

§ 3. Maße für Lage und Streuung

Motivation: Datensätze mit Hilfe von (eindimensionalen) Kennzahlen beschreiben.

Hier nehmen wir explizit quantitative Daten mit Wertebereich $W \subseteq \mathbb{R}$. Die relevanten Informationen, die wir in diesem Abschnitt verstehen wollen sind (1) wo befinden sich die Daten und (2) wie weit verstreut sind die Daten.

Beispiel: Verteilung von Temperaturdaten.

wo: wenn die Daten alle bei um die 20°C liegen, dann handelt es sich um einen warmen Ort.

wie weit: Auf den Kanarischen Inseln ist das ganze Jahr über ca. 25°C . In Berlin schwankt die Temperatur über das Jahr zwischen -10°C und $+30^\circ\text{C}$.

Teil 1: Lageparameter

Definition (Rangwerte)

Seien $x_1, \dots, x_N \in \mathbb{R}$ Beobachtungen eines quantitativen Merkmals, so dass $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(N)}$. Wir nennen:

- $x_{(n)}$ den n -ten Rangwert.

Wir nennen: $x_{(1)}$ das Minimum und $x_{(N)}$ das Maximum.

Beispiel: Daten $x_1 = 3, x_2 = 6, \underline{x_3 = 1}, \underline{x_4 = 1}, x_5 = 4$.

Rangwerte, $x_{(1)} = 1, x_{(2)} = 1, x_{(3)} = 3, x_{(4)} = 4, x_{(5)} = 6$.

Rangwerte liefern Information darüber, in welchem Bereich sich die Daten befinden. Die Lage innerhalb dieses Bereichs wird durch Quantile beschrieben.

Idee Für $0 < p < 1$ ist das p -Quantil ein Punkt $\tilde{x_p}$, so dass $\approx p \cdot 100\%$ der Daten kleiner als $\tilde{x_p}$ sind.

$$x_{(1)} \leq \dots \leq x_{(k)} \leq \tilde{x_p} \leq x_{(k+1)} \leq \dots \leq x_{(N)}, \text{ sodass } k \approx p \cdot N.$$

Beispiel: $p = \frac{1}{2}$: ($\frac{1}{2}$ -Quantil = Median).

$$x_1 = 1, x_2 = 4, x_3 = 1, x_4 = 6$$

Rangwerte $x_{(1)} = 1 \leq x_{(2)} = 1 \leq x_{(3)} = 4 \leq x_{(4)} = 6$

$$\tilde{x_{\frac{1}{2}}} := 2.5$$

Jetzt haben wir zusätzlich $x_{(5)} = 10$

$$x_{(1)} = 1 \leq x_{(2)} = 1 \leq \underbrace{x_{(3)} = 4}_{= \tilde{x_{\frac{1}{2}}}} \leq x_{(4)} = 6 \leq x_{(5)} = 10$$

Definition (Quantile)

Seien $x_1, \dots, x_N \in \mathbb{R}$ Beobachtungen eines Merkmals x .

Sei $0 < p < 1$. Das p -Quantil \tilde{x}_p ist definiert als

$$\tilde{x}_p = \begin{cases} x_{(k)} & , \text{ falls } p \cdot N \notin \mathbb{N} \text{ und } pN < k < pN+1. \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & , \text{ falls } p \cdot N = k \in \mathbb{N}. \end{cases}$$

Definition

Das $\frac{1}{2}$ -Quantil heißt **Median**.

Das $\frac{1}{4}$ -Quantil heißt **unteres Quartil**

Das $\frac{3}{4}$ -Quantil heißt **oberes Quartil**

In R:

`median(...)`

`quantile(...)`

Definition (Mittelwert)

Seien $x_1, \dots, x_N \in \mathbb{R}$ Beobachtungen eines Merkmals x .

Der **Mittelwert** der Daten ist definiert als

$$\bar{x} := \frac{1}{N} (x_1 + x_2 + \dots + x_N).$$

In R: `mean(...)`.

Mittelwert und Median sind beides Maße für das „Zentrum“ der Daten.

Beispiel $x_1 = 3, x_2 = 10, x_3 = 1, x_4 = 1, x_5 = 5$

Dann (a) Median: $N = 5, p = \frac{1}{2} \rightsquigarrow \underbrace{N \cdot p}_{=2,5} \notin N \rightsquigarrow 2,5 < k < 3,5 \Rightarrow k = 3$
 $\Rightarrow \tilde{x}_{1,5} = x_{(3)} = 3$

(b) Mittelwert: $\frac{1}{5}(3 + 10 + 1 + 1 + 5) = \frac{20}{5} = 4$

Beobachtung Extremwerte in den Beobachtungen ziehen \bar{x} von $\tilde{x}_{1,2}$ weg.
($= x_2 = 10$).

Teil 2 Streuungsparameter

Definition (Streuungsparameter)

Seien $x_1, \dots, x_N \in \mathbb{R}$ Beobachtungen eines Merkmals X . Wir kennen:

$R := x_{(N)} - x_{(1)}$ die Spannweite

$Q_1 = \tilde{x}_{1,4} - \tilde{x}_{3,4}$ den Quartilsabstand

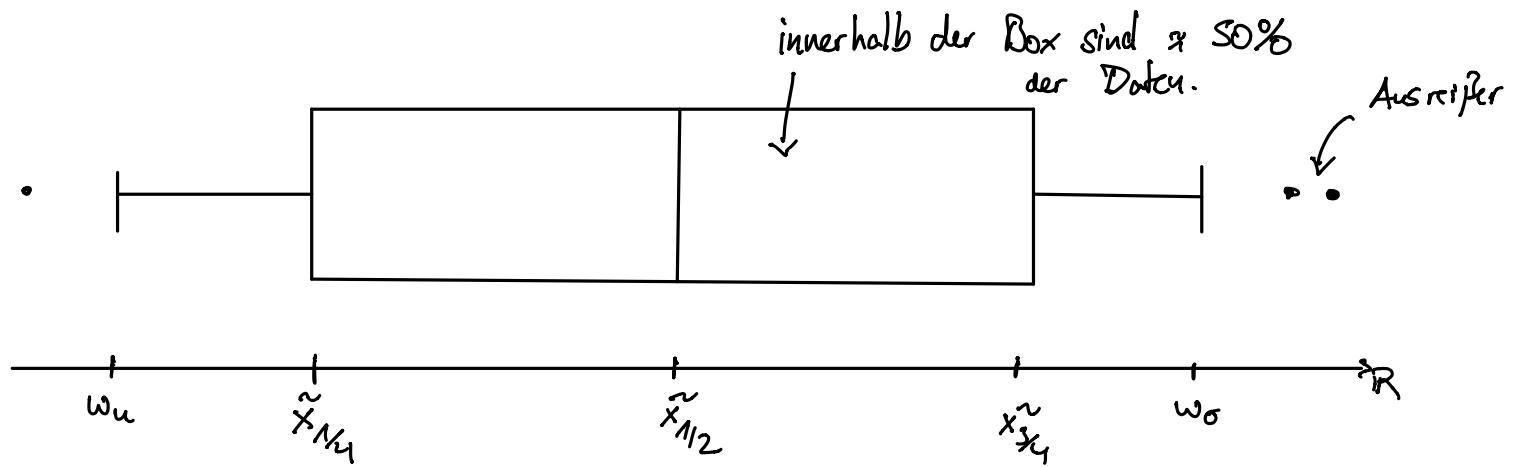
$S := \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$ die Standardabweichung der Daten.

Außerdem, nennen wir $s^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x} - x_i)^2$ die Varianz der Daten, oder auch Stichprobenvarianz.

In R: $sd(\dots)$ für die Standardabweichung
 $var(\dots)$ für die Varianz.

Varianz und Standardabweichung messen den Mittelwert der quadratischen Abweichung der Beobachtungen x_i zum Mittelwert \bar{x} .

Visualisierung: Die Informationen über Lage und Streuung werden in einem **Boxplot** zusammengefasst.



wobei $w_u :=$ kleinste Beobachtung $> \tilde{x}_{1/4} + \frac{3}{2} (\overbrace{\tilde{x}_{3/4} - \tilde{x}_{1/4}}^Q)$.
 $w_o :=$ größte Beobachtung $< \tilde{x}_{3/4} + \frac{3}{2} (\tilde{x}_{3/4} - \tilde{x}_{1/4})$.

w_u, w_o heißen **unterer und oberer Whisker**.

Ausreißer = Daten die $< w_u$ oder $> w_o$ sind.