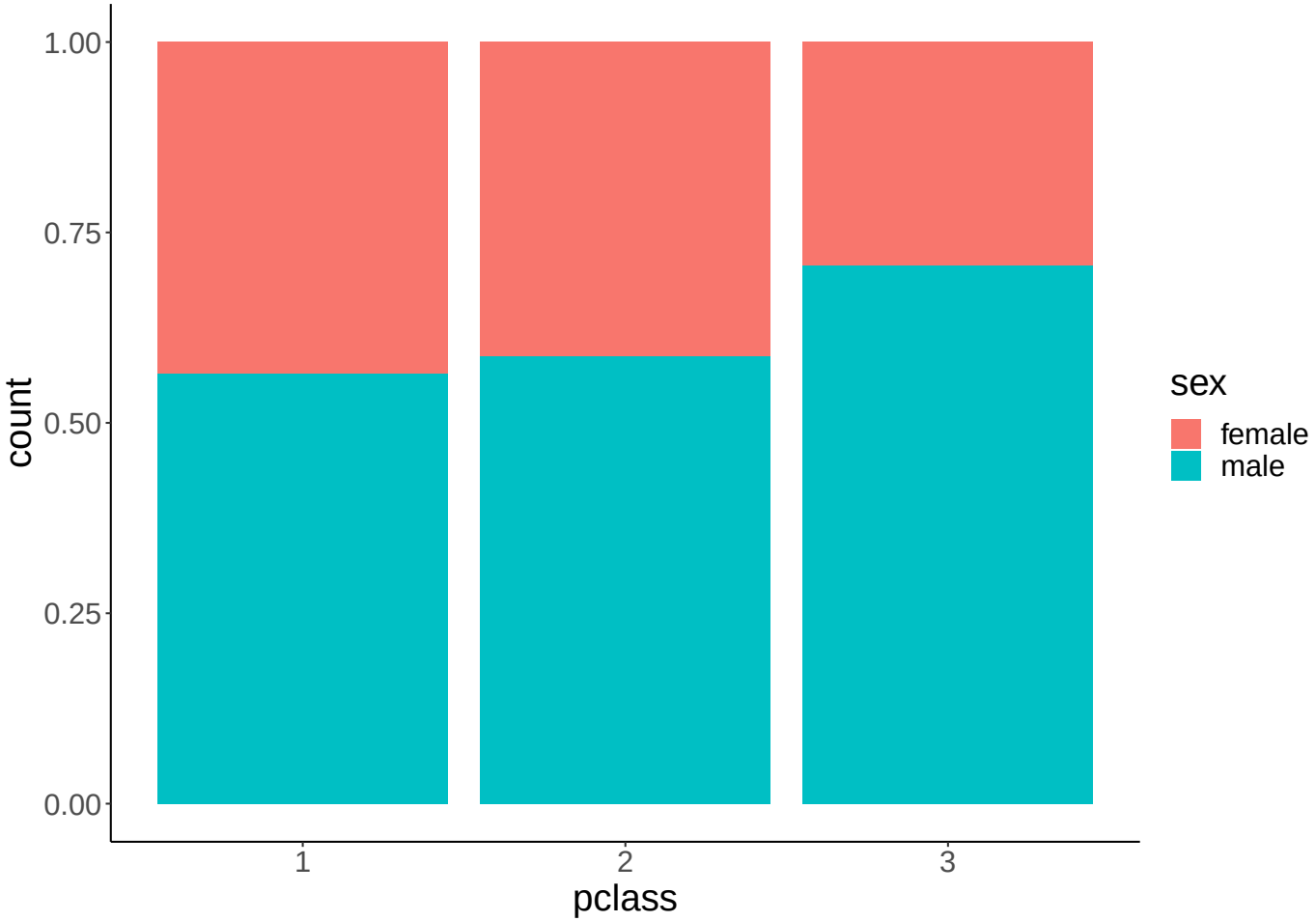


Figure 4: Komponenten-Balkendiagramm (relative Häufigkeit)

Komponenten-Balkendiagramm (relative Häufigkeit) (1)

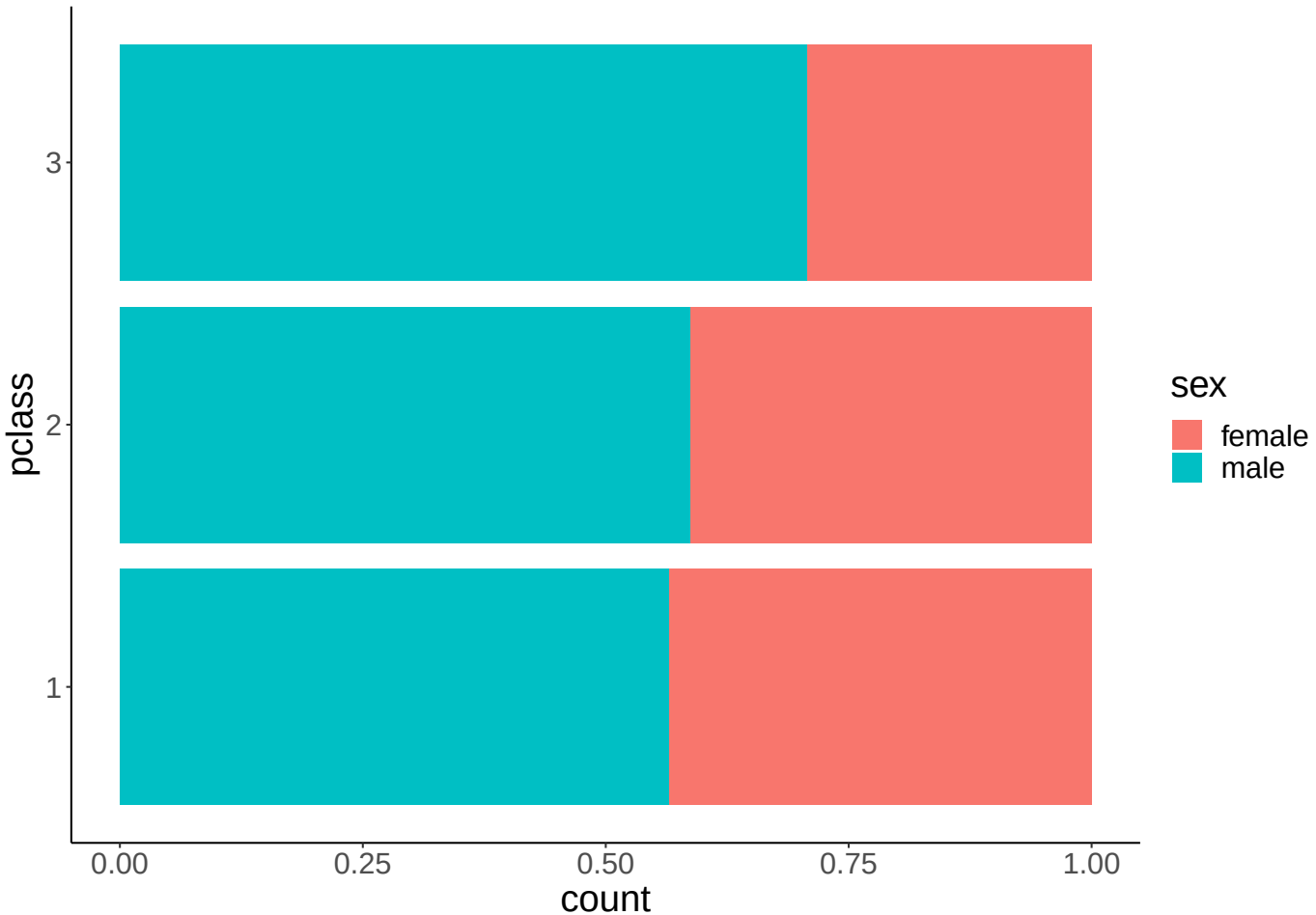


Figure 5: Komponenten-Balkendiagramm (relative Häufigkeit)

Daten auf der Zahlengerade

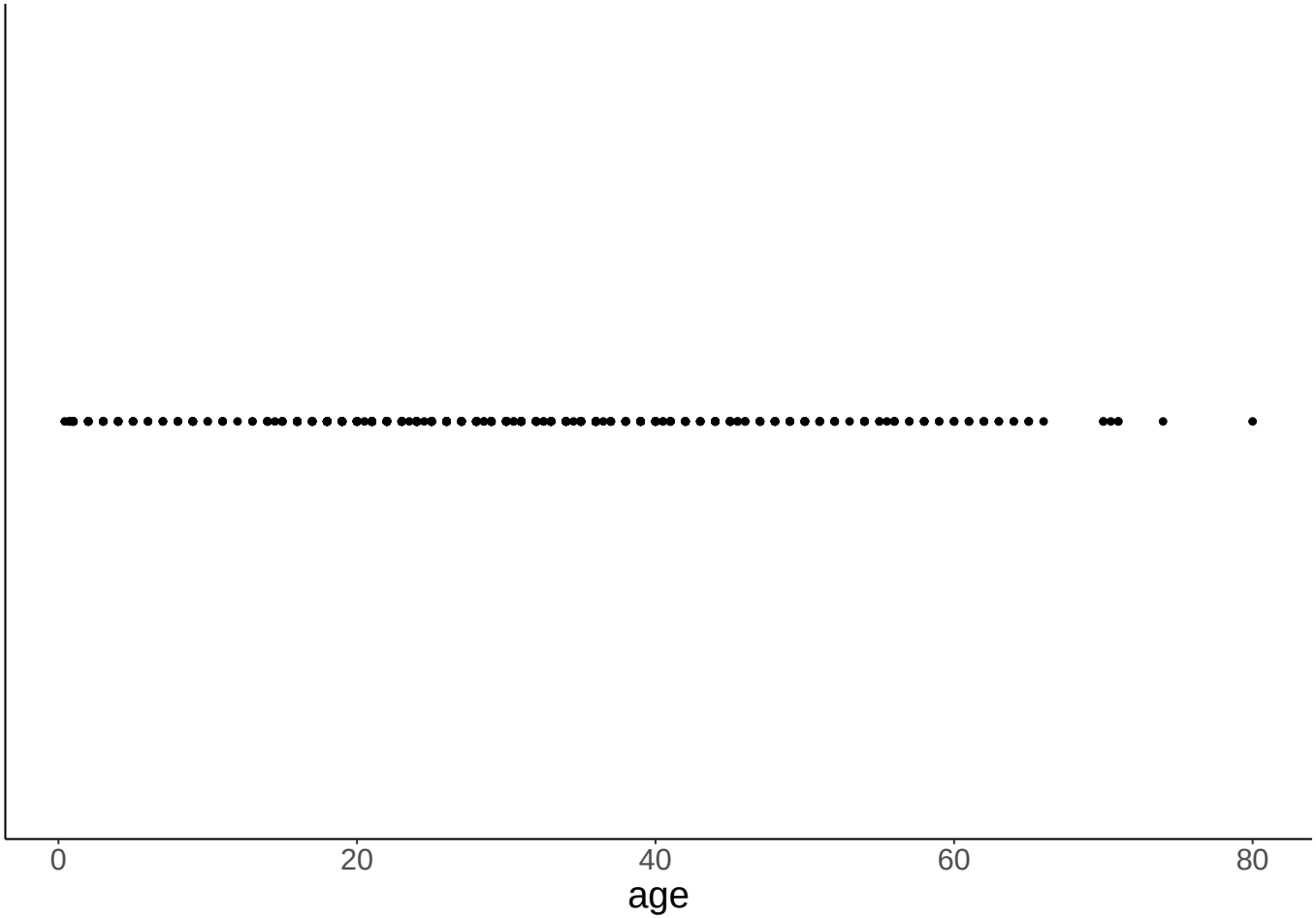


Figure 6: Daten auf der Zahlengerade

Histogramm

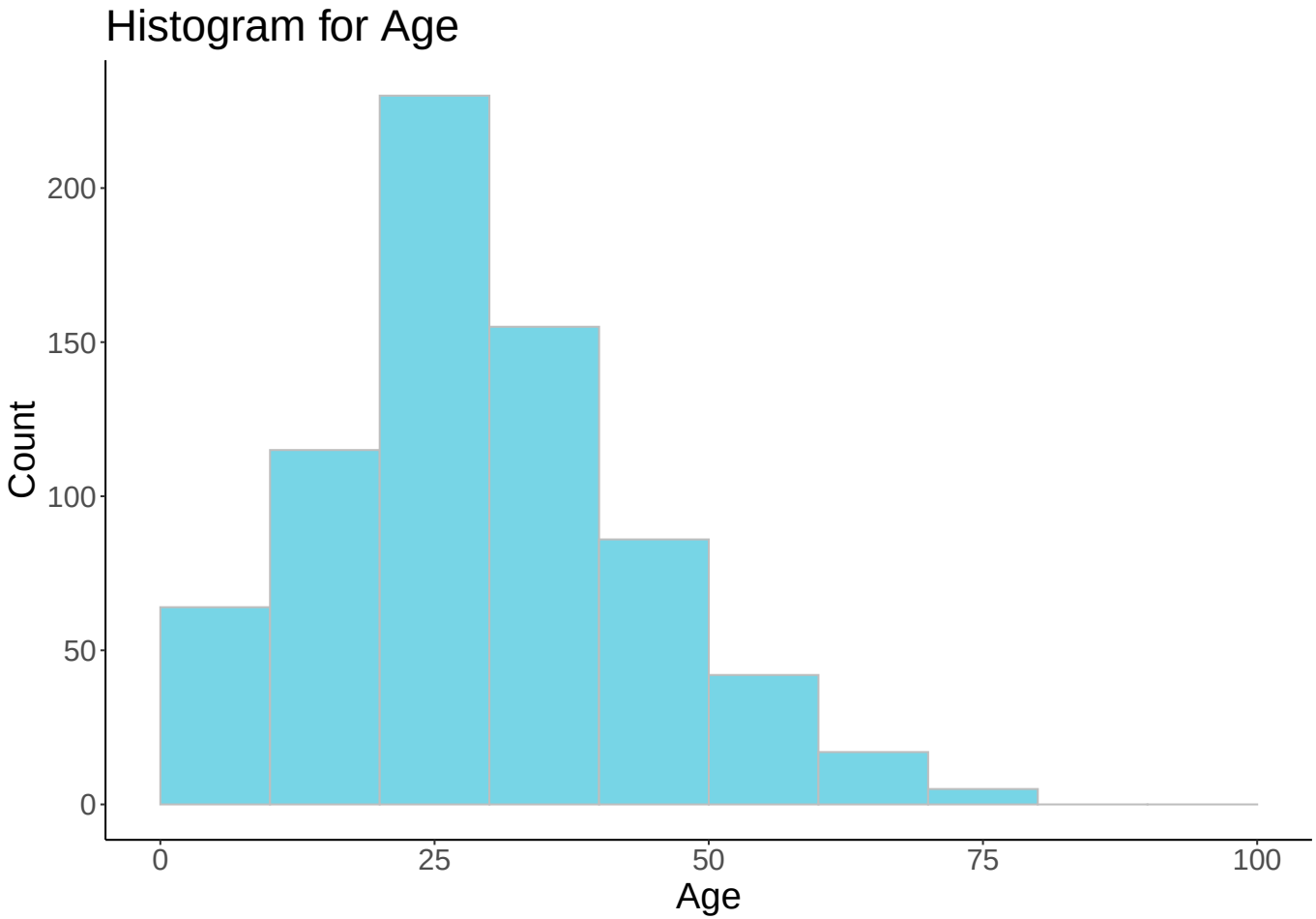


Figure 7: Histogramm

Histogramm (relative Häufigkeit)

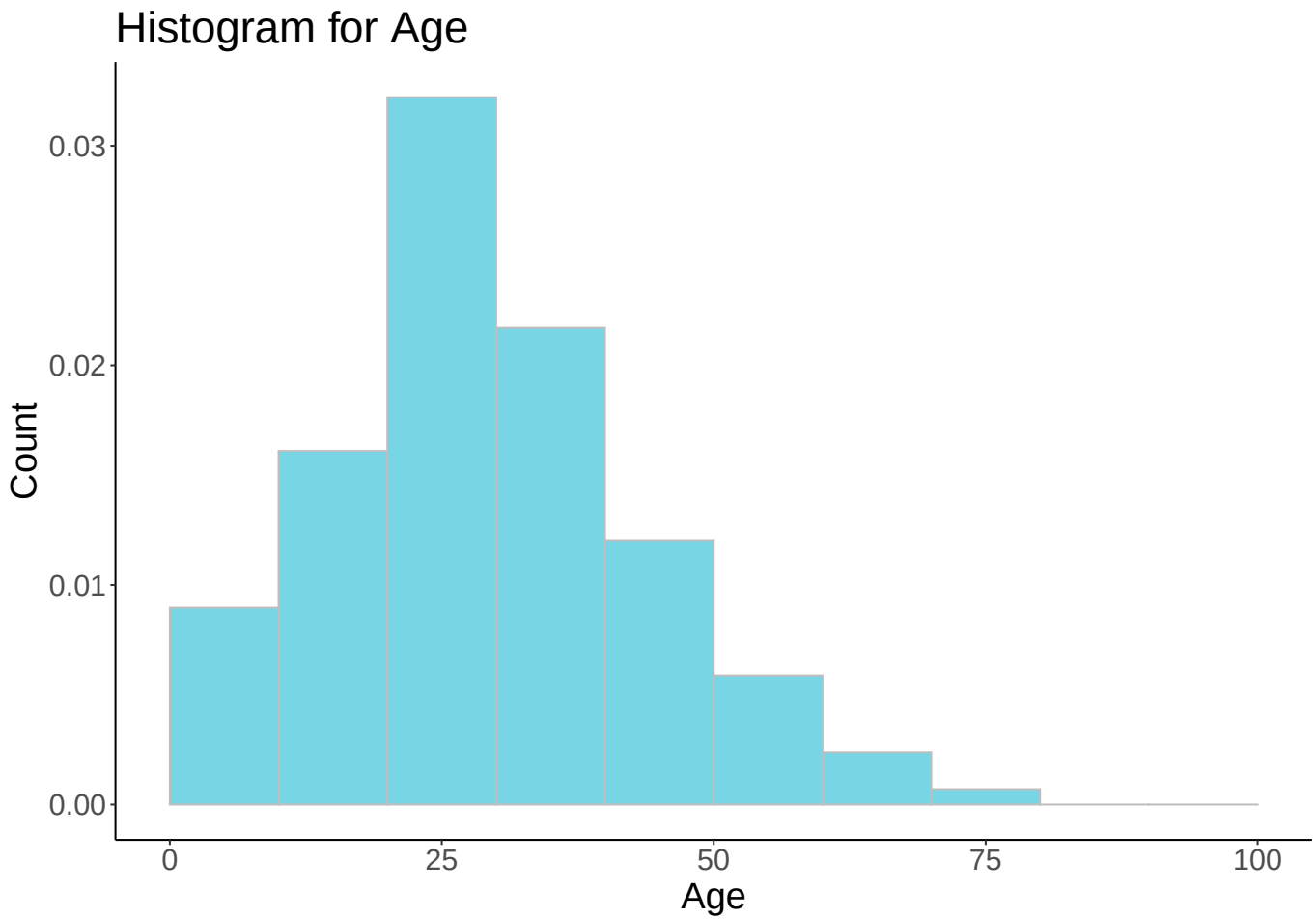


Figure 8: Histogramm (relative Häufigkeit)

Histogramm (relative Häufigkeit) (2)

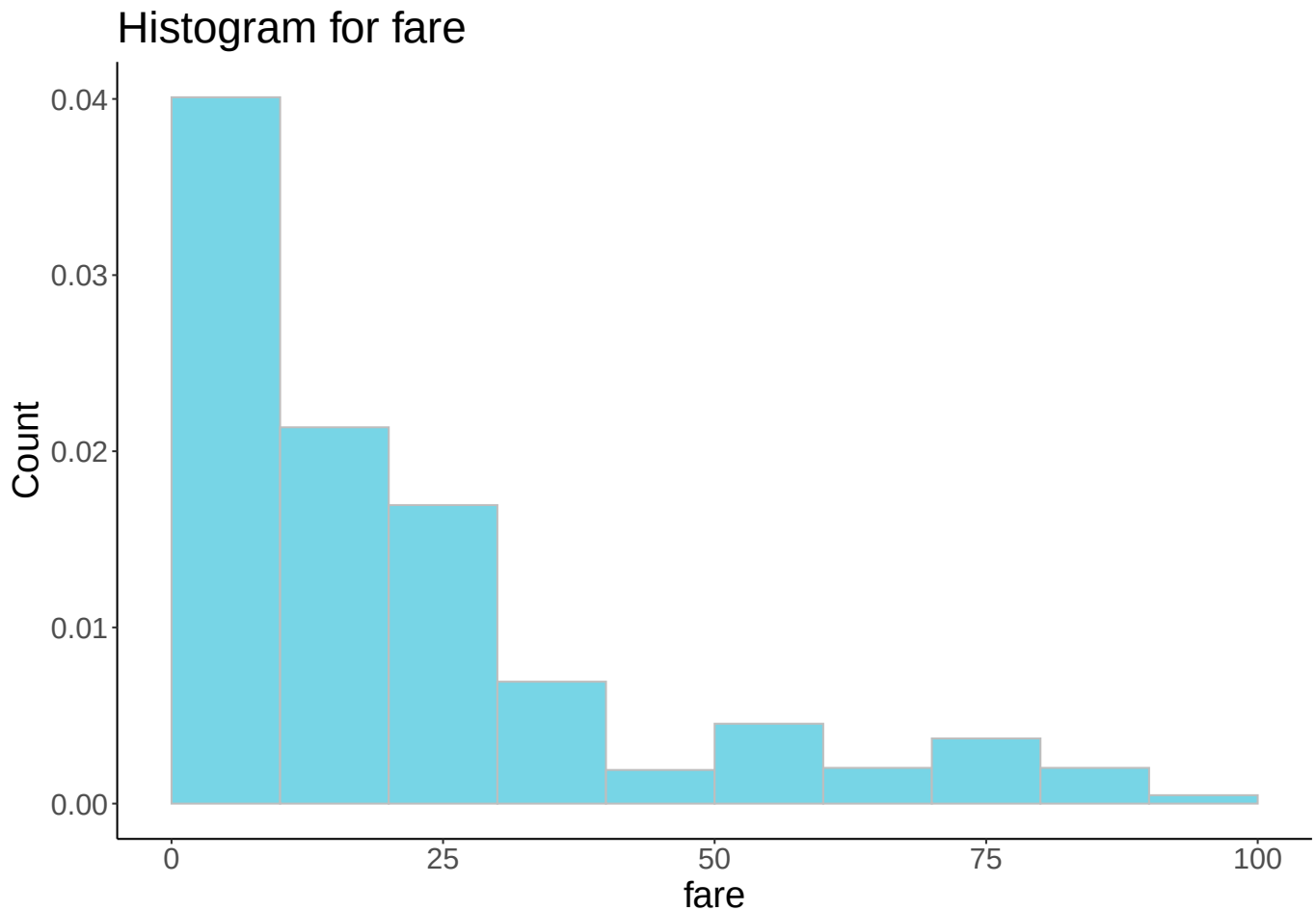


Figure 9: Histogram fare

Gruppenbildung

Klassenbreite b nach von Sturges

$$b = \frac{V}{1 + 3.32 \cdot \lg n} \approx \frac{V}{5 \lg n}$$

n - der Stichprobenumfang (Anzahl der Messwerte)

V - die Variationsbreite (Spannweite)

$\lg n$ - Zehnerlogarithmus von n

Histogramm (relative Häufigkeit) (3)

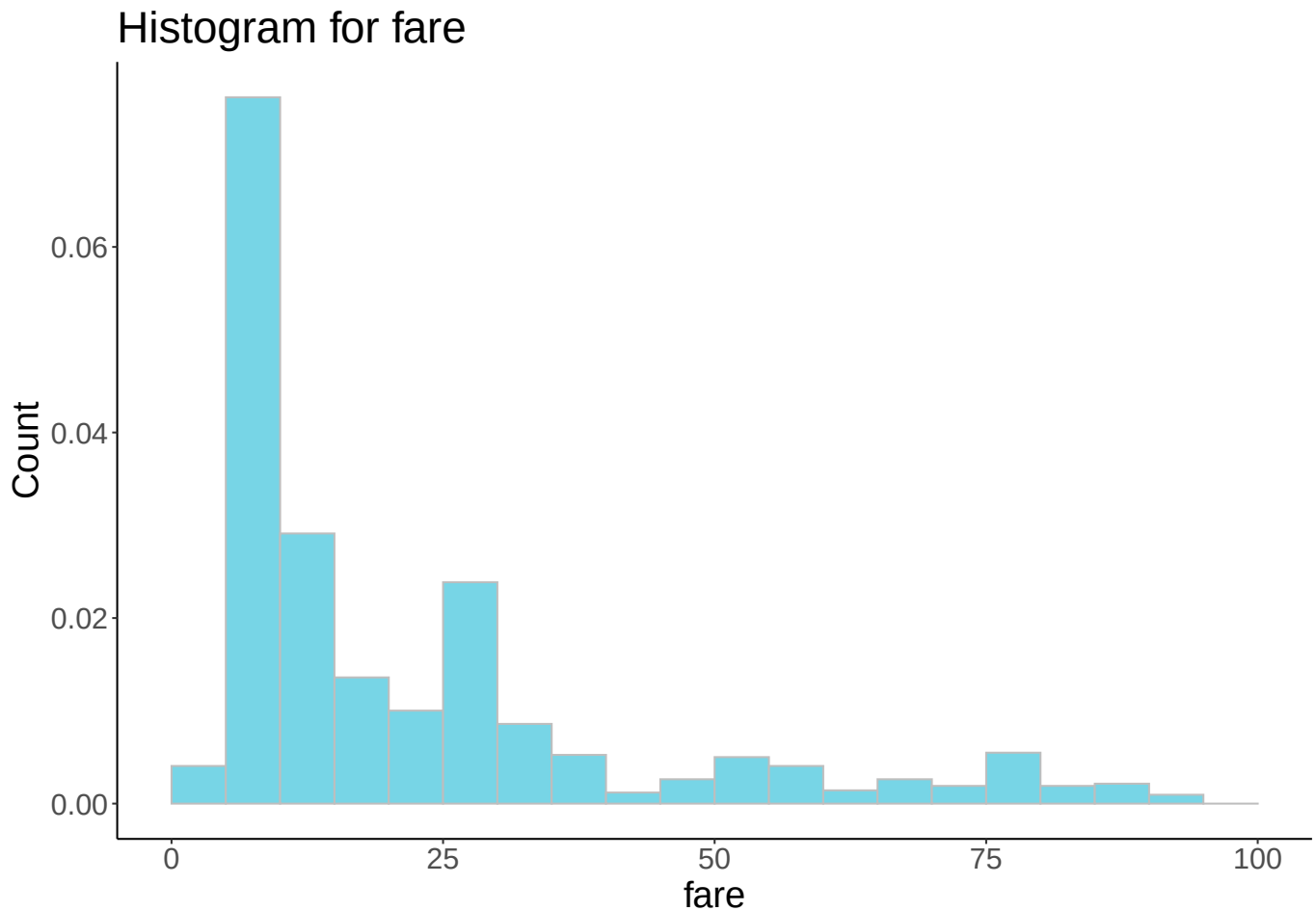


Figure 10: Histogram fare

Maßzahlen

Maßzahlen

Lageparameter

Streuungmaße

Lageparameter

Variationsbreite (Range)

(arithmetischer) Mittelwert

Median

Modus oder Modaler Wert

geometrischer Mittelwert

harmonischer Mittelwert

Streuungsmaße

Varianz und Standardabweichung

Quantile

Variationskoeffizient

Variationsbreite (Range)

$$V_x = \max(x) - \min(x)$$

(arithmetischer) Mittelwert

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Nutzliche Eigenschaft

$$\sum_{i=1}^{N} cx_i = c \left(\sum_{i=1}^{N} x_i \right)$$

Der Mittelwert ist sehr empfindlich was extreme Werte betrifft

Median

Zentralwert

Ungerade Anzahl der Beobachtungen

$\left(\frac{n+1}{2}\right)$. Beobachtungswert

Gerade Anzahl der Beobachtungen

Mittelwert von $\left(\frac{n}{2}\right)$. und $\left(\frac{n}{2}+1\right)$. Beobachtungswerten

Der Median ist hauptsächlich bestimmt durch die Werte in der Mitte der Stichprobe und ist wenige abhängig von den extremen Werten

Modus oder Modaler Wert

Dichtemittel

Der häufigste Wert. Wenn alle Werte nur einmal vorkommen, gibt es keinen Modus.

Geometrischer Mittelwert

$$G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln x_i}$$

wird z.B. für die Bestimmung der MIC benutz ($2^k c$, $k = 1, 2, \ldots$)

Harmonischer Mittelwert

$$H = \frac{1}{n}\left(\frac{1}{x_1} + \ldots + \frac{1}{x_n}\right)^{-1}$$

Vergleich vom Maßzahlen

Histogramm (relative Häufigkeit) (3)

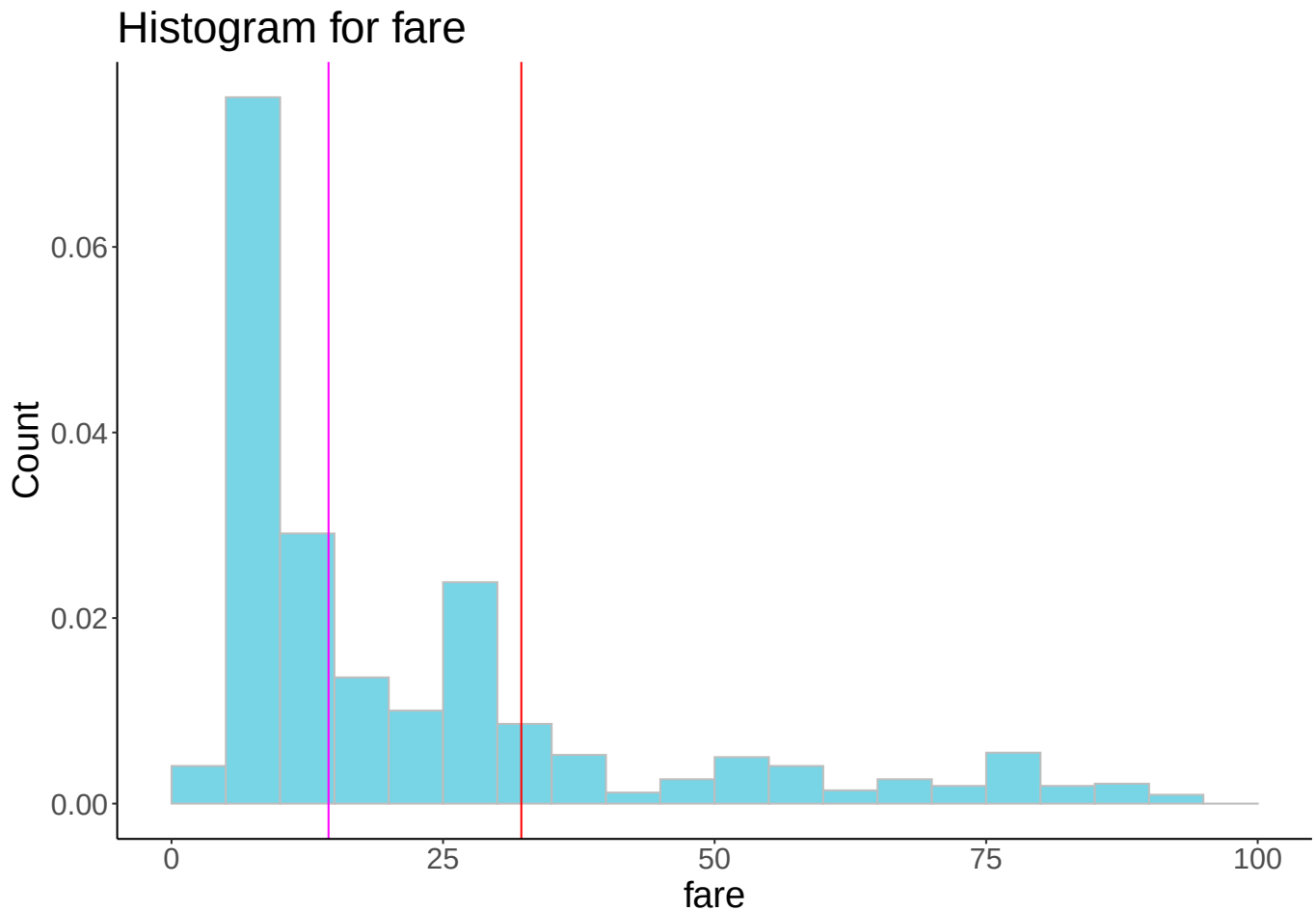


Figure 11: Histogram fare

Streuungsmaße

Varianz und Standardabweichung

Varianz

$$s_x^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2$$

$n - 1$ - Freiheitsgrad

Standardabweichung

$$s_x = \sqrt{\frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2}$$

Der mittlere Fehler des Mittelwertes

Streuung von $\bar{x}$ um den wahren Mittelwert μ der Grundgesamtheit

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Variationskoeffizient

Das Verhältnis der Standardabweichung zum Mittelwert

$$\mathrm{CV} = \frac{s}{|\bar{x}|}$$

Erlaubt unambhängig vom Mittelwert die Streuung der Daten zu Vergleichen

Quantile

p. Perzentil

- $(k+1)$. Datenpunkt wenn $\frac{np}{100}$ nicht ganzzahlig ist $(k < \frac{np}{100}, k \in \mathbb{N})$
- Durchschnitt von k. und $(k+1)$. Datenpunkt wenn $\frac{np}{100}$ ganzzahlig ist

1. Quantil (Q_1) - Der Datenpunkt wo 25% Messpunkte unterhalb und 75% oberhalb liegen
2. Quantil (Q_3) - Der Datenpunkt wo 75% Messpunkte unterhalb und 25% oberhalb liegen
3. Quantil (Q_2) - Median

Interquartilsabstand und MAD

(Inter)quartilsabstand

$$IQR = Q_{0.75} - Q_{0.25}$$

Median-Abweichung (Mean Abslolute Deviation)

$$|x_i - Q_{0.5}|$$

Kumulative Verteilung

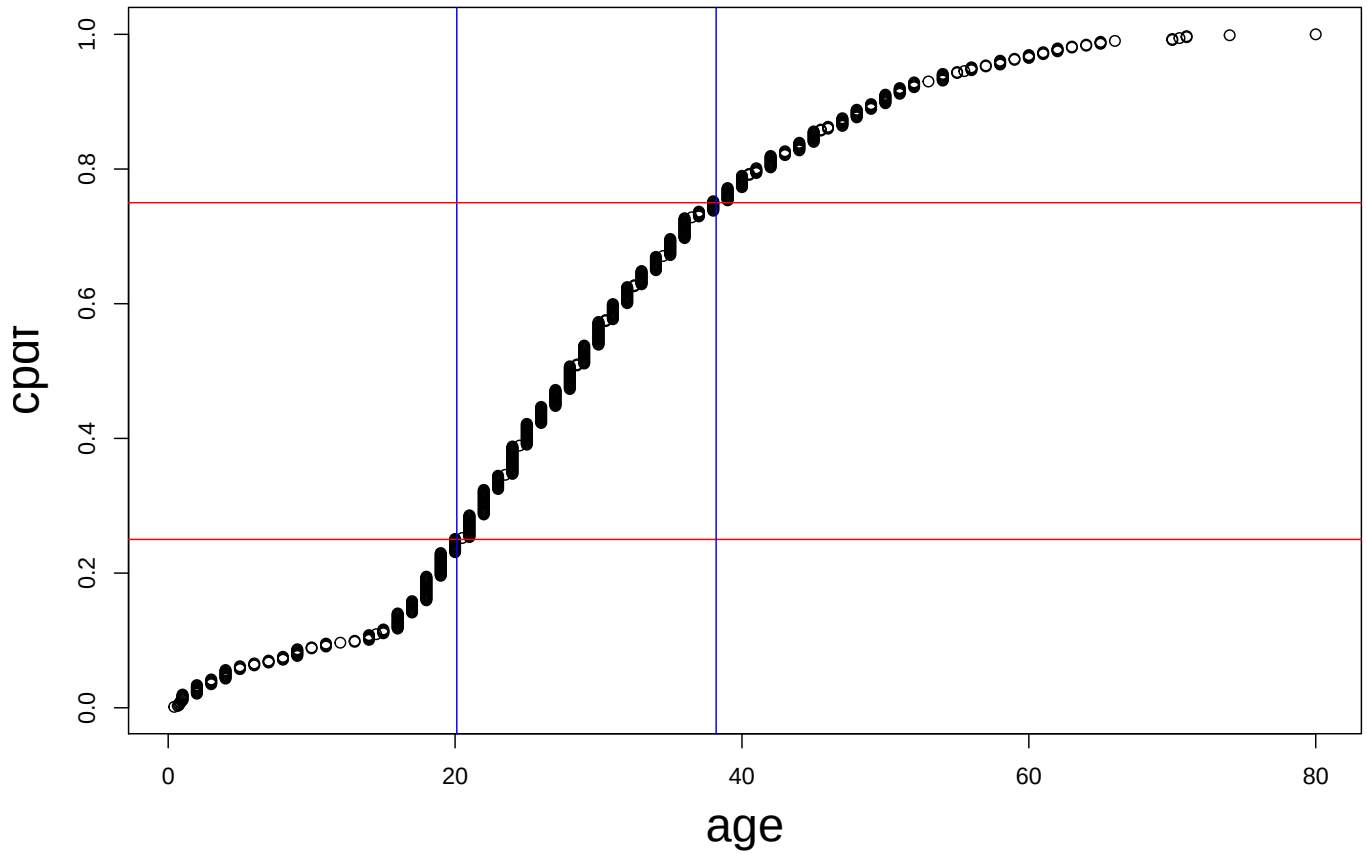
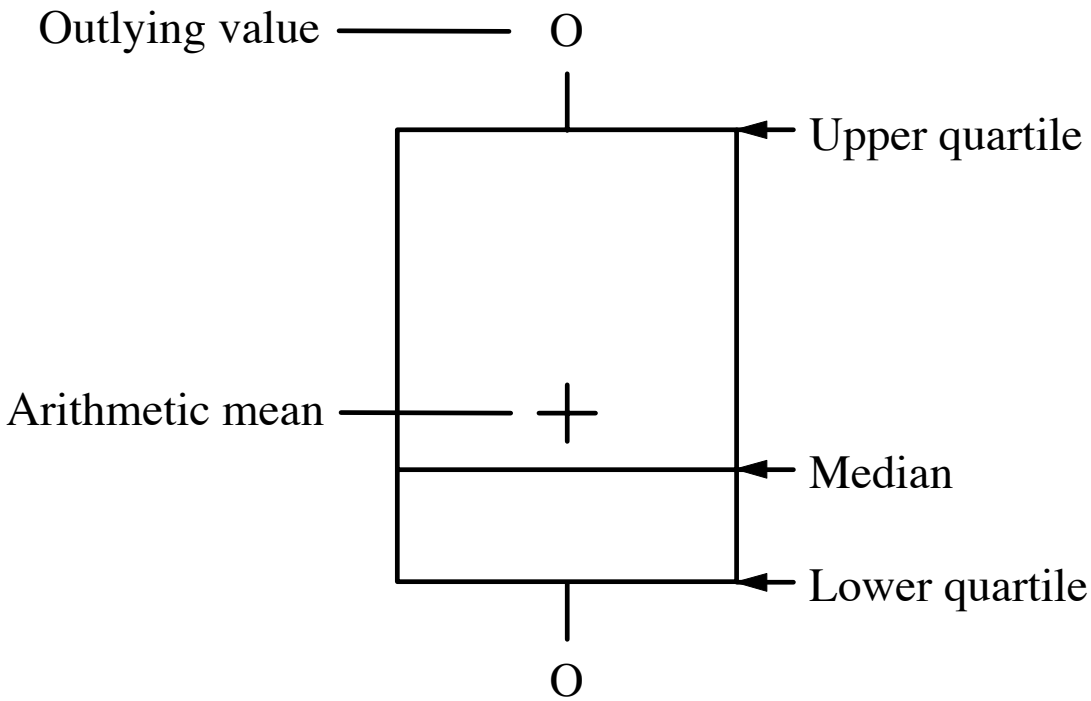


Figure 12: Kumulative Verteilung

Boxplot


Outlying value
O
Upper quartile
Arithmetic mean
Median
Lower quartile
O

Boxplot

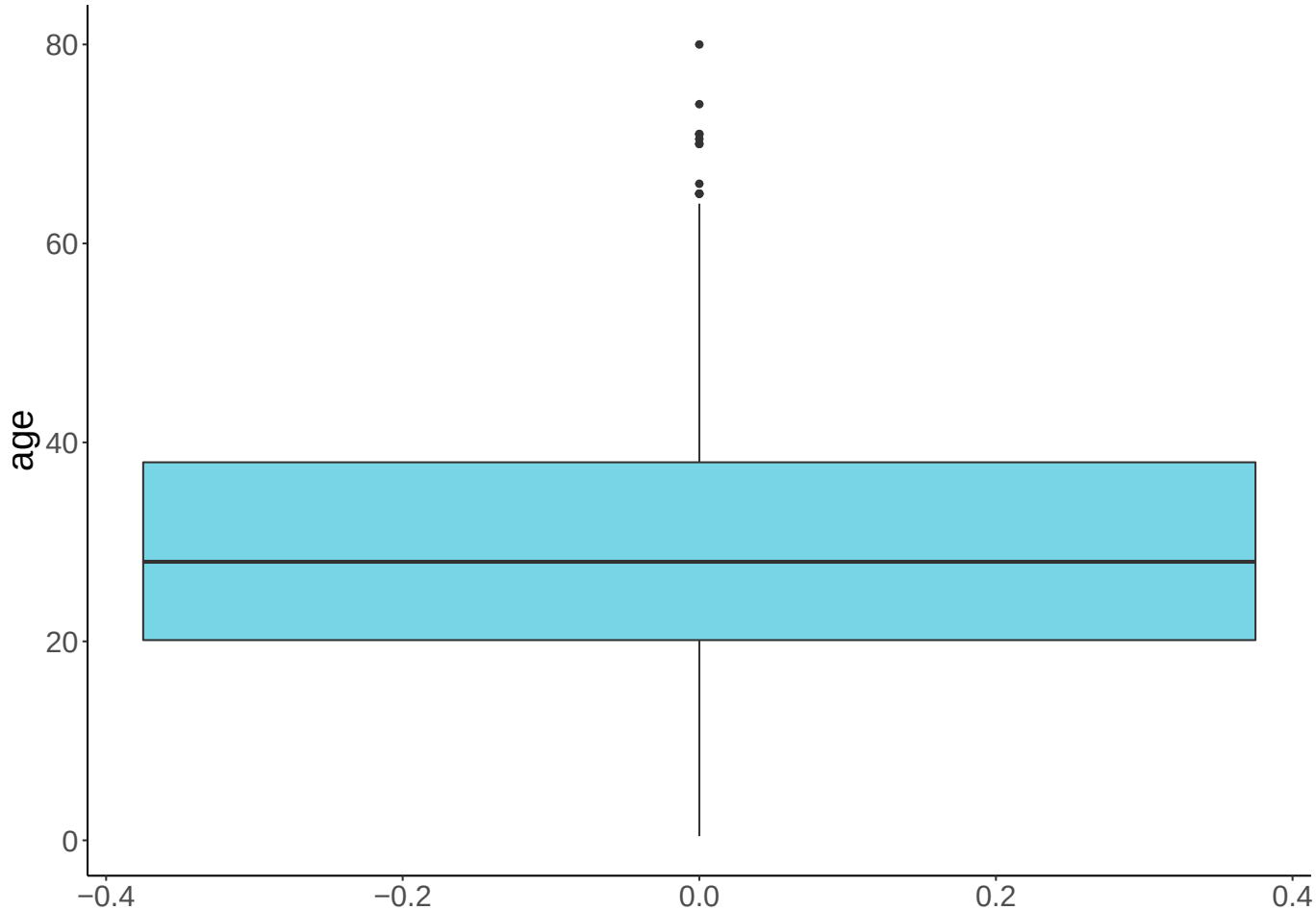


Figure 13: Boxplot

Boxplot

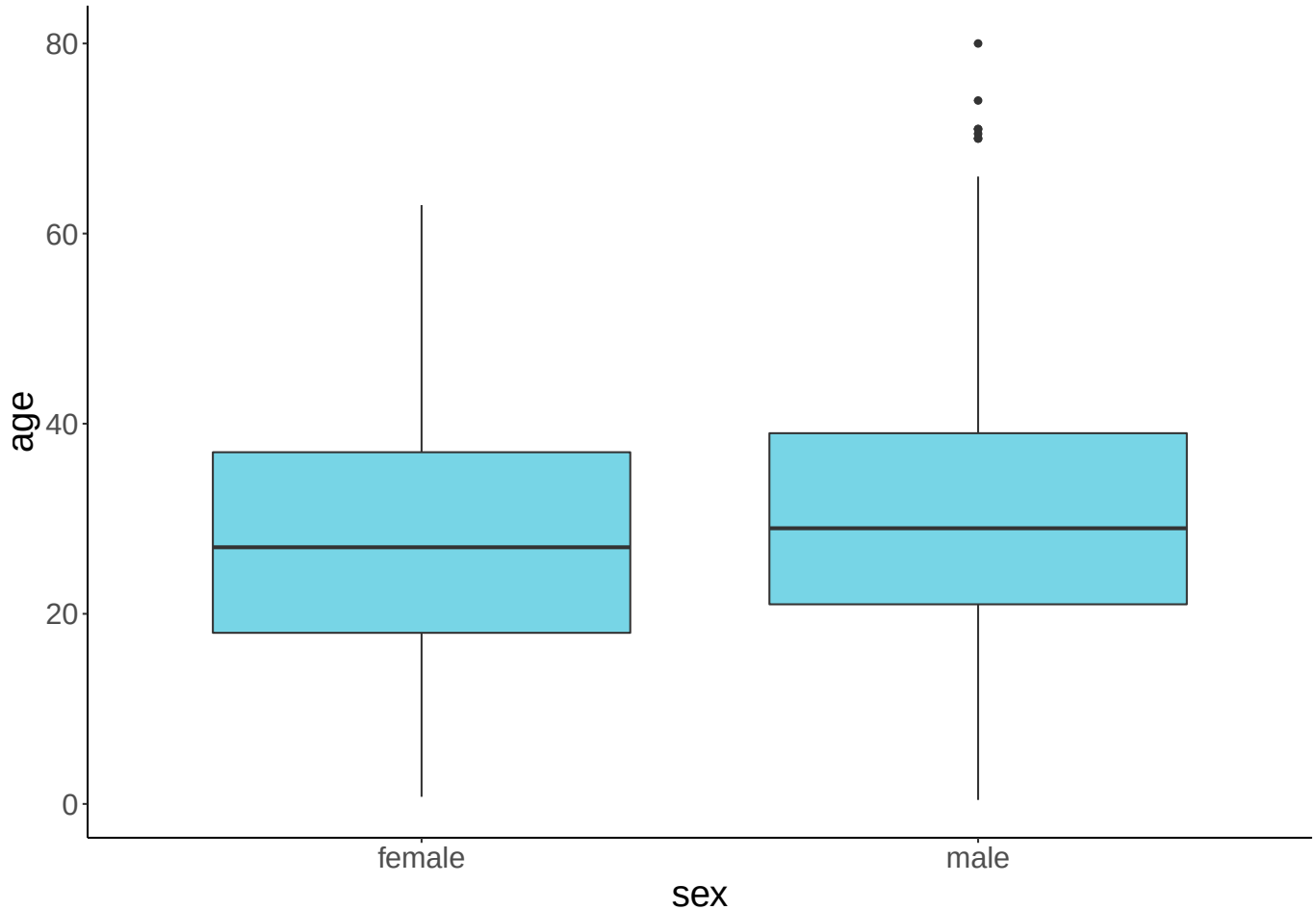


Figure 14: Boxplot

Diversitätsindizes

$$S = \sum_i p_i \ln p_i$$