# Daten und Zufall

Marco Rehmeier

Universität Bielefeld

Wintersemester 2026/2027

Version vom 11.11.2026

Der Vorlesungstitel *Daten und Zufall* ist eine Umschreibung der mathematischen Disziplin der *Stochastik*. Die Stochastik beschäftigt sich mit der Beschreibung und Analyse des Zufalls und dessen Gesetzmäßigkeiten. Die Stochastik besteht aus den Teilgebieten der **Wahrscheinlichkeitstheorie** sowie der **Statistik**; letztere gliedert sich in *deskriptive* und *schließende* Statistik.

**Wahrscheinlichkeitstheorie:** axiomatische Beschreibung des Zufalls und Untersuchung von zusammengesetzten Zufallsexperimenten bei bekannten Grundeigenschaften (Mehfacher Wünzwurf bei fairer Münze: Wie wahrscheinlich ist "3 mal Kopf"bei fünf Würfen?; Wie lange muss man im Mittel warten bis zum zweiten mal 'Zahl' fällt?, usw.)

**Statistik:** Beschäftigt sich mit 'umgekehrten' Fragestellungen: Beobachtungen aus einem Zufallsexperiment sind gegeben, Ziel ist es, durch deren Analyse begründete Rückschlüsse auf Grundeigenschaften des Experiments zu ziehen (Eine Münze wurde 100 mal geworfen und zeigt 27 mal Kopf. Ist es vernünftig anzunehmen, dass die Münze 'fair' ist?).

*Deskriptive Statistik:* Aufbereitung der Daten, graphische/tabellarische Darstellung der für die jeweilige Fragestellung wesentlichen Aspekte.

*Schließende Statistik:* Durch Analyse von vorliegenden Daten soll Rückschluss auf die Grundgesamtheit gewonnen werden (Aus einer Stichprobe von 1000 Erwachsenen wählen 398 Partei X; was ist eine vernünftige Prognose für das Wahlergebnis dieser Partei bei einer deutschlandweiten Wahl?)

# Inhaltsverzeichnis

**1 Deskriptive Statistik** **1**
1.1 Einführung in die deskriptive Statistik . . . . . 1
1.2 Empirische Häufigkeitsverteilung . . . . . 3
1.3 Lagemaße . . . . . 6
1.4 Streumaße . . . . . 12
1.5 Lineare Regression . . . . . 13

# 1 Deskriptive Statistik

## 1.1 Einführung in die deskriptive Statistik

Die Objekte, auf die sich eine statistische Analyse bezieht, heißen *Untersuchungseinheiten*. Die Menge aller Untersuchungeinheiten bildet die *Grundgesamtheit*.

**Beispiel 1.1.1.** (a) Untersuchung sozialer Verhältnisse in Deutschland. Jede(r) Bürger*in ist hier eine Untersuchungseinheit; die Gesamtbevölkerung ist die Grundgesamtheit.

(b) Analyse der Wirtschaftskraft der Nahrungsindustrie. Jede Firma der Branche ist eine Untersuchungseinheit; die Menge aller dieser Firmen bildet die Grundgesamtheit.

Für statistische Untersuchungen sind typischerweise nur gewissen Eigenschaften von Untersuchungseinheiten relevant. Diese Eigenschaften werden *Merkmale* genannt. Mögliche Werte eines Merkmals heißen *Merkmalsausprägungen.*

**Beispiel 1.1.2** (Fortführung vorheriges Beispiel)**.** (a) Mögliche Merkmale sind (i) Einkommen, (ii) Vermögen oder (iii) Bildungsgrad. Merkmalsausprägungen für (i) sind Werte $x \in [0, \infty)$, $x \in (-\infty, \infty)$ für (ii) und etwa 'Abitur', "Hochschulabschluss", "Hauptschulabschluss", 'keiner' für (iii).

(b) Mögliche Merkmale sind etwa 'Umsatz des Vorjahres' oder 'Mitarbeiterzahl', mit Merkmalsausprägungen $x \in [0, \infty)$, bzw. $x \in \{0, 1, 2, \dots\}$.
Weitere Beispiele:

(c) Altersverteilung in Deutschland: Merkmal = Alter einer Person in Jahren. Merkmalsausprägungen $n \in \{0, 1, 2, \dots, 100, 101, \dots\}$. Da es nur sehr wenige Leute mit $n \geqslant 100$ gibt, kann man die Ausprägungen $\{100, 101, 102, \dots\}$ zu einer einzigen Ausprägung '$\geqslant 100$' zusammenfassen. So gibt es dann nur endlich viele Ausprägungen $0, 1, 2, \dots, 99, ' \geqslant 100'$.

(d) Demographische Analyse. Merkmal: Familienstand einer Person; Ausprägungen: ledig, liiert, verheiratet, usw.

Es kann sein, dass für eine Untersuchung mehrere Merkmale von Bedeutung sind. Falls wir gleichzeitig die Merkmale 'Alter' und 'Familienstand' erheben, so erhalten wir 2-dimensionale Merkmale mit Ausprägungen $(0, \text{ledig})$, $(18, \text{verheiratet})$, etc.

**Klassen von Merkmalen.**

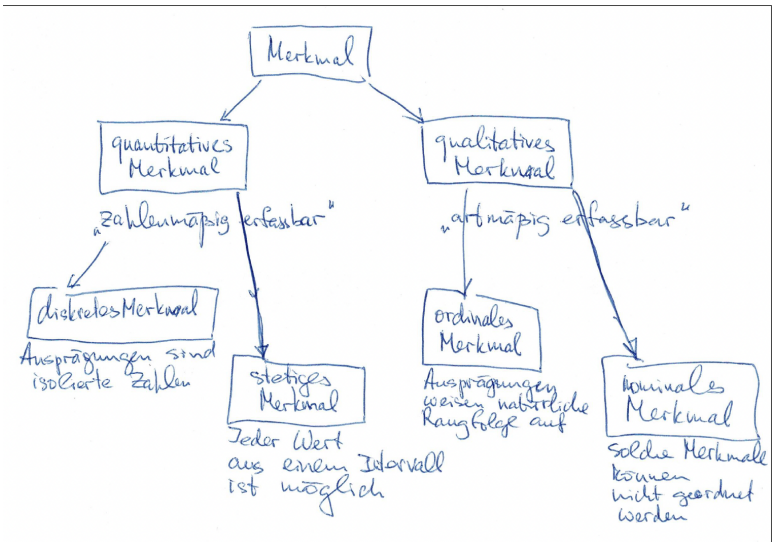



Abbildung 1.1: Klassifizierung von Merkmalen

**Beispiel 1.1.3.** Nominale Merkmale: Familienstand, Geschlecht, Parteipräferenz
Ordinale Merkmale: Schulnote (sehr gut, gut, ..., können durch Zahlwerte ersetzt und somit geordnet werden; dennoch kein quantitatives Merkmal, da Abstände zwischen den Notenzahlen nicht interpretiert werden können. Es gilt 'sehr gut besser als 'gut', 'gut' besser als 'befriedigend', aber ist der Abstand zwischen beiden Paaren gleich groß? Unklar!).
Diskretes Merkmal: Alter einer Person.
Stetiges Merkmal: Jahresumsatz eines Unternehmens.

Bei statistischen Untersuchungen kann man typischerweise nicht die Merkmale der Grundgesamtheit erfassen (Sie können realistischerweise nicht *jede(n)* Bürger*in nach seiner/ihrer Parteipräferenz befragen). Stattdessen wählt man zufällig eine endliche Teilmenge $M$ der Grundgesamtheit und erhebt Merkmale der Untersuchungseinheiten aus $M$. Eine solche Teilmenge von Ausprägungen heißt *Stichprobe*; die Anzahl der Elemente einer Stichprobe heißt *Umfang (der Stichprobe)*.

**Notation.** Typischerweise werden Merkmale mit $X, Y, Z, ...$ bezeichnet. Mathematisch kann man Merkmale als Funktionen auf der Grundgesamtheit betrachten; bei jeder Untersuchungseinheit ist die Merkmalsausprägung der Wert dieser Funktion. Falls wir eine endliche Stichprobe vom Umfang $n$ und ein Merkmal $X$ haben, so werden die Ausprägungen mit $x_1, x_2, ..., x_n$ bezeichnet; $x_k$ ist also die Ausprägung des $k$-ten Elements der Stichprobe. $x_k$ kann je nach Art des Merkmals eine Zahl (quantitatives Merkmal) oder ein Wert (Schulnote, Parteipräferenz) sein (qualitatives Merkmal).

## 1.2 Empirische Häufigkeitsverteilung

Sei $X$ ein Merkmal, das genau $s$ mögliche verschiedene Ausprägungen $a_1, a_2, \ldots, a_s$ hat.

**Definition 1.2.1.** Sei $x_1, x_2, \ldots, x_n$ eine Stichprobe des Merkmals $X$ vom Umfang $n$. Für jedes $j \in \{1, 2, \ldots, s\}$ setzen wir

$$H(a_j) := |\{i \in \{1, 2 \ldots, n\} : x_i = a_j\}| = \text{ Anzahl der Elemente der Stichprobe mit Ausprägung } a_j$$

(hier bezeichnet $|A|$ die Anzahl der Elemente in einer endlichen Menge $A$). $H(a_j)$ heißt die *absolute Häufigkeit* der Ausprägung $a_j$ (bezogen auf die betrachtete Stichprobe).

**Bemerkung 1.2.2.** *Die absolute Häufigkeit hängt von der Stichprobe ab; ändert man die Stichprobe, so wird sich möglicherweise die absolute Häufigkeit einzelner Ausprägungen ändern.*

Sei $M$ eine Menge und $N \subseteq M$ eine Teilmenge. Bezeichne mit $\mathbb{1}_N$ die Indikatorfunktion dieser Teilmenge, d.h.

$$\mathbb{1}_N(x) := \begin{cases} 1, \text{ falls } x \in N \\ 0, \text{ sonst.} \end{cases} \tag{1.2.1}$$

Dann kann man $H(a_j)$ wie folgt schreiben:

$$H(a_j) = \sum_{k=1}^{n} \mathbb{1}_{\{a_j\}}(x_k)$$

**Definition 1.2.3.** Die Funktion

$$H : \{a_1, a_2, \ldots, a_s\} \to \{0, 1, 2, \ldots, n\}, \quad a_j \mapsto H(a_j)$$

heißt *absolute* oder *empirische Häufigkeitsverteilung* der Stichprobe.

**Definition 1.2.4.** Sei erneut $n \in \mathbb{N}$ der Stichprobenumfang. Dann heißt $h(a_j) := \frac{H(a_j)}{n}$ die *relative Häufigkeit* der Ausprägung $a_j$ der Stichprobe.

Es gilt

$$\sum_{j=1}^{s} H(a_j) = n, \quad \sum_{j=1}^{s} h(a_j) = 1.$$

Die zweite Gleichheit rechtfertigt, dass relative Häufigkeiten auch als Prozentwerte angegeben werden.

**Bemerkung 1.2.5.** *Falls man zwei Stichproben eines Merkmals mit verschiedenen Umfängen $n$ und $m$ hat, so lassen sich nur* relative *Häufigkeiten sinnvollerweise vergleichen: wenn Sie 30 Menschen nach ihrer Parteipräferenz befragen und Ihre Kommilitonen 3000 Menschen befragt, ist der Vergleich wie viele Menschen Ihnen, bzw. Ihrer Kommilitonen 'Partei X' geantwortet haben, nicht sinnvoll, die relativen Häufigkeiten, also bspw. $\frac{7}{30}$ vs. $\frac{979}{3000}$ hingegen schon.*

Man kann die empirischen Häufigkeiten auf unterschiedliche Weisen darstellen. Diese Darstellungen sollen die Analyse der Daten erleichtern:

(a) Tabelle:

| Ausprägung | $a_1$ | $a_2$ | ... | $a_3$ |
|---|---|---|---|---|
| abs. Häufigkeit | $H(a_1)$ | $H(a_2)$ | ... | $H(a_3)$ |

Genauso kann man auch die relativen Häufigkeiten darstellen.

(b) Stabdiagramm: Absolute (oder relative) Häufigkeiten werden als Funktion der Ausprägung angezeigt, wobei $H(a_j)$ (bzw. $h(a_j)$) die Länge des Stäbchens über $a_j$ ist.

(c) Kreisdiagramm: Kreisfläche wird in Sektoren unterteilt, deren jeweilige Fläche proportional zu den absoluten (bzw. den relativen) Häufigkeiten der jeweiligen Ausprägung ist.

Falls der Umfang $n$ wesentlich kleiner als die Anzahl $s$ der möglichen Ausprägungen ist, so wird es mehrere Ausprägungen mit $H(a_j) = 0$ geben. Über solche Ausprägungen kann man dann keine quantitative Aussage treffen. Dieses Problem tritt insbesondere auf, wenn $s = \infty$, wenn es also unendlich viele verschiedene Ausprägungen gibt (bspw., falls Merkmal = privates Vermögen oder = persönliche Lieblingszahl $k \in \mathbb{N}$).

Als Lösung kann man *Merkmalsklassen* einführen. Dafür zerlegt man die Menge der möglichen Ausprägungen in paarweise disjunkte Mengen $K_1, K_2, \ldots, K_m$. Im Fall eines stetigen Merkmals mit Ausprägungswerten $[0, \infty)$ oder $(-\infty, \infty)$ wählt man $K_i$ typischerweise als Intervalle $K_1 = (t_0, t_1]$, $K_2 = (t_1, t_2]$, ..., $K_{m-1} = (t_{m-2}, t_{m-1}]$, $K_m = (t_{m-1}, t_m]$.

**Definition 1.2.6.** Ein *Histogramm* zur Klasseneinteilung $K_1, K_2, \ldots, K_m$ ist eine graphische Darstellung der relativen Häufigkeitsverteilung einer Stichprobe $x_1, x_2, \ldots, x_n$ eines quantitativen Merkmals mit möglichen Ausprägungen $A \subseteq [a, b]$ für gewisse $a, b \in \mathbb{R}$, die wir erhalten, indem wir über jedem Intervall $K_1, K_2, \ldots, K_m$ ein Rechteck errichten, dessen Flächen proportional zu den jeweiligen *relativen Klassenhäufigkeiten* $h(K_1), h(K_2), \ldots, h(K_m)$ sind, wobei

$$h(K_j) := \frac{\text{Anzahl der Stichprobenwerte in } K_i}{n} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{K_j}(x_k)$$

**Bemerkung 1.2.7.** *(a) Die Höhe $d_j$ des Rechtecks über $K_j$ bestimmt man aus der Formel*

$$d_j \cdot (t_j - t_{j-1}) = h(K_j), \quad j \in \{1, 2, \ldots, m\}.$$

(b) *Die Anzahl $m$ der Klassen kann frei gewählt werden. Wählt man zu wenig Klassen, erhält man zu grobe Informationen; wählt man zu viele, bleibt das ursprüngliche Problem bestehen. Bewährte Faustregel:* $m \approx \sqrt{n}$.

(c) *Klassen (für uns meist Intervalle) müssen nicht gleich groß sein (die Intervalle müssen also nicht gleich breit sein).*

**Stamm-und-Blatt-Diagramm.** *Ein* Stamm-und-Blatt-Diagramm *(SBD) liefert eine kompakte Veranschaulichung der Stichprobe mit geringem Informationsverlust; ein solches Diagramm funktioniert nur für quantitative Daten.*

1. Schritt: *Sortiere die Stichprobenwerte $x_1, \ldots, x_n$ der Größe nach. Bezeichne mit $x_{(1)}$ den kleinsten Stichprobenwert, mit $x_{(2)}$ den nächst kleinsten, ..., mit $x_{(n)}$ den größten.*

2. Schritt: *Wähle eine geeignete Skala.*

3. Schritt: *Schreibe alle bzgl. der gewählten Skala auftretenden Vorkommastellen von der kleinsten bis zur größten in die erste Spalte (Stamm).*

4. Schritt: *Schreibe Nachkommastellen jedes Stichprobenwerts hinter die entsprechende Vorkommastelle (Blätter).*

**Beispiel 1.2.8.** Länge des Schulwegs von Kindern aus einem Dorf in Metern, Stichprobenumfang = 8.

1014, 1223, 1130, 1047, 1351, 1234, 1407, 1170

1014, 1047, 1130, 1170, 1223, 1234, 1351, 1407

Skala: 100 Meter

| Stamm | Blätter |
|---|---|
| 10 | 14 47 |
| 11 | 30 70 |
| 12 | 23 34 |
| 13 | 51 |
| 14 | 07 |

**Bemerkung 1.2.9.** (a) *Aus einem SBD kann man die geordnete Liste $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eindeutig erstellen.*

(b) *Dreht man das SBD um 90, so erhält man das zugehörige Histogramm mit Klassenbreite von einer Skala-Einheit.*

## 1.3 Lagemaße

Gegeben sei eine Stichprobe $x_1, x_2, \ldots, x_n$ eines quantitativen oder ordinalen Merkmals, also $\{x_j\}_{j=1}^n \subseteq \mathbb{R}$. Wir wollen eine reelle Zahl $\ell(x_1, x_2, \ldots, x_n) \in \mathbb{R}$ finden, die die 'grobe Lage' der gesamten Stichprobe beschreibt.

Natürliche Anforderung: Verschiebt man alle Stichprobenwerte um denselben Wert $a \in \mathbb{R}$, so soll sich auch $\ell(x_1, x_2, \ldots, x_n)$ um $a$ verschieben.

**Definition 1.3.1.** Eine Funktion $\ell : \mathbb{R}^n \to \mathbb{R}$ heißt *Lagemaß*, falls für alle $x_1, x_2, \ldots, x_n \in \mathbb{R}$ und $a \in \mathbb{R}$ gilt:

$$\ell(x_1 + a, x_2 + a, \ldots, x_n + a) = \ell(x_1, x_2, \ldots, x_n) + a. \tag{1.3.1}$$

**Definition 1.3.2.** Die Zahl

$$\bar{x} := \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{k=1}^{n} x_k$$

heißt *arithmetisches Mittel* der Stichprobe $x_1, \ldots, x_n$ (Umgangssprachlich: Mittelwert, Durchschnitt).

Man kann $\bar{x}$ mit Hilfe der relativen Häufigkeiten darstellen: Seien $a_1, \ldots, a_s$ die möglichen verschiedenen Ausprägungen, dann gilt

$$\begin{aligned}\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k &= \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{s} a_j \mathbb{1}_{\{a_j\}}(x_k) \\ &= \frac{1}{n}\sum_{j=1}^{s} a_j \sum_{k=1}^{n} \mathbb{1}_{\{a_j\}}(x_k) \\ &= \sum_{j=1}^{s} a_j \big(\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{a_j\}}(x_k)\big) = \sum_{j=1}^{s} a_j jh(a_j).\end{aligned}$$

**Satz 1.3.3.** *(i) Das arithmetische Mittel ist ein Lagemaß.*

*(ii) Es gilt* $x_{(1)} \leqslant \bar{x} \leqslant x_{(n)}$.

*Beweis.* (i) Definiere $y_k := x_k + a$ für $k = 1, 2, \ldots, n$. Wir sehen

$$\begin{aligned}\ell(x_1 + a, \ldots, x_n + a) = \frac{1}{n}\sum_{k=1}^{n} y_k &= \frac{1}{n}\sum_{k=1}^{n}(x_k + a) \\ &= \frac{1}{n}\sum_{k=1}^{n} x_k + \frac{1}{n}\sum_{k=1}^{n} a = \bar{x} + \frac{1}{n}na = \bar{x} + a = \ell(x_1, \ldots, x_n) + a.\end{aligned}$$

(ii) Übung.

□

**Definition 1.3.4.** Seien $x_1, x_2, \ldots, x_n \in (0, \infty)$. Wir nennen

(a) $\bar{x}_g := \sqrt[n]{x_1 x_2 \cdots x_n}$ *geometrisches Mittel* von $x_1, \ldots, x_n$;

(b) $\bar{x}_h := \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots \frac{1}{x_n}}$ *harmonisches Mittel* von $x_1, \ldots, x_n$.

**Bemerkung 1.3.5.** *(a) $\bar{x}_g$ und $\bar{x}_h$ sind* keine *Lagemaße.*

*(b) Es gilt $x_{(1)} \leqslant \bar{x}_g \leqslant x_{(n)}$ und $x_{(1)} \leqslant \bar{x}_h \leqslant x_{(n)}$.*

**Proposition 1.3.6** (Eigenschaften des arithmetischen Mittels)**.** *(a)* $\sum_{k=1}^{n}(x_k - \bar{x}) = 0$.

*(b) Betrachte die Funktion $g_{x_1,\ldots,x_n} : \mathbb{R} \to [0, \infty)$, definiert durch $g_{x_1,\ldots,x_n}(c) := \sum_{k=1}^{n}(x_k - c)^2$. Dann ist $c = \bar{x}$ das eindeutige Minimum von $g_{x_1,\ldots,x_n}$.*

*Beweis.* (a) $\sum_{k=1}^{n}(x_k - \bar{x}) = \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$.

(b) Wir rechnen:

$$\begin{aligned}
&\sum_{k=1}^{n}(x_k - c)^2 = \sum_{k=1}^{n}(x_k - \bar{x} + \bar{x} - c)^2 \\
&= \sum_{k=1}^{n}\big((x_k - \bar{x})^2 + 2(\bar{x} - c)(x_k - \bar{x}) + (\bar{x} - c)^2\big) \\
&= \sum_{k=1}^{n}(x_k - \bar{x})^2 + \sum_{k=1}^{n} 2(\bar{x} - c)(x_k - \bar{x}) + \sum_{k=1}^{n}(\bar{x} - c)^2 \\
&= \sum_{k=1}^{n}(x_k - \bar{x})^2 + 2(\bar{x} - c)\Big(\underbrace{\sum_{k=1}^{n}(x_k - \bar{x})}_{=0}\Big) + n(\bar{x} - c)^2 \\
&= \sum_{k=1}^{n}(x_k - \bar{x})^2 + n(\bar{x} - c)^2 > \sum_{k=1}^{n}(x_k - \bar{x})^2 \text{ für alle } c \neq \bar{x}.
\end{aligned}$$

□

**Nun fragen wir:** Welche quantitativen Charakteristiken kann man für *qualitative* Merkmale einführen und nutzen?

**Definition 1.3.7.** Sei $H$ die absolute Häufigkeitsverteilung einer Stichprobe $x_1, \ldots, x_n$. Eine Ausprägung $a$ heißt *Modalwert*, falls

$$H(a) \geqslant H(x_i) \quad \text{für alle } i \in \{1, 2, \ldots, n\}.$$

Wir schreiben für den Modalwert einer Stichprobe auch $x_{\text{mod}}$ .

Der Modalwert kann für alle Merkmale, insbesondere also für qualitative Merkmale, definiert werden. Er muss nicht eindeutig sein, es kann also (offensichtlich) mehrere Modalwerte $x^1_{\text{mod}}$, $x^2_{\text{mod}}$, ... innerhalb einer Stichprobe geben.

Ist der Modalwert einer Stichprobe eindeutig, so heißt die zugehörige Häufigkeitsverteilung *unimodal*.

**Bemerkung 1.3.8.** *Sei $X$ ein quantitatives Merkmal. Dann ist der Modalwert ein Lagemaß, da für jedes $a \in \mathbb{R}$ gilt: für jedes $x_i$ ist die absolute Häufigkeit $H(x_i)$ bzgl. der Stichprobe $x_1, \ldots, x_n$ gleich der absoluten Häufigkeit $H(x_i + a)$ bzgl. der verschobenen Stichprobe $x_1 + a, \ldots, x_n + a$.*

Wir werden sagen, dass ein Merkmal $X$ *ordinal-skaliert* ist, falls $X$ entweder quantitativ oder ordinal ist. Jede Stichprobe $x_1, \ldots, x_n$ eines ordinal-skalierten Merkmals kann geordnet werden $(x_{(1)}, \ldots, x_{(n)})$.

**Definition 1.3.9.** Der *(empirische) Median* einer Stichprobe $x_1, \ldots, x_n$ eines ordinalskalierten Merkmals ist eine Ausprägung von $X$, bezeichnet $x_{0,5}$, definiert als

$$x_{0,5} := \begin{cases} x_{(\frac{n+1}{2})}, & \text{falls } n \text{ ungerade} \\ a, & \text{falls } n \text{ gerade und } a \text{ eine Ausprägung mit } a \in [x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}] \end{cases}$$

**Bemerkung 1.3.10.** *(a) Ist der Stichprobenumfang $n$ ungerade (z.B. 39), so ist $x_{0,5}$ also als der eindeutige Stichprobenwert definiert, für den jeweils mindestens $\frac{n+1}{2}$ Stichprobenwerte größer gleich, bzw. kleiner gleich $x_{0,5}$ sind (im Beispiel also jeweils mindestens 20 Stichprobenwerte größer gleich, bzw. kleiner gleich als $x_{0,5}$ sind).*

*(b) Ist $n$ gerade (z.B. 40), so ist $x_{0,5}$ i. A. nicht eindeutig bestimmt, sondern ein beliebiger Wert zwischen den beiden Werten des 'mittleren geordneten Wertepaares' (im Beispiel also ein beliebiger Wert zwischen $x_{20}$ und $x_{21}$). Oft wählt man dann*

$$x_{0,5} = x_{(\frac{n}{2})}$$

*(im Beispiel also $x_{0,5} = x_{(20)}$) oder, bei quantitativen stetigen Merkmalen,*

$$x_{0,5} = \frac{1}{2}\big(x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}\big).$$

***Achtung:** Für $n$ gerade muss $x_{0,5}$* keine *der in $x_1, \ldots, x_n$ beobachteten Ausprägungen sein. Verlangt wird per Definition jedoch, dass $x_{0,5}$ eine Ausprägung ist.*

**Beispiel 1.3.11.** (a) Merkmal $X$: Körpergröße eines Menschen in cm, mögliche Ausprägungen $a \in [0, 300]$, Stichprobenumfang $n = 9$:

| Stichprobe | Ausprägungswert |
|---|---|
| $x_1$ | 119 |
| $x_2$ | 201 |
| $x_3$ | 180 |
| $x_4$ | 167 |
| $x_5$ | 191 |
| $x_6$ | 167 |
| $x_7$ | 181 |
| $x_8$ | 177 |
| $x_9$ | 160 |

Also, nach Ordnung:

| Stichprobe | Ausprägungswert |
|---|---|
| $x_{(1)}$ | 119 |
| $x_{(2)}$ | 160 |
| $x_{(3)}$ | 167 |
| $x_{(4)}$ | 167 |
| $x_{(5)}$ | 177 |
| $x_{(6)}$ | 180 |
| $x_{(7)}$ | 181 |
| $x_{(8)}$ | 191 |
| $x_{(9)}$ | 201 |

Per Definition ist also $x_{0,5} = x_{(\frac{n+1}{2})} = x_{(5)} = 177$. Wir sehen: Jeweils $\frac{n+1}{2} = 5$ Werte sind größer gleich, bzw. kleiner gleich $x_{0,5} = 177$.

(b) Obige Stichprobe um einen zehnten Wert $x_{10} = 188$ erweitert. Nach Ordnung:

| Stichprobe | Ausprägungswert |
|---|---|
| $x_{(1)}$ | 119 |
| $x_{(2)}$ | 160 |
| $x_{(3)}$ | 167 |
| $x_{(4)}$ | 167 |
| $x_{(5)}$ | 177 |
| $x_{(6)}$ | 180 |
| $x_{(7)}$ | 181 |
| $x_{(8)}$ | 188 |
| $x_{(9)}$ | 191 |
| $x_{(10)}$ | 201 |

Da nun $n = 10$ gerade, müssen wir für $x_{0,5}$ einen Ausprägungswert aus dem Intervall $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}] = [x_{(5)}, x_{(6)}] = [177, 180]$ wählen. Jeder solche Wert ist eine mögliche Ausprägung (auch wenn nur die Werte 177 und 180 innerhalb der Stichprobe beobachtet wurden), also könnten wir $x_{0,5} \in [177, 180]$ beliebig wählen, etwa $x_{0,5} = 177$, $x_{0,5} = 179,6785849$, usw. Eine naheliegende Wahl scheint $\frac{1}{2}(177 + 180) = 178,5$ zu sein.

In jedem Fall sehen wir, dass $x_{0,5}$ größer gleich dem empirischen Median $x_{0,5} = 177$ aus Beispiel a) ist. Dies ist plausibel, schließlich haben wir die Stichprobe aus a) um einen Wert der größer 177 ist erweitert.

Hätten wir hingegen als Ausprägungen nur ganze cm-Werte zugelassen, also $0, 1, \ldots, 177, 178, 179, 180, \ldots$, aber nicht 179,6785849, dann kämen für $x_{0,5}$ nur $177, 178, 179, 180$ infrage.

Umgangssprachlich ist der empirische Median also ein Wert für den "jeweils mindestens die Hälfte der Stichprobenwerte größer, bzw. kleiner" als dieser Wert ist.

Im folgenden Satz diskutieren wir Charakterisierungen des Medians, d.h. wir formulieren notwendige und gleichzeitig hinreichende Bedingungen wann eine Ausprägung $a$ der Median einer gegebenen Stichprobe ist.

**Satz 1.3.12.** *Sei $X$ ein ordinal-skaliertes Merkmal mit endlich vielen Ausprägungen $a_1, a_2, \ldots, a_s$.*

*(a) Eine Ausprägung $a$ ist genau dann ein Median der Stichprobe $x_1, \dots, x_n$, wenn*

$$\sum_{j:a_j \leqslant a} h(a_j) \geqslant \frac{1}{2} \quad \textit{und} \quad \sum_{j:a_j \geqslant a} h(a_j) \geqslant \frac{1}{2}. \tag{1.3.2}$$

*(b) Ist $X$ ein quantitatives Merkmal, so gilt: die Summe $\sum_{k=1}^{n} |x_k - c|$ ist genau dann minimal, wenn $c$ ein Median ist.*

*Beweis.* (a) Sei zunächst $n$ ungerade, also $n = 2m + 1$ für ein $m \in \{0, 1, 2, \dots\}$. Dann ist $x_{(\frac{n+1}{2})}$ der Median und es gilt

$$\sum_{j:a_j \leqslant x_{(\frac{n+1}{2})}} h(a_j) = \frac{1}{n}\left|\{x_k : x_k \leqslant x_{(\frac{n+1}{2})}\}\right| \geqslant \frac{n+1}{2n} > \frac{1}{2} \tag{1.3.3}$$

und

$$\sum_{j:a_j \geqslant x_{(\frac{n+1}{2})}} h(a_j) = \frac{1}{n}\left|\{x_k : x_k \geqslant x_{(\frac{n+1}{2})}\}\right| = \frac{n+1}{2n} > \frac{1}{2}. \tag{1.3.4}$$

Sei nun $n$ gerade. Für jeden Median $a$ gilt $a \geqslant x_{(\frac{n}{2})}$ und $a \leqslant x_{(\frac{n}{2}+1)}$. Daraus folgt

$$\sum_{j:a_j \leqslant a} h(a_j) \geqslant \frac{1}{2} \quad \text{und} \quad \sum_{j:a_j \geqslant a} h(a_j) \geqslant \frac{1}{2}. \tag{1.3.5}$$

Also gilt (1.3.2) für jeden Median $a$.

Gelte nun umgekehrt (1.3.2) für eine Ausprägung $a$. Wir zeigen, dass $a$ ein Median der Stichprobe ist. Wegen (1.3.2) gilt

$$|\{k : x_k \leqslant a\}| \geqslant \frac{n}{2}, \text{ also } \{1, 2, \dots, \lceil \frac{n}{2} \rceil\} \subseteq \{i : x_{(i)} \leqslant a\}, \tag{1.3.6}$$

wobei $\lceil \frac{n}{2} \rceil$ die kleinste natürlich Zahl $\geqslant \frac{n}{2}$ bezeichnet (*obere Gaußklammer*). Analog

$$|\{k : x_k \geqslant a\}| \geqslant \frac{n}{2}, \text{ also } \{1, 2, \dots, n - \lceil \frac{n}{2} \rceil + 1\} \subseteq \{i : x_{(i)} \geqslant a\}. \tag{1.3.7}$$

Ist $n$ ungerade, so gilt $\lceil \frac{n}{2} \rceil = \frac{n+1}{2}$ und $n - \lceil \frac{n}{2} \rceil + 1 = \frac{n+1}{2}$. Also

$$\frac{n+1}{2} \in \{i : x_{(i)} \geqslant a\} \cap \{i : x_{(i)} \leqslant a\} = \{i : x_{(i)} = a\},$$

und somit ist $x_{(\frac{n+1}{2})} = a$ der Median der Stichprobe. Ist $n$ gerade, so ist $\lceil \frac{n}{2} \rceil = \frac{n}{2}$ und $n - \lceil \frac{n}{2} \rceil + 1 = \frac{n}{2} + 1$, also $x_{(\frac{n}{2})} \leqslant a$ und $x_{(\frac{n}{2}+1)} \geqslant a$. Also ist $a$ auch in diesem Fall ein Median.

(b) Setze $f(c) = \sum_{k=1}^{n} |x_k - c|$ und sei $x_{0,5}$ ein Median. Wir wollen zunächst zeigen:

$$f(c) - f(x_{0,5}) \geqslant 0, \quad \forall c \in \mathbb{R}.$$

Wir beschränken uns auf $c \leqslant x_{0,5}$, der Fall $c \geqslant x_{0,5}$ verläuft analog. Sei also $c \leqslant x_{0,5}$ beliebig.

Es ist $f(c) - f(x_{0,5}) = \sum_{k=1}^{n} \big(|x_k - c| - |x_k - x_{0,5}|\big)$. Für $x_k \leqslant c$ gilt

$$|x_k - c| - |x_k - x_{0,5}| = (c - x_k) - (x_{0,5} - x_k) = c - x_{0,5},$$

und für $c < x_k < x_{0,5}$

$$|x_k - c| - |x_k - x_{0,5}| = (x_k - c) - (x_{0,5} - x_k) = 2x_k - c - x_{0,5} = (c - x_{0,5}) + 2(x_k - c).$$

Im dritten und letztmöglichen Fall $x_k \geqslant x_{0,5}$ haben wir

$$|x_k - c| - |x_k - x_{0,5}| = (x_k - c) - (x_k - x_{0,5}) = x_{0,5} - c.$$

Aus diesen drei Umformungen folgt

$$\begin{aligned} f(c) - f(x_{0,5})) &= \sum_{k:x_k \leqslant c} (c - x_{0,5}) + \sum_{k:c<x_k<x_{0,5}} \big((c - x_{0,5}) + 2(x_k - c)\big) + \sum_{k:x_k \geqslant x_{0,5}} (x_{0,5} - c) \\ &= \sum_{k:x_k<x_{0,5}} (c - x_{0,5}) + \sum_{k:c<x_k<x_{0,5}} 2(x_k - c) + \sum_{k:x_k \geqslant x_{0,5}} (x_{0,5} - c) \\ &=: f_1(c) + f_2(c), \end{aligned}$$

mit

$$f_1(c) := \sum_{k:x_k<x_{0,5}} (c - x_{0,5}) + \sum_{k:x_k \geqslant x_{0,5}} (x_{0,5} - c)$$

und

$$f_2(c) := 2 \sum_{k:c<x_k<x_{0,5}} (x_k - c).$$

Dann gilt

$$f_1(c) = \underbrace{(x_{0,5} - c)}_{\geqslant 0} \big( \underbrace{|\{k : x_k \geqslant x_{0,5}\}|}_{\geqslant \frac{n}{2}} - \underbrace{|\{k : x_k < x_{0,5}\}|}_{\leqslant \frac{n}{2}} \big),$$

wobei die Ungleichungen "$\geqslant \frac{n}{2}$", bzw. "$\leqslant \frac{n}{2}$"nach Teil a) gelten. Also ist $f_1(c) \geqslant 0$. $f_2(c) \geqslant 0$ gilt per Definition. Somit ist also $f(c) \geqslant f(x_{0,5})$, also ist $x_{0,5}$ eine Minimumsstelle von $f$.

Ist nun $c \leqslant x_{0,5}$ mit $f(c) = f(x_{0,5})$, so bleibt zu zeigen, dass $c$ ebenfalls Median ist. Wegen $f(c) = f(x_{0,5})$ gilt $f_1(c) = 0 = f_2(c)$. Ist $n$ ungerade, so ist

$$|\{k : x_k \geqslant x_{0,5}\}| - |\{k : x_k < x_{0,5}\}| \geqslant \frac{n+1}{2} - \frac{n-1}{2} > 0,$$

also folgt $x_{0,5} = c$. Ist andererseits $n$ gerade und $c \leqslant x_{0,5}$ mit $f_1(c) = 0 = f_2(c)$, so ist $|\{k : c < x_k < x_{0,5}\}| = 0$. Da für $n$ gerade per Definition gilt

$x_{0,5} \in [x_{(\frac{n}{2})}, x_{(\frac{n+1}{2})}]$, folgt im Fall $x_{0,5} > x_{(\frac{n}{2})}$ bereits $c \in [x_{(\frac{n}{2})}, x_{(\frac{n+1}{2})}]$, also dass $c$ ebenfalls Median ist. Ist $x_{0,5} = x_{(\frac{n}{2})}$, so folgt

$$|\{k : x_k \geqslant x_{0,5}\}| - |\{k : x_k < x_{0,5}\}| \geqslant \frac{n+2}{2} - \frac{n-2}{2} > 0,$$

also folgt aus $f_1(c) = 0$ in diesem Fall bereits $c = x_{0,5}$. Also ist jede Minimumsstelle von $f$ ein Median. Dies schließt den Beweis.

□

**Bemerkung 1.3.13.** *Der Median ist ein Lagemaß.*

**Definition 1.3.14.** Sei $p \in [0,1]$ und $x_1, x_2, \ldots, x_n$ eine Stichprobe eines ordinalskalierten Merkmals. Eine Ausprägung $x_p$ hießt *p-Quantil*, falls

$$\sum_{j:a_j \leqslant x_p} h(a_j) \geqslant p \quad \text{und} \quad \sum_{j:a_j \geqslant x_p} \geqslant 1 - p.$$

$x_{0,25}$ und $x_{0,75}$ heißen *unteres*, bzw. *oberes Quantil*.

Nach Satz 1.3.12 (a) sind $0,5$-Quantile also gerade Mediane. Insofern sind $p$-Quantile also eine Verallgemeinerung des Begriffs des Medians. Im folgenden Satz bezeichnet $\lfloor x \rfloor$ die *untere Gaußklammer* für eine reelle Zahl $x$, also die größte ganze Zahl $z$ mit $z \leqslant x$. Insbesondere gilt $\lfloor x \rfloor = x$ falls $x$ eine ganze Zahl ist.

**Satz 1.3.15.** *Sei $p \in (0,1)$. Eine Ausprägung $a$ ist genau dann ein $p$-Quantil der Stichprobe $x_1, x_2, \ldots, x_n$, falls*

$$a = x_{(\lfloor np+1 \rfloor)}, \quad \textit{falls } np \notin \mathbb{N}$$

*oder*

$$x_{(np)} \leqslant a \leqslant x_{(np+1)}, \quad \textit{falls } np \in \mathbb{N}.$$

*Beweis.* Übung, analog zu Median. □

**Bemerkung 1.3.16.** *Im Fall $np \in \mathbb{N}$ wählt man das (in diesem Fall nicht-eindeutige) $p$-Quantil häufig wie beim Median, also als $x_{(np)}$ bei qualitativen Merkmalen und $\frac{1}{2}\big(x_{(np)} + x_{(np+1)}\big)$ bei quantitativen Merkmalen.*

## 1.4 Streumaße

Betrachte die folgende Situation: Bei einer Klausur haben zwei Klassen denselben Notendurchschnitt. In einer Klasse gibt es fast nur 3en, in der anderen hingegen viele 1en und 5en. Ein Lagemaß wird für beide Stichproben einen ähnlichen Wert nah an 3 liefern. Andererseits unterscheidet sich die Abweichungen von diesem Lagemaß erheblich. Solche Charakteristiken werden mit Hilfe von *Streumaßen* beschrieben.

**Definition 1.4.1.** Ein *Streumaß* ist eine Funktion $s : \mathbb{R}^n \to \mathbb{R}$ mit der Eigenschaft

$$s(x_1 + a, \ldots, x_n + a) = s(x_1, \ldots, x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R} \text{ und } a \in \mathbb{R}.$$

**Beispiel 1.4.2.** (i) $x_{(n)} - x_{(1)}$ (also der Abstand zwischen dem größten und kleinsten Wert einer Stichprobe eines ordinal-skalierten Merkmals) ist ein Streumaß.

(ii) $x_{0,75} - x_{0,25}$ (also der Abstand zwischen dem $0,75$- und $0,25$-Quantil der Stichprobe eines ordinal-skalierten Merkmals) ist auch ein Streumaß.

(iii) $\frac{1}{n}\sum_{k=1}^{n} |x_k - x_{0,5}|$ (also die mittlere absolute Abweichung vom Median) ist ein Streumaß.
Beweis: Übung!

Erinnerung: Für eine Stichprobe $x_1, \ldots, x_n$ eines ordinal-skalierten Merkmals schreiben wir $\bar{x} := \frac{1}{2}\sum_{k=1}^{n} x_k$ (arithmetisches Mittel).

**Definition 1.4.3.** Sei $x_1, \ldots, x_n$ eine Stichprobe eines ordinal-skalierten Merkmals. Die *mittlere quadratische Abweichung* von $\bar{x}$ ist durch

$$s^2 := \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

definiert und heißt *(empirische) Varianz* der Stichprobe. $\sqrt{s^2}$ heißt *(empirische) Standardabweichung* der Stichprobe.

**Satz 1.4.4.** *Varianz und Standardabweichung sind Streumaße.*

*Beweis.* Sei $a \in \mathbb{R}$ und $x_1, \ldots, x_n \in \mathbb{R}$. Setze $y_1 := x_1 + a, \ldots, y_n := x_n + a$. Dann gilt $\bar{y} = \bar{x} + a$, und daher auch

$$s_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k + a - (\bar{x} + a))^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = s_x^2.$$

□

**Satz 1.4.5.** *Für die Varianz $s^2$ gilt*

$$s^2 = \left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right) - (\bar{x})^2.$$

*Beweis.* Übung. □

## 1.5 Lineare Regression

Betrachte $n$ Untersuchungseinheiten und erhebe zu jeder Einheit zwei ordinalskalierte Merkmale $X$ und $Y$. Als Resultat erhält man eine zweidimensionale Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$. Unser Ziel ist es, zu verstehen, ob es einen Zusammenhang zwischen den Merkmalen $X$ und $Y$ gibt.

Als Beispiel betrachte folgende Situation: Untersuchungseinheiten sind Schüler*innen einer Klasse, $X$ und $Y$ seien die Mathe- bzw. Sport-Note einer/eines Schüler*in. Vermutung: *je besser die Leistung in Mathe, desto schlechter die Leistung in Sport.*

Zweites Beispiel: Untersuchungseinheiten sind Patient*innen; $X$ und $Y$ sind Alter, bzw. Herzfrequenz. Vermutung: *je höher das Alter, desto kleiner die Herzfrequenz.*

Die einfachste Form solch funktionaler Abhängigkeiten ist das *lineare Modell*:

$$Y = mX + t.$$

Fragen:

(i) Lässt sich die Vermutung durch ein lineares Modell beschreiben?

(ii) Falls ja, wie sollten $m$ und $t$ gewählt werden? Und wie findet man gute Werte für $m$ und $t$?

(iii) Wie gut ist die Beschreibung des Zusammenhangs von $X$ und $Y$ durch das gewählte lineare Modell?

Wenn wir die zweidimensionalen Stichproben $(x_k, y_k)$ als Punkte in der zweidimensionalen Ebene interpretieren, können wir eine Wolke aus $n$ Punkten visualisieren. Wenn das lineare Modell (für geeignete Werte $m$ und $t$) den Zusammenhang zwischen $X$ und $Y$ "gut"beschreibt, so sollte sich die Punktwolke um den Graph der Gerade $y = mx + t$ häufen.

Wie findet man nun also Werte $m$ und $t$, sodass die Punktwolke bestmöglich durch die entsprechende Gerade "getroffen"wird? Eine typische Möglichkeit: *Methode der kleinsten Quadrate*. Hier nutzt man die Summe der quadratischen Abweichungen zwischen den Stichprobenelementen und der Gerade als Qualitätsmaß der Approximation:

Setze für $x \in \mathbb{R}$

$$g(x) := mx + t$$

und betrachte

$$\sum_{k=1}^{n} \big(y_k - g(x_k)\big)^2.$$

Wir wollen $m$ und $t$ so wählen, dass diese Summe minimiert wird.

**Bemerkung 1.5.1.** *(i)*

*Falls $m$ und $t$ so gewählt werden können, dass die obige Summe gleich 0 ist, so liegen die Stichproben $(x_1, y_1), \ldots, (x_n, y_n)$ exakt auf einer Gerade. Dies ist im allgemeinen nicht zu erwarten (wenn Sie etwa von 1000 Patient*innen Alter und Herzfrequenz notieren, erwarten Sie nicht, dass für* jede(n) *Patient*in derselbe lineare Zusammenhang, etwa* Alter $= \frac{1}{3}\times$ *Herzfrequenz + 2 gilt).*

*(ii) Natürlich gibt es andere Möglichkeiten die Güte der Approximation der Punktwolke durch eine Gerade zu messen, etwa durch*

$$\sum_{k=1}^{n} |y_k - g(x_k)| \quad \textit{oder} \quad \sum_{k=1}^{n} \big(y_k - g(x_k)\big)^4, \ \textit{usw. ...}$$

*Es stellt sich jedoch heraus, dass die Methode der kleinsten Quadrate technisch besonders einfach ist.*

Definiere die Funktion

$$H(m,t) := \sum_{k=1}^{n} \big(y_k - g(x_k)\big)^2, \quad \forall m,t \in \mathbb{R}.$$

Wir wollen nun eine Minimumsstelle $(m,t)$ von $H$ finden. Dies gelingt in zwei Schritten:

1) Zunächst wird $m$ fixiert und der optimale Wert für $t$ gefunden. Dafür betrachte die Funktion

$$h(t) := \sum_{k=1}^{n} \big(y_k - mx_k - t\big)^2 = \sum_{k=1}^{n} (t + z_k)^2,$$

wobei $z_k := mx_k - y_k$. Wir können $h(t)$ wie folgt darstellen:

$$\begin{aligned} h(t) &= \sum_{k=1}^{n} (t^2 + 2z_k t + z_k^2) = n(t^2 + 2t\bar{z}) + \sum_{k=1}^{n} z_k^2 \\ &= n(t^2 + 2t\bar{z} + (\bar{z})^2) + n\big[\frac{1}{n}\sum_{k=1}^{n} z_k^2 - (\bar{z})^2\big] \\ &= n(t + \bar{z})^2 + n\big[\frac{1}{n}\sum_{k=1}^{n} z_k^2 - (\bar{z})^2\big], \end{aligned}$$

wobei $n\big[\frac{1}{n}\sum_{k=1}^{n} z_k^2 - (\bar{z})^2\big]$ nicht von der (einzigen) Variable $t$ abhängt. Dies zeigt, dass $h$ eindeutig an der Stelle $t_m := -\bar{z}$ minimiert wird. Beachte, dass $z_1, \ldots, z_n$, und somit auch $\bar{z}$ von $m$ abhängen. Wir sehen

$$t_m = -\bar{z} = -\frac{1}{n}\sum_{k=1}^{n} (mx_k - y_k) = \bar{y} - m\bar{x}.$$

2) Nun minimieren wir $H(m, t_m)$ in der Variable $m$. Es gilt

$$\begin{aligned} H(m,t_m) &= \sum_{k=1}^{n} (y_k - mx_k - \bar{y} + m\bar{x})^2 \\ &= \sum_{k=1}^{n} \Big( (y_k - \bar{y})^2 - 2(y_k - \bar{y})m(x_k - \bar{x}) + m^2(x_k - \bar{x})^2 \Big) \\ &= n\Big[\frac{1}{n}\sum_{k=1}^{n} (y_k - \bar{y})^2 - 2m\frac{1}{n}\sum_{k=1}^{n} (y_k - \bar{y})(x_k - \bar{x}) + m^2\frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2\Big]. \end{aligned}$$

Definiere jetzt

$$s_{xx} := \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2 \quad \text{(empirische Varianz von } x_1, \ldots, x_n)$$

$$s_{xy} := \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$s_{yy} := \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2 \quad \text{(empirische Varianz von } y_1, \ldots, y_n\text{)}.$$

Damit hat man

$$H(m, t_m) = n\Big[s_{xx}m^2 - 2s_{x,y}m + s_{yy}\Big].$$

Man sieht leicht (Übung!), dass diese Funktion eindeutig an der Stelle

$$m^* = \frac{s_{xy}}{s_{xx}}$$

minimal ist. Hierzu müssen wir annehmen, dass $s_{xx} \neq 0$ gilt, also dass $s_{xx} > 0$.

**Definition 1.5.2.** Die Größe

$$s_{xy} := \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

heißt die *(empirische) Kovarianz* der Stichproben $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$.

**Bemerkung 1.5.3.** *(i) Die Kovarianz von $x_1, \ldots, x_n$ und $x_1, \ldots, x_n$ ist gleich der Varianz von $x_1, \ldots, x_n$.*

*(ii) Für die Kovarianz gilt*

$$s_{xy} := \left(\frac{1}{n}\sum_{k=1}^{n} x_k y_k\right) - \bar{x}\bar{y}.$$

Durch die Rechnungen in den obigen Schritten haben wir folgenden Satz gezeigt.

**Satz 1.5.4.** *Gegeben seien $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ sodass nicht alle $x_k$ gleich sind ($\iff s_{xx} > 0$). Dann gibt es eindeutig bestimmte Zahlen $m$ und $t$ sodass*

$$\sum_{k=1}^{n}(y_k - mx_k - t)^2$$

*minimal ist. Diese Werte sind durch*

$$m^* = \frac{s_{xy}}{s_{xx}}, \quad t^* = \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x}$$

*gegeben.*

Die Gerade $g(x) = m^*x + t^*$ heißt *Regressionsgerade (der Stichproben $x_1, \ldots, x_n$, $y_1, \ldots, y_n$)*.

**Nächste natürliche Frage:** Wie gut beschreibt die Regressionsgerade mit diesen Werten für $m$ und $t$ die gegebene Punktwolke? Dies wird mit Hilfe der folgenden Größe beschrieben.

**Definition 1.5.5.** Sei $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}$ gegeben, sodass nicht alle $x_k$ und nicht alle $y_k$ gleich sind ( $\iff$ $s_{xx} > 0$ und $s_{yy} > 0$). Dann heißt

$$r_{xy} := \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$$

der *Korrelationskoeffizient der Stichprobe.*

**Satz 1.5.6.** *Es gilt*

*(i)* $s_{xy}^2 \leqslant s_{xx} s_{yy}$,

*(ii)* $|r_{xy}| \leqslant 1$,

*(iii)* $|r_{xy}| = 1 \iff$ *alle* $(x_k, y_k)$ *liegen auf einer Gerade.*

*Beweis.* Im Beweis von Satz 1.5.4 haben wir gesehen, dass für alle $m \in \mathbb{R}$ und zugehörige gerade $g_m(x) = mx + t_m$ (mit $t_m$ wie in Schritt 1 der Rechnung zum Beweis von Satz 1.5.4) gilt

$$0 \leqslant \sum_{k=1}^{n} \big(g_m(x_k) - y_k\big)^2 = n\big[s_{xx} m^2 - 2 s_{xy} m + s_{yy}\big].$$

Einsetzen von $m^*$ liefert

$$0 \leqslant \frac{s_{xy}^2}{s_{xx}} - 2\frac{s_{xy}^2}{s_{xx}} + s_{yy} = s_{yy} - \frac{s_{xy}^2}{s_{xx}}$$

und somit $s_{xy}^2 \leqslant s_{yy} s_{xx}$, was (i) beweist. Bzgl. (ii) beachte

$$r_{xy}^2 = \frac{s_{xy}^2}{s_{xx} s_{yy}} \leqslant 1 \implies |r_{xy}| \leqslant 1.$$

Für (iii) argumentieren wir wie folgt: Ist $|r_{xy}| = 1$, so gilt $s_{xy}^2 = s_{xx} s_{yy}$. Daraus folgt nach der obigen Rechnung zu Teil (i)

$$\sum_{k=1}^{n} \big(g_{m^*}(x_k) - y_k\big)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = 0,$$

also $g_{m^*}(x_k) = y_k$ für alle $k \in \{1, \ldots, n\}$. Liegen umgekehrt alle Punkte auf einer gemeinsamen Gerade, so gilt $y_k = m^* x_k + t^*$, also

$$\begin{aligned} s_{yy} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2 = \frac{1}{n}\sum_{k=1}^{n}(m^* x_k + t^* - \bar{y})^2 &= \frac{1}{n}\sum_{k=1}^{n}\Big(\frac{s_{xy}}{s_{xx}} x_k + \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x} - \bar{y}\Big)^2 \\ &= \frac{1}{n}\Big(\frac{s_{xy}}{s_{xx}}\Big)^2 \sum_{k=1}^{n}(x_k - \bar{x})^2 \\ &= \frac{s_{xy}^2}{s_{xx}}, \end{aligned}$$

also $|r_{xy}| = 1$. □

**Bemerkung 1.5.7.** *$|r_{xy}|$ kann also als Maß der Plausibilität eines linearen Zusammenhangs zwischen $X$ und $Y$ interpretiert werden:*

*(i)* *$|r_{xy}| = 1$ bedeutet perfekten linearen Zusammenhang (der Stichproben $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$, dies bedeutet nicht, dass für alle möglichen Wertepaare $(x, y)$ aus $X \times Y$ derselbe oder überhaupt ein perfekter linearer Zusammenhang vorliegt).*

*(ii)* *Ist $|r_{xy}|$ groß, bspw. $> \frac{4}{5}$, so liegt eine* starke Korrelation *zwischen den Stichproben aus $X$ und $Y$ vor.*

*(iii)* *Ist $|r_{xy}|$ klein (etwa $< \frac{3}{10}$), so liegt* schwache Korrelation *vor.*

*(iv)* *Ist $r_{xy} = 0$, so nennt man die Stichproben* unkorreliert.

*(v)* *Das Vorzeichen von $r_{xy}$ gibt die "Richtung"der Regressionsgerade vor.*

**Bemerkung 1.5.8.** *Aus einer Korrelation (auch aus einer starken) darf nicht allgemein auf einen* kausalen *Zusammenhang geschlossen werden, also nicht auf einen* Ursache-Wirkung-*Zusammenhang. Beachte, dass $r_{xy}$ symmetrisch in $x$ und $y$ ist, also*

$$r_{xy} = r_{yx}.$$

*Bei Annahme eines kausalen Zusammenhangs kann also aus dem Korrelationskoeffizienten nicht abgelesen werden was "Ursache" und was "Wirkung" sein sollte.*