**Teil 1**

# Deskriptive Statistik

Moodle

Lehrbuch

# Teil 1: Deskriptive Statistik

**Lernziele**

- Kennenlernen zentraler Grundbegriffe der Statistik
- Beherrschung wichtiger Lage- Streuungskenngrößen sowie deren Interpretation
- Fähigkeit zur Analyse mehrdimensionalen Datenmaterials mithilfe von Zusammenhangsmaßen
- Erlangung elementaren Wissens hinsichtlich Bestimmung und Interpretation von Preisindizes

# Teil 1: Deskriptive Statistik

# Kapitel 2

# Grundbegriffe der Datenerhebung

# Merkmal, Merkmalsträger, Merkmalsausprägung

**Merkmal**

Die Eigenschaften, über welche die Daten Informationen enthalten.

*Beispiele:* Einkommen, #Beschäftigte, Einwohner, BIP

**Merkmalsträger**

Das Objekt, dessen Daten erhoben werden.

*Beispiele:* Personen, Firmen, Städte, Staaten

**Merkmalsausprägungen**

Die möglichen Werte die ein Merkmal annehmen kann.

*Beispiele:* 17.889€, 25, 609546, 3870 Mrd.€

# Grundgesamtheit, Totalerhebung, Stichprobenerhebung

Die Menge aller Merkmalsträger heißt **Grundgesamtheit**.

*Beispiel:* Alle in Deutschland lebenden Personen

**Totalerhebung**

Vollständige Erfassung der Grundgesamtheit

**Teil–** oder **Stichprobenerhebung**

Unvollständige Erfassung der Grundgesamtheit

# Datentypen

**qualitative Merkmale**

Die Merkmalsausprägungen sind Wörter.

*Beispiele:* Farbe „blau“, Gemütszustand „gut“, Wetter „regnerisch“

**quantitative Merkmale**

Die Merkmalsausprägungen sind Zahlen.

Durch **Quantifizierung** werden qualitativen Merkmalen Zahlen zugeordnet.

*Beispiele:* atheistisch: 0, röm. kath.: 1, protestantisch: 2, andere: 3

# Skalierung

Die Skalierung ist bei quantitativen bzw, quantifizierten Ausprägungen wichtig.

**Nominalskala**

Die Zahlen dienen nur der Unterscheidung und Identifizierung.

**Ordinalskala**

Die Zahlen dienen außerdem einer Reihung der Ausprägungen in einer Rangordnung.

**Kardinalskala**

Zusätzlich zur Rangordnung können die Abstände zwischen den Zahlen sinnvoll interpretiert werden.

# Diskrete und stetige Merkmale

**Diskretes Merkmal**

Die Anzahl der möglichen Ausprägungen ist abzählbar.

*Beispiel:* {Wirtschaftswissenschaften, Biologie, Sportwissenschaften, Maschinenbau}

**Stetiges Merkmal**

Es gibt überabzählbar unendlich viele mögliche Ausprägungen.

*Beispiele:* Temperatur, Zeit mit Sonnenlicht, Entfernung

# quasistetige Merkmale und klassierte Daten

**Quasistetiges Merkmal**

Diskretes Merkmal mit sehr vielen Ausprägungen.

*Beispiele:* Einkommen, Gewicht in Gramm

**Klassierte Daten**

Unterteilung von stetigen oder quasistetigen Merkmalen in endlich viele Klassen

*Beispiele:* Einkommensklassen, Gewichtsklassen

# Kapitel 3

# Auswertungsmethoden für eindimensionale Daten

# Merkmalswerte und Urliste

"Stichprobengröße"

Für ein quantitatives oder quantifiziertes Merkmal $X$ von $n$ Merkmalsträgern heißen die erhobenen Zahlen $x_1, \ldots, x_n$ **Merkmalswerte**.

$x_i$ ist die beim $i$-ten Merkmalsträger beobachtete Merkmalsausprägung des Merkmals $X$.

**Urliste**

$n$-Tupel $(x_1, \ldots, x_n)$ aller $n$ Merkmalswerte

# Urliste

Die Urliste umfasst oft unübersichtlich viele Daten.

*Beispiel:*

Krankmeldungen der letzten 20 Tage:

| $x_1$ | $x_2$ | $x_3$ | | | | | | $x_{19}$ | $x_{20}$ |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 4 | 1 | 0 | 5 | 3 | 4 | 2 | 0 |
| 1 | 1 | 3 | 3 | 4 | 2 | 5 | 3 | 2 | 3 |

Bsp. $a_j = 3$

$\Rightarrow$ $h(a_j = 3) = 6$

$f(a_j = 3) = \frac{6}{20} = \frac{3}{10} = 30\%$

Dieser Urliste kann nur schwer sinnvolle Information entlesen werden.

# Absolute und relative Häufigkeiten

Wir ordnen die in der Urliste vorkommenden $k$ verschiedenen Ausprägungen der Größe nach:

$$a_1 < a_2 < \ldots < a_k$$

**Absolute Häufigkeit** von $a_j$

$h(a_j)$: Anzahl der $x_i$ für welche $x_i = a_j$ gilt.

**Relative Häufigkeit** von $a_j$

$f(a_j) = \frac{1}{n} \cdot h(a_j)$: Die absolute Häufigkeit dividiert durch die Stichprobengröße $n$

Es gilt:

$$\sum_{j=1}^{k} h(a_j) = n \text{ und } \sum_{j=1}^{k} f(a_j) = 1$$

# Beispiel: Krankmeldungen pro Monat in einer Firma

Urliste:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 4 | 1 | 0 | 5 | 3 | 4 | 2 | 0 |
| 1 | 1 | 3 | 3 | 4 | 2 | 5 | 3 | 2 | 3 |

Verschiedene Merkmalsausprägungen: $\{0, 1, \ldots, 5\}$

$n = 20$ Beobachtungen (Werktage)

**Häufigkeitstabelle**

| $a_j$ | 0 | 1 | 2 | 3 | 4 | 5 | $\sum$ |
|---|---|---|---|---|---|---|---|
| $h(a_j)$ | 2 | 3 | 4 | 6 | 3 | 2 | 20 |
| $f(a_j)$ | 0,1 | 0,15 | 0,2 | 0,3 | 0,15 | 0,1 | 1 |

# Häufigkeitsverteilung

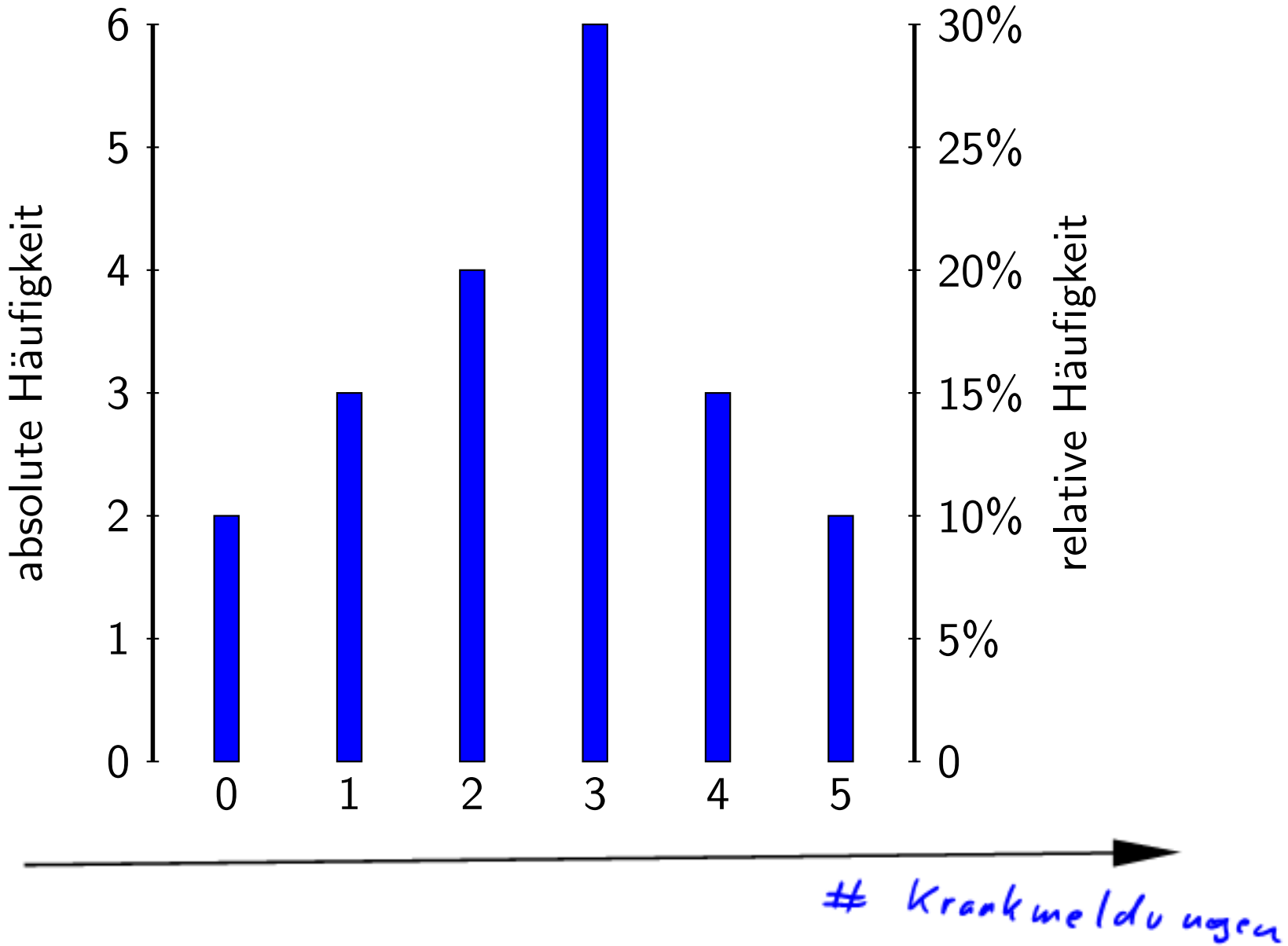
Die Häufigkeitstabelle fasst die **Häufigkeitsverteilung** zusammen.

Wir unterscheiden zwischen der Verteilung der absoluten und der relativen Häufigkeiten.

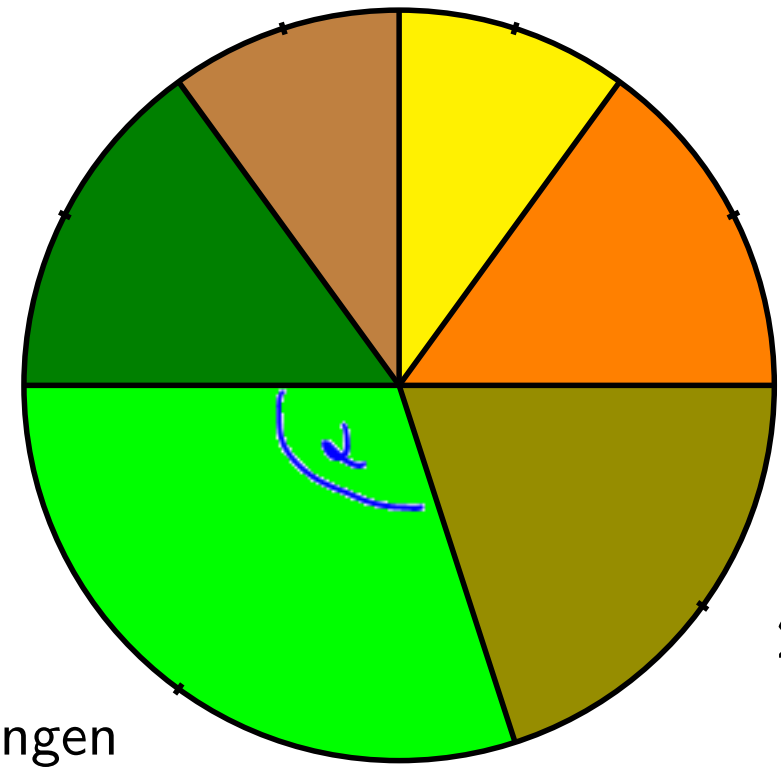
Graphische Darstellung der Häufigkeitsverteilung:

- Säulendiagramm
- Kreissektorendiagramm
- Histogramm

# Säulendiagramm (Bsp. Krankmeldungen)



# Kreissektorendiagramm (Bsp. Krankmeldungen)




$\alpha$ : 30% von 360°

$\rightsquigarrow \alpha = (360°/100) \cdot 30$

$= 360° \cdot \frac{30}{100} = 36° \cdot 3 = 108°$

# Histogramm

In Fällen mit vielen beobachteten Ausprägungen und jeweils nur kleinen Häufigkeiten sind Säulendiagramme nicht hilfreich.

**Beispiel 3.3: Lebensdauern von 30 Ventilen**

Das Merkmal $X$ ist die Lebensdauer in Stunden.

Urliste:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 110 | 520 | 490 | 30 | 120 | 290 | 370 | 305 | 415 | 170 |
| 280 | 70 | 540 | 460 | 260 | 345 | 150 | 220 | 435 | 425 |
| 470 | 350 | 130 | 380 | 230 | 320 | 360 | 240 | 330 | 580 |

# Säulendiagramm: Negativbeispiel

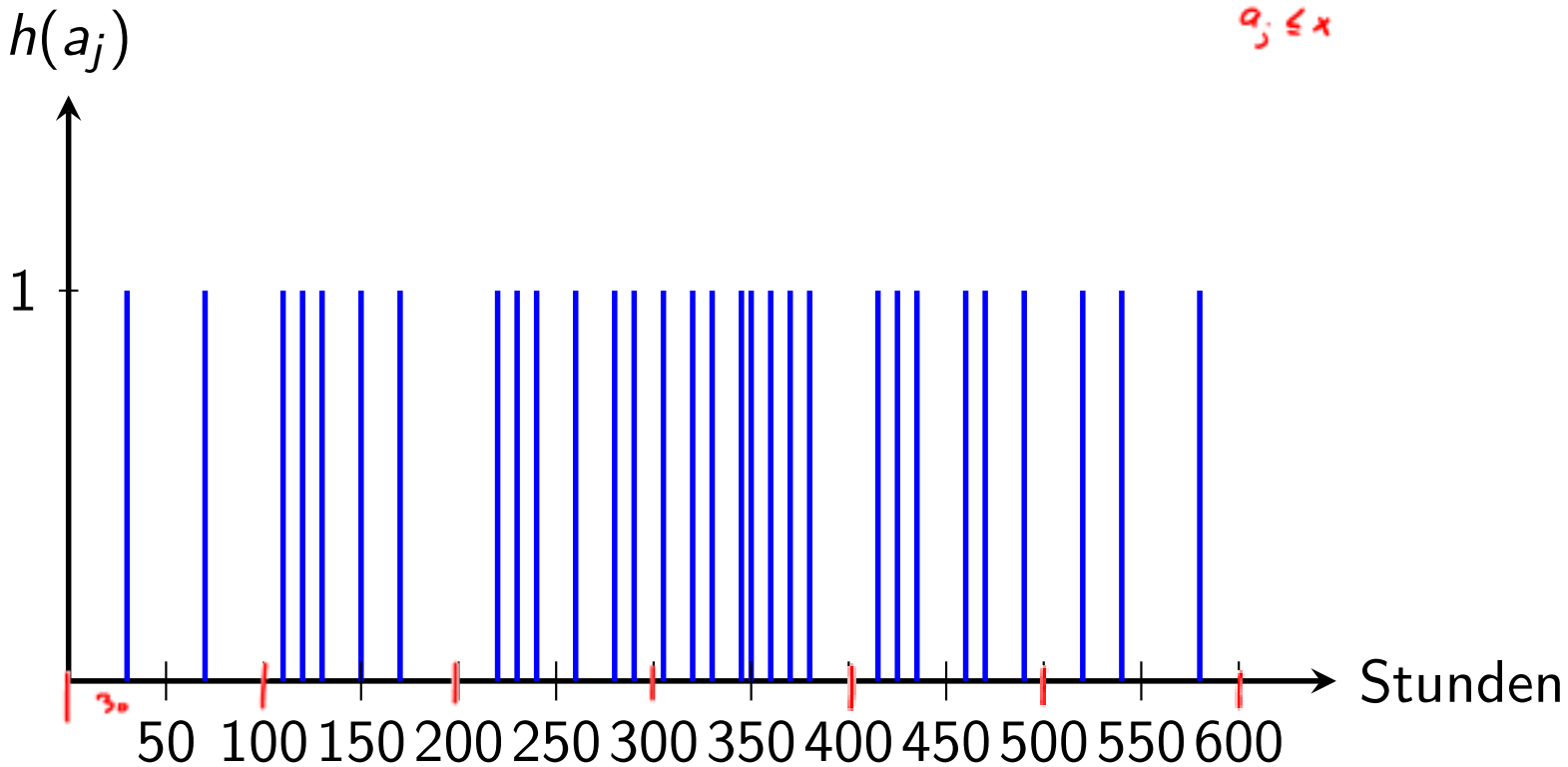
## Lebensdauer von Ventilen in Betriebsstunden

(Folie 23)

Bsp. kumulierte Häufigkeitsverteilung

$H(x) = \sum_{a_j \leq x} h(a_j)$

$H(130) = h(30) + h(70) + \cdots + h(120) = 7$



# Histogramm: Klassenbildung

Wir fassen jeweils mehrere Merkmalsausprägungen zu einer Klasse zusammen („Klassierung“).

Lege fest:

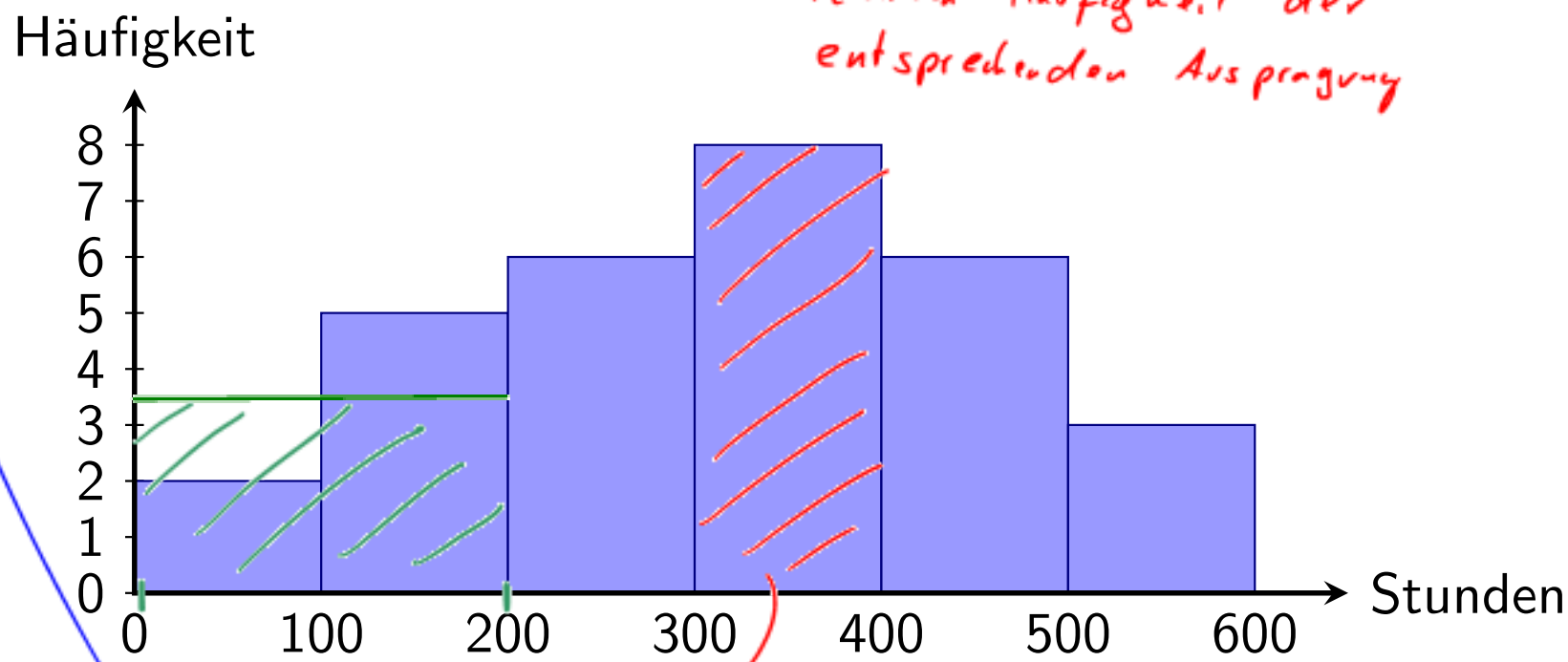
- die Anzahl der Klassen (hier: 6)
- deren jeweilige Breite (hier: 100)

| Klasse | Häufigkeit |
|---|---|
| 0 bis unter 100 | 2 |
| 100 bis unter 200 | 5 |
| 200 bis unter 300 | 6 |
| 300 bis unter 400 | 8 |
| 400 bis unter 500 | 6 |
| 500 bis unter 600 | 3 |

Wichtig:

- Zu grob: Informationsverlust
- Zu fein: wesentliche Charakteristika sind nicht gut sichtbar

# Histogramm: Lebensdauer in Stunden




700 : Anzahl der Beobachtungen · Prop.-Faktor (=100)

Die Fläche einer Säule ist proportional zu der relativen Häufigkeit der entsprechenden Ausprägung

Falls 1. & 2. Klasse zusammengefasst: [0-200]

h(0-200) = 7

(200 − 0) · Höhe = 700 ?

⇒ Höhe = 700/200 = 3,5

Fläche (400 − 300) · 8 = 800

h(300-400) = 8

Proportionalitätsfaktor = 100

# Kumulierte Häufigkeitsverteilung

$f(a_j) = \frac{h(a_j)}{n}$

**Absolute kumulierte Häufigkeitsverteilung**

reelle Zahl; $a_j$ z.B.: 0, 1, 2, 3, ...

$$H(x) = \sum_{a_j \leq x} h(a_j)$$

**Relative kumulierte Häufigkeitsverteilung**

$$F(x) = \sum_{a_j \leq x} f(a_j) = \sum_{a_j \leq x} \frac{h(a_j)}{n} = \frac{1}{n} \underbrace{\sum_{a_j \leq x} h(a_j)}_{= H(x)} = \frac{1}{n} H(x)$$

$F$ wird häufig als **empirische Verteilungsfunktion** bezeichnet.

$F(x)$: Anteil der Beobachtungen, die höchstens den Wert $x$ haben.

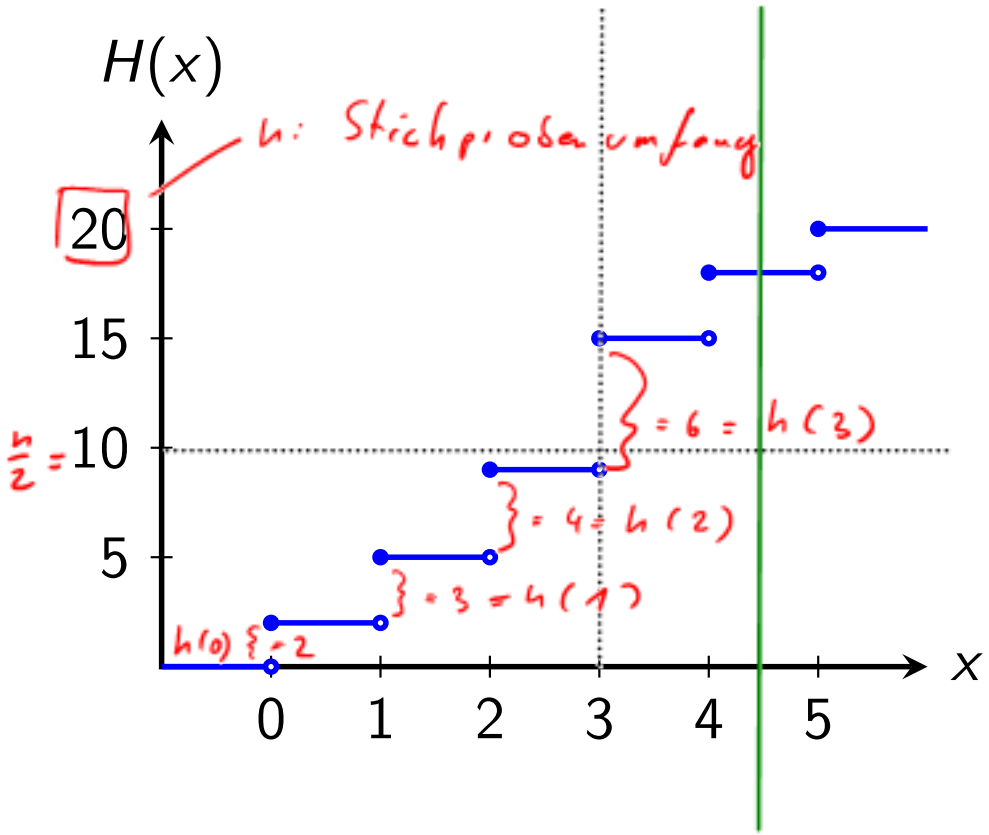
$H$ und $F$ sind monoton wachsende **Treppenfunktionen**.

# Abs. kum. Häufigkeitsverteilung: Krankmeldungen

| $a_j$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $h(a_j)$ | 2 | 3 | 4 | 6 | 3 | 2 |
| $H(a_j)$ | 2 | 5 | 9 | 15 | 18 | 20 |

z.B. $H(4) = 18$

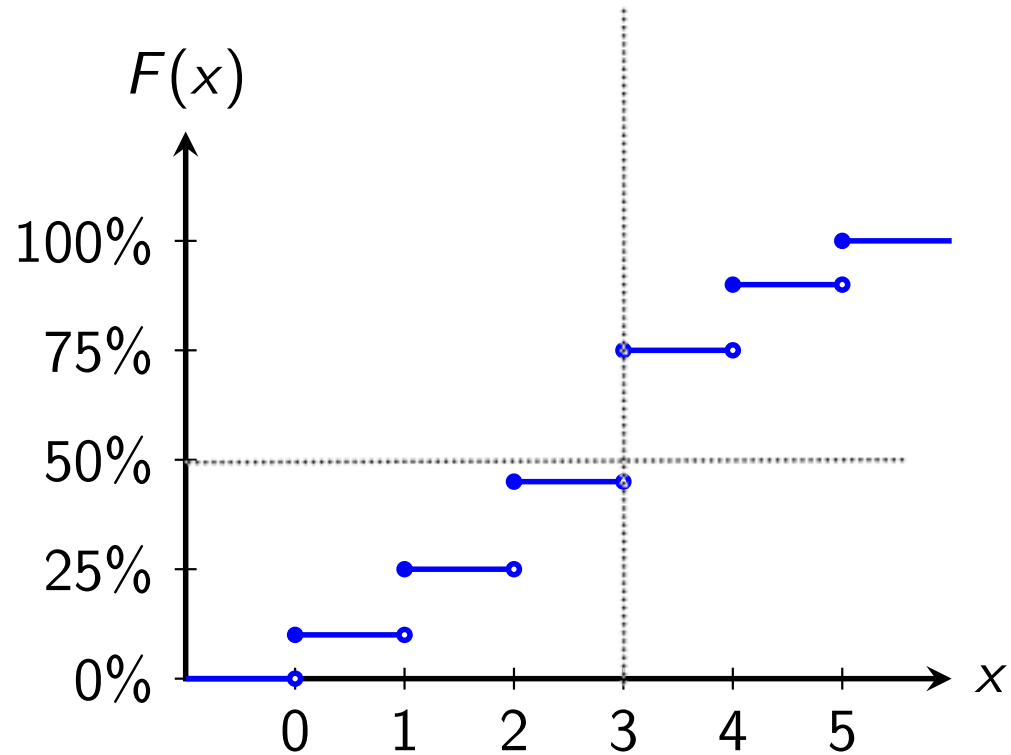
$$H(x) = \begin{cases} 0 & \text{für} \quad x < 0 \\ 2 & \text{für} \quad 0 \leq x < 1 \\ 5 & \text{für} \quad 1 \leq x < 2 \\ 9 & \text{für} \quad 2 \leq x < 3 \\ 15 & \text{für} \quad 3 \leq x < 4 \\ 18 & \text{für} \quad 4 \leq x < 5 \\ 20 & \text{für} \quad 5 \leq x \end{cases}$$



# Rel. kum. Häufigkeitsverteilung: Krankmeldungen

| $a_j$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $f(a_j)$ | .10 | .15 | .20 | .30 | .15 | .10 |
| $F(a_j)$ | .10 | .25 | .45 | .75 | .90 | 1 |

$$F(x) = \begin{cases} 0 & \text{für} \quad x < 0 \\ .10 & \text{für} \quad 0 \leq x < 1 \\ .25 & \text{für} \quad 1 \leq x < 2 \\ .45 & \text{für} \quad 2 \leq x < 3 \\ .75 & \text{für} \quad 3 \leq x < 4 \\ .90 & \text{für} \quad 4 \leq x < 5 \\ 1.0 & \text{für} \quad 5 \leq x \end{cases}$$



# Wachstumsfaktor und Wachstumsrate

Seien $x_1$ und $x_2$ zwei quantitative Beobachtungen.

**Wachstumsfaktor** (von $x_1$ zu $x_2$):

$$\frac{x_2}{x_1} = \frac{\text{neuer Wert}}{\text{alter Wert}}$$

**Wachstumsrate** (von $x_1$ zu $x_2$):

$$\frac{x_2 - x_1}{x_1} = \frac{\text{neuer Wert} - \text{alter Wert}}{\text{alter Wert}} = \frac{x_2}{x_1} - 1$$

**Bemerkung**:

Für positive Merkmale (Preise, Gewichte, Größen, etc.) müssen Wachstumsfaktoren positiv sein – Wachstumsraten aber nicht.

Die Wachstumsrate ist genau dann positiv, wenn der Wachstumsfaktor größer ist als 1.

# Lageparameter

- Modalwert
- Median
- Arithmetisches Mittel
- Geometrisches Mittel

# Lageparameter

**Ziel**

Beschreibung der Daten durch einen einzigen Wert.

Die Daten werden zu einer Kenngröße verdichtet, welche die Lage der Daten beschreibt.

# Modalwert

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit den Ausprägungen $a_1, \ldots, a_k$.

Dann ist der Modalwert $x_{Mod}$ jener Wert $a_j$, der am häufigsten in der Stichprobe auftritt.

Es gilt also:

$$h(x_{Mod}) \geq h(a_j) \text{ für alle } j = 1, \ldots, k$$

Der Modalwert muss nicht eindeutig sein, zum Beispiel kommt bei der Lebensdauer der Ventile jede Ausprägung gleich häufig vor.

# Der Median

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit ordinaler Skalierung mit den Ausprägungen $a_1 < \ldots < a_k$.

Für den Median $x_{Med}$ gilt:

Mindestens 50% aller Merkmalswerte sind kleiner oder gleich $x_{Med}$
und
mindestens 50% aller Merkmalswerte sind größer oder gleich $x_{Med}$.

# Berechnung des Medians

Es bezeichne $x_{(i)}$ den $i$-ten Wert der aufsteigend geordneten Daten. Es gilt also:

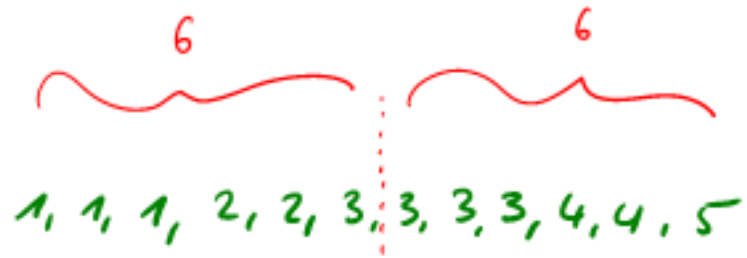
$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

**Falls $n$ ungerade:** $x_{Med} = x_{\left(\frac{n+1}{2}\right)}$.

**Falls $n$ gerade,** so erfüllt jeder Wert im Intervall $\left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right]$ die Median-Bedingung. In diesem Fall wird der Median oft als Intervallmitte, d.h. als $x_{Med} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$ festgesetzt.

# Beispiel Median: „Sterne“ von 12 Hotels

$x_{Mod} = 3$ ✓




Urliste:

3, 2, 1, 4, 3, 5, 4, 1, 1, 2, 3, 3

Für welche Zahl $x_{Med}$ zwischen 1 und 5 gilt, dass mindestens die Hälfte der Merkmalswerte kleiner oder gleich $x_{Med}$ und mindestens die Hälfte aller Mermalswerte größer oder gleich $x_{Med}$ sind?

$[\#\, x_i \text{ mit } x_i \leq 3] = 9 \geq 6$ ✓

$H(3)$

$[\#\, x_i \text{ mit } x_i \geq 3] = 7 \geq 6$ ✓

Häufigkeitstabelle

| $a_j$ | $a_1$ 1 | $a_2$ 2 | $a_3$ 3 | $a_4$ 4 | $a_5$ 5 |
|---|---|---|---|---|---|
| $h(a_j)$ | 3 | 2 | 4 | 2 | 1 |
| $H(a_j)$ | 3 | 3+2 = 5 | 3+2+4 = 9 | 3+2+4+2 = 11 | 3+2+4+2+1 = 12 |

$H(a_j) \geq 6 = \frac{n}{2}$

Falls $x_{Med} = a_j$ :

$H(a_j) \geq \frac{n}{2}$

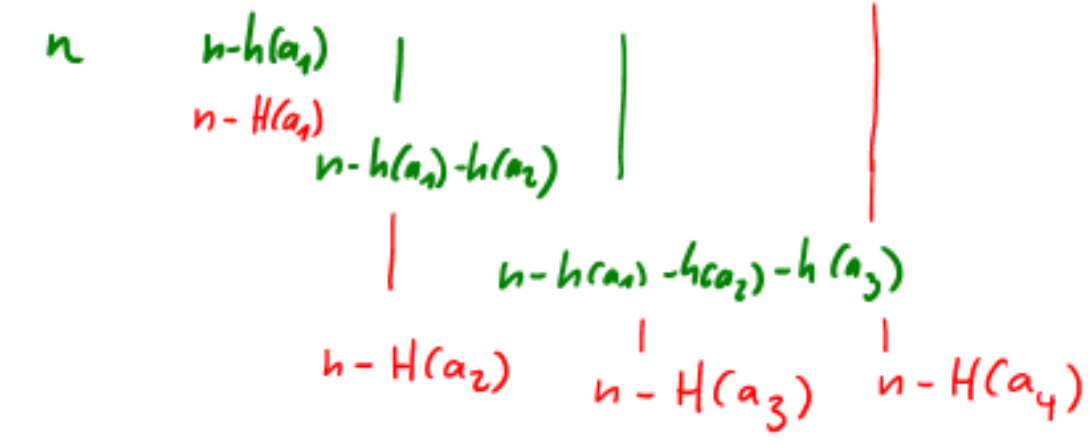
$H(a_{j-1}) \leq \frac{n}{2}$

12 — 12-3 — 12-3-2 — 12-3-2-4 — 12-3-2-4-2

12 — 9 — 7 — 3 — 1

$\#x_i$ mit $x_i \geq a_j$

$n - H(a_{j-1}) \geq \frac{n}{2} \Leftrightarrow n - \frac{n}{2} \geq H(a_{j-1})$



neue Stichprobe


Vorschlag 1 $x_{Med} = 2{,}5$

$\#\, x_i \text{ mit } x_i \leq 2{,}5 : 6 \geq \frac{n}{2}$ ✓

$\#\, x_i \text{ mit } x_i \geq 2{,}5 : 6 \geq \frac{n}{2}$ ✓

---

Vorschlag 2 $x_{Med} = 2$

$\#\, x_i \text{ mit } x_i \leq 2 : 6 \geq \frac{n}{2}$ ✓

$\#\, x_i \text{ mit } x_i \geq 2 : 9 \geq \frac{n}{2}$ ✓

Vorschlag 3 $x_{Med} = 3$

$\#\ x_i \text{ mit } x_i \leq 3 : \quad 9 \geq \frac{n}{2}$ ✓

$\#\ x_i \text{ mit } x_i \geq 3 : \quad 6 \geq \frac{n}{2}$ ✓

Alle Zahlen im Intervall $[2, 3]$ sind ein Median.

# Beispiel Median: „Sterne“ von 12 Hotels

Urliste:

$$3, 2, 1, 4, 3, 5, 4, 1, 1, 2, 3, 3$$

Geordnete Urliste:

$$\begin{array}{cccccccccccc} x_{(1)} & x_{(2)} & x_{(3)} & x_{(4)} & x_{(5)} & x_{(6)} & x_{(7)} & x_{(8)} & x_{(9)} & x_{(10)} & x_{(11)} & x_{(12)} \\ 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 5 \end{array}$$

$n = 12$ gerade, also ist jeder Wert im Intervall $\left[x_{(6)}, x_{(7)}\right] = \{3\}$ ein Median.

$$x_{Med} = 3$$

# Median: Interpretation

Falls $a_j$ ein Median ist:

„Mindestens 50% aller Merkmalswerte sind **kleiner oder gleich** $a_j$“

bedeutet: $H(a_j) \geq \frac{n}{2}$ bzw. $F(a_j) \geq \frac{1}{2}$. = 50%

„Mindestens 50% aller Merkmalswerte sind **größer oder gleich** $a_j$“

bedeutet: $n - H(a_j) + h(a_j) = n - H(a_{j-1}) \geq \frac{n}{2} \Leftrightarrow H(a_{j-1}) \leq \frac{n}{2}$
bzw. $F(a_{j-1}) \leq \frac{1}{2}$. = 50%

Die relative kumulierte Häufigkeitsverteilung „überspringt“beim Median den 50%-Wert.

# Arithmetisches Mittel

$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x^2}$

$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \overline{xy}$

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals mit kardinaler Skalierung.

**Arithmetisches Mittel** $\bar{x}$:

$x_1 + x_2 + x_3 + \cdots + x_n$

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

Es seien $a_j$ die gemessenen Ausprägungen und $f(a_j)$ die relative Häufigkeit von $a_j$ für $j = 1, \ldots, k$, dann gilt:

$$\bar{x} = \sum_{j=1}^{k} a_j f(a_j) = \sum_{j=1}^{k} a_j \frac{h(a_j)}{n}$$

($f(a_j)$)

$$= \frac{1}{n}\sum_{j=1}^{k} a_j h(a_j)$$

# Median & arithmetisches Mittel: Lineare Transformationen

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\overbrace{\sum_{i=1}^{n} a}^{=n\cdot a} + \frac{1}{n}\sum_{i=1}^{n} b x_i = \frac{1}{n}\cdot n a + b\,\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}} = a + b\bar{x} \quad \checkmark$$

Seien $X$ Merkmale mit kardinaler Skalierung.

Lineare Transformation (mit $a, b \in \mathbb{R}$):

$$y_i = a + bx_i$$

Dann gilt

$$\bar{y} = a + b\bar{x}$$

und

$$y_{Med} = a + bx_{Med}$$

Warum?

# Geometrisches Mittel

Seien $x_1, \ldots, x_n$, mit $x_i \geq 0$ für alle $i$ Ausprägungen eines Merkmals $X$ mit kardinaler Skalierung.

**Geometrisches Mittel** $x_{Geom}$:

$$x_{Geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

$$\Leftrightarrow \quad x_{Geom}^{n} = x_1 \cdot x_2 \cdot \cdots \cdot x_n$$

$$\underbrace{x_{Geom} \cdot x_{Geom} \cdot \ldots \cdot x_{Geom}}_{n\text{-mal}}$$

Typische Anwendung: Wachstumsfaktoren

# Lageparameter im Vergleich

| Skalierung | zu verwendender Lageparameter |
|---|---|
| nominal | Modalwert $x_{Mod}$ |
| ordinal | Median $x_{Med}$ |
| kardinal | arithm. Mittel $\bar{x}$ oder geom. Mittel $x_{Geom}$ |

Falls $x_1, \ldots, x_n \geq 0$, so gilt:

$$\bar{x} \geq x_{Geom} \ ,$$

wobei die beiden Ausdrücke gleich sind, wenn alle $x_i$ gleich sind.

# Optimalitätseigenschaften von $\bar{x}$ und $x_{Med}$

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals mit kardinaler Skalierung. Dann gilt:

- Das arithmetische Mittel minimiert die Summe der **quadrierten Abweichungen** der Daten $x_1, \ldots, x_n$ von einer Zahl:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - \lambda)^2 \text{ für alle } \lambda \in \mathbb{R}$$

"lambda"

- Der Median minimiert die Summe der **absoluten Abweichungen** der Daten $x_1, \ldots, x_n$ von einer Zahl:

$$\sum_{i=1}^{n}|x_i - x_{Med}| \leq \sum_{i=1}^{n}|x_i - \lambda| \text{ für alle } \lambda \in \mathbb{R}$$

Absolutwert

Warum berechnet man nicht

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

# Streuungsmaße

- Spannweite
- Durchschnittliche Abweichung
- Mittlere quadratische Abweichung
- Standardabweichung

# Streuungsmaße

**Ziel:**

Charakterisierung der Repräsentativität von Mittelwerten

Wichtige Ergänzung zu Lageparametern, da Daten mit gleichem Lageparameter beliebig unterschiedliche Streuung aufweisen können.

Nur für quantitative Merkmale mit kardinaler Skalierung definiert.

# Spannweite

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit kardinaler Skalierung.

**Spannweite**

$$SP = \max_i \; x_i - \min_i \; x_i = x_{(n)} - x_{(1)}$$

$SP$ ist das einfachste Streuungsmaß.

Sehr anfällig gegenüber Ausreißern.

# Durchschnittliche Abweichung

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit kardinaler Skalierung und Ausprägungen $a_1, \ldots, a_k$.

**Durchschnittliche Abweichung** (von einem Lageparameter $\lambda$):

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \lambda|$$

$\bar{s}$ ist für beliebige Lageparameter $\lambda$ definiert.

Wir wissen bereits: $\bar{s}$ ist für $\lambda = x_{Med}$ minimal.

Es gilt auch:

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \lambda| = \sum_{j=1}^{k} |a_j - \lambda| f(a_j)$$

# Mittlere quadratische Abweichung, Standardabweichung

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit kardinaler Skalierung und Ausprägungen $a_1, \ldots, a_k$.

Einheit von $X$
z.B. €

**Mittlere quadratische Abweichung**:

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Einheit von $s^2 = €^2$

Es gilt ebenfalls:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{j=1}^{k} (a_j - \bar{x})^2 f(a_j)$$

**Standardabweichung**: $s = \sqrt{s^2}$

Einheit von $s = €$

$s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ + $\frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i - \bar{x})^2}$

# Die mittlere quadratische Abweichung und lineare Transformationen

Sei $s_x^2$ für die Beobachtungen $x_1, \ldots, x_n$ definiert und sei

$$y_i = a + bx_i, \;\; i = 1, \ldots, n$$

Dann gilt:

$$s_y^2 = b^2 s_x^2$$

Für die Standardabweichungen gilt:

$$s_y = |b| s_x$$

$$s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$y_i = a + b x_i \quad \text{für} \quad i = 1, \ldots, n$$

$$\bar{y} = a + b \cdot \bar{x}$$

$$s_y^2 = \frac{1}{n} \sum_{i=1}^{n} \left( a + b x_i - (a + b\bar{x}) \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (a + b x_i - a - b\bar{x})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (b x_i - b\bar{x})^2 \equiv \frac{1}{n} \sum_{i=1}^{n} b^2 (x_i - \bar{x})^2$$

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

$s_x^2$

$$(b(x_i - \bar{x}))^2$$

$$= (b(x_i - \bar{x}))(b(x_i - \bar{x}))$$

$$= b(x_i - \bar{x}) \cdot b(x_i - \bar{x})$$

$$= b \cdot b \cdot (x_i - \bar{x})(x_i - \bar{x})$$

$$= b^2 (x_i - \bar{x})^2$$

$$s_y^2 = b^2 \cdot s_x^2$$

$$s_y = \sqrt{s_y^2} = \sqrt{b^2 \cdot s_x^2} = \sqrt{b^2} \cdot \sqrt{s_x^2} = \sqrt{b^2} \cdot s_x$$

$$= \begin{cases} b \cdot s_x, & b \geq 0 \\ -b \cdot s_x, & b < 0 \end{cases} = |b| \cdot s_x$$

$x_i$ : 0 0 1 3 16

$\bar{x} = \frac{1}{5}(0 + 0 + 1 + 3 + 16) = \frac{1}{5} \cdot 20 = 4$

$x_i - \bar{x}$ : -4 -4 -3 -1 12

$(x_i - \bar{x})^2$ : 16 16 9 1 144

$\sum_{i=1}^{5} (x_i - \bar{x})^2$ : $32 + 10 + 144 = 186 = 185 + 1$

$\frac{1}{5} \sum_{i=1}^{5} (x_i - \bar{x})^2 = s_x^2 = 37{,}2 = \frac{186}{5} = \frac{185}{5} + \frac{1}{5} = 37 + 0{,}2$

$s_x = \sqrt{37{,}2} = 6{,}099$ ⟸ (*)

$y_i = 16 + 1 \cdot x_i \rightsquigarrow s_y^2 = 1^2 \cdot s_x^2$ , $s_y = |1| \cdot s_x = s_x$

# Der Verschiebungssatz

Es gilt:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$$

$\overline{x^2} - \bar{x}^2$

und

$$\sum_{j=1}^{k}(a_j - \bar{x})^2 f(a_j) = \left(\sum_{j=1}^{k} a_j^2 f(a_j)\right) - \bar{x}^2$$

$\sum_{j=1}^{k} f(a_j) = 1$

$x_i$: 0 0 1 3 16

$x_i^2$: 0 0 1 9 256

$\sum_{i=1}^{n} x_i^2 = 266$ $\quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{5} 266 = 53,2$

$\overline{x^2} - \bar{x}^2 = 53,2 - 16 = 37,2$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{S_x^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$a^2 - 2ab + b^2$ ($a = x_i$, $b = \bar{x}$)

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i^2}_{\overline{x^2}} - \frac{1}{n}\sum_{i=1}^{n}2x_i\cdot\bar{x} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2}_{n\cdot\bar{x}^2}$$

$$= \overline{x^2} - 2\cdot\bar{x}\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i}_{\bar{x}} + \frac{1}{n}\cdot n\,\bar{x}^2$$

$$= \overline{x^2} - 2\cdot\bar{x}\cdot\bar{x} + \bar{x}^2$$

$$= \overline{x^2} - \bar{x}^2$$

# Konzentrationsmaße

- Variationskoeffizient
- Lorenzkurve
- Gini-Koeffizient
- Herfindahl-Index

# Konzentrationsmaße

Die behandelten Streuungsmaße geben uns keine Auskunft über Fragen der Ungleichheit oder Konzentration (von z.B. Vermögen).

Konzentration und Ungleichheit und der Effekt von Politik auf Ungleichheit ist eine zentrale Frage in vielen Bereichen.

# Maße für Konzentration und Ungleichheit

**Beispiel: Einkommensverteilungen**

| | | | | | | |
|---|---|---|---|---|---|---|
| X: | 0 | 0 | 1 | 3 | 16 | $\bar{x} = 4$ |
| Y: | 16 | 16 | 17 | 19 | 32 | $\bar{y} = 20$ |

Es gilt $s_x = s_y = 6,099$.

Die Standardabweichung ignoriert die Verschiedenartigkeit dieser beiden Stichproben.

Wie können wir diese vergleichen und feststellen welche gleicher ist und welche ungleicher?

# Variationskoeffizient

Seien $x_1, \ldots, x_n$ Merkmalswerte eines Merkmals $X$ mit kardinaler Skalierung, Standardabweichung $s$ und arithmetischem Mittel $\bar{x}$.

**Variationskoeffizient**:

$$V = \frac{s}{\bar{x}}$$

Der Variationskoeffizient setzt die Standardabweichung in Relation zum arithmetischen Mittel.

Beachte: $s$ und $\bar{x}$ werden in den gleichen Einheiten gemessen. Der Variationskoeffizient hat daher keine Einheiten und kann zwischen Stichproben verglichen werden.

# Beispiel: Einkommensverteilungen

| | | | | | | |
|---|---|---|---|---|---|---|
| X: | 0 | 0 | 1 | 3 | 16 | $\bar{x} = 4$ |
| Y: | 16 | 16 | 17 | 19 | 32 | $\bar{y} = 20$ |

Es gilt $s_x = s_y = 6,099$. ← siehe (*)

Es gilt $V \geq 0$ (falls $\bar{x} > 0$)

Variationskoeffizient von X:

$$V_x = \frac{6,099}{4} \approx 1,52$$

$V = 0 \iff s = 0 \iff x_i = \bar{x}$ für alle $i$

Variationskoeffizient von Y:

$$V_y = \frac{6,099}{20} \approx 0,30$$

Die beiden Variationskoeffizienten deuten darauf hin, dass die Ungleichheit in X größer ist als in Y.

Wir wollen dies aber genauer untersuchen!

# Beispiel: Einkommensverteilungen

$\frac{16+16+17}{100} = \frac{49}{100} = 0{,}49$

| i: | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| X: | 0 | 0 | 1 | 3 | 16 | $\sum_{i=1}^{5} x_i = 20$ |
| Y: | 16 | 16 | 17 | 19 | 32 | $\sum_{i=1}^{5} y_i = 100$ |

$\frac{0+0+1}{20} = \frac{1}{20} = 0{,}05$

16+16

| | in X | in Y |
|---|---|---|
| die 20% Ärmsten haben | 0% von 20 | 16% von 100 |
| die 40% Ärmsten haben | 0% von 20 | 32% von 100 |
| die 60% Ärmsten haben | 5% von 20 | 49% von 100 |
| die 80% Ärmsten haben | 20% von 20 | 68% von 100 |
| alle 100% haben | 100% von 20 | 100% von 100 |

→ x-Achse → y-Achse

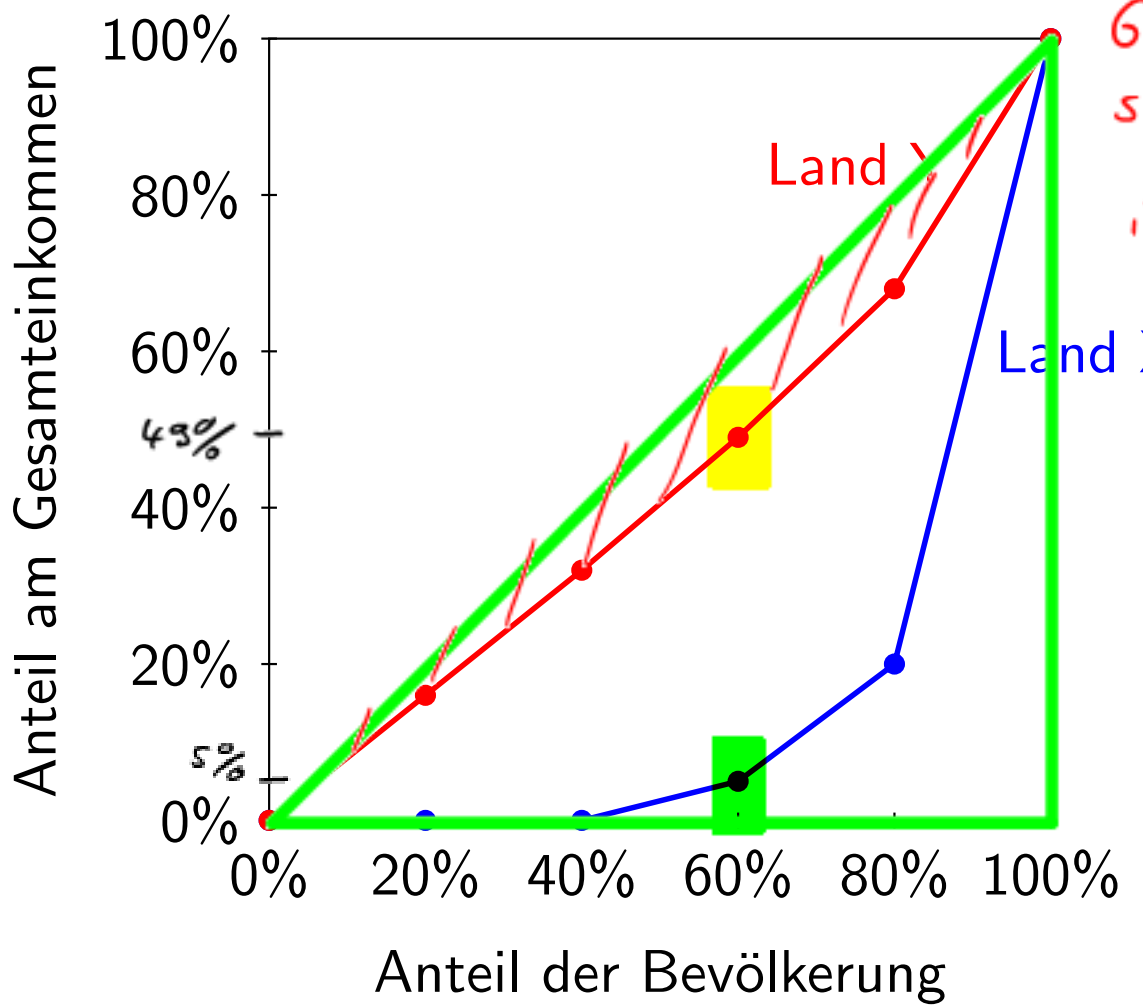
⇒ Lorenz Kurve

# Graphische Darstellung („Lorenzkurve“)

Fläche von △ = 1/2

Ginikoeffizient setzt

in das Verhältnis zu △



# Lorenzkurve

$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

Seien $x_{(1)}, \ldots, x_{(n)}$ der Größe nach geordnete Ausprägungen eines Merkmals $X$ mit kardinaler Skalierung und $\sum_{i=1}^{n} x_i > 0$.

Anteil an der gesamten Merkmalssumme, den die $k$ kleinsten Merkmalsträger haben:

$x_{(i)}$ : $i$.-kleinster Merkmalswert

Y-Koordinaten

$$v_k = \frac{\sum_{i=1}^{k} x_{(i)}}{\sum_{i=1}^{n} x_i}$$

Anteil, den die $k$ kleinsten Merkmalsträger haben:

X-Koordinaten

$$u_k = \frac{k}{n}$$

**Lorenzkurve**:

Polygonzug durch die Punkte

$$(0,0) = (u_0, v_0), (u_1, v_1), \ldots, (u_n, v_n) = (1,1)$$

# Eigenschaften der Lorenzkurve

a) Die Lorenzkurve verläuft durch die Punkte (0,0) und (1,1).

b) Die Lorenzkurve verläuft nirgends oberhalb der Diagonale.

c) Die Lorenzkurve ist monoton und konvex.

d) Die Lorenzkurve verändert sich nicht, wenn alle Merkmalsausprägungen mit dem gleichen Faktor ($\neq 0$) multipliziert werden.

e) Die Lorenzkurve stimmt mit der Diagonale genau dann überein, wenn alle Merkmalsausprägungen gleich sind.

f) Die Ungleichheit ist umso größer, je weiter die Lorenzkurve von der Diagonale entfernt ist.

# Der Gini-Koeffizient

Die Eigenschaft der Lorenzkurve

„Die Ungleichheit ist umso größer, je weiter die Lorenzkurve von der Diagonalen entfernt ist.“

möchten wir durch den Gini-Koeffizienten quantifizieren.

Je weiter die Lorenzkurve von der Diagonalen entfernt ist, desto größer ist die **Fläche zwischen der Lorenzkurve und der Diagonalen**.

# Gini-Koeffizient

Die Fläche des Quadrates

$$\text{Anteil an Beobachtungen} \times \text{Anteil an Merkmalssumme}$$

beträgt eins.

Die Fläche des Dreiecks unter der Diagonalen beträgt $\frac{1}{2}$.

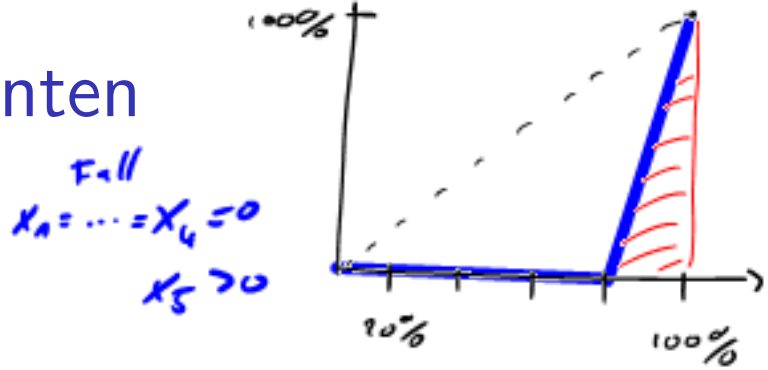
Wir setzen die Fläche zwischen der Lorenzkurve und der Diagonalen in Relation zur Fläche des Dreiecks.

**Gini-Koeffizient**:

$$G = \frac{\text{Fläche zwischen der Lorenzkurve und der Diagonalen}}{\text{Fläche unter der Diagonalen}}$$

$$= 2 \cdot \text{Fläche zwischen der Lorenzkurve und der Diagonalen}$$

# Eigenschaften des Gini-Koeffizienten




a) $0 \leq G \leq \frac{n-1}{n} < 1$:
Der Gini-Koeffizient liegt immer zwischen 0 und 1.

b) $G = 0 \Leftrightarrow$ die Lorenzkurve stimmt mit der Diagonalen überein.

c) $G = 0 \Leftrightarrow$ alle $x_i$ sind gleich.

d) $G = \frac{n-1}{n} \Leftrightarrow x_{(1)} = \ldots = x_{(n-1)} = 0 < x_{(n)}$.

e) Normierter Gini-Koeffizient: $G_* = \frac{n}{n-1} G$.

f) Werden alle Merkmalswerte $x_i$ mit einem konstanten Faktor $a$, $a \neq 0$, multipliziert, d. h. $y_i = ax_i$, $i = 1, \ldots, n$, ändert sich der Gini-Koeffizient nicht, also $G_y = G_x$.

# Berechnung des Gini-Koeffizienten

**Berechnung I**

$$G = 2\frac{\sum_{i=1}^{n} i \cdot x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}$$

**Berechnung II**

Es seien $(u_i, v_i)$, $i = 0, 1, \ldots, n$ die Stützpunkte der Lorenzkurve mit:

$$u_k = \frac{k}{n} \text{ und } v_k = \frac{\sum_{i=1}^{k} x_{(i)}}{\sum_{i=1}^{n} x_i}$$

Dann gilt:

$$G = \sum_{i=2}^{n} u_{i-1}v_i - u_i v_{i-1}$$

# Gini-Koeffizient für X: 0 0 1 3 16

$$
\begin{aligned}
G_x &= 2\frac{\sum_{i=1}^{n} i \cdot x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n} \\
&= 2\frac{1\cdot 0 + 2\cdot 0 + 3\cdot 1 + 4\cdot 3 + 5\cdot 16}{5(0+0+1+3+16)} - \frac{6}{5} \\
&= 2\frac{95}{100} - \frac{120}{100} = \frac{190}{100} - \frac{120}{100} = \frac{70}{100} = 0,7
\end{aligned}
$$

# Gini-Koeffizient für X: 0 0 1 3 16

$$G_x = \sum_{i=2}^{n} u_{i-1} v_i - u_i v_{i-1}$$

| $i$ | $u_i$ | $v_i$ | $u_{i-1} v_i$ | $u_i v_{i-1}$ |
|---|---|---|---|---|
| 0 | 0 | 0 | – | – |
| 1 | 0,2 | 0 | 0 | 0 |
| 2 | 0,4 | 0 | 0 | 0 |
| 3 | 0,6 | 0,05 | 0,02 | 0 |
| 4 | 0,8 | 0,2 | 0,12 | 0,04 |
| 5 | 1 | 1 | 0,8 | 0,2 |
| $\sum$ | | | 0,94 | 0,24 |

$$G_x = 0,94 - 0,24 = 0,7$$

Übung: berechne $G_y = 0,14$ für Y: 16 16 17 19 32!

# Kritikpunkte bzgl. des Gini-Koeffizienten

Unterschiedliche Lorenzkurven können zum gleichen Wert des Gini-Koeffizienten führen.

Der Gini-Koeffizient ist nur ein Maß für die relative Ungleichheit und nicht für die absolute:

- $G = 0$ bei 2 Firmen mit je 50% Marktanteil
- $G = 0$ bei 20 Firmen mit je 5% Marktanteil

Dies sind offenbar sehr unterschiedlich strukturierte Märkte.

# Herfindahl-Index

Seien $x_1, \ldots, x_n$ Ausprägungen eines Merkmals $X$ mit kardinaler Skalierung.

Definiere relative Merkmalswerte:

$$p_i = \frac{x_i}{\sum_{j=1}^{n} x_j} \text{ für } i = 1, \ldots, n$$

**Herfindahl-Index**:

$$H = \sum_{i=1}^{n} p_i^2$$

Es gilt:

- $\frac{1}{n} \leq H \leq 1$
- $H = \frac{1}{n} \Leftrightarrow$ alle $x_i$ sind gleich.
- $H = 1 \Leftrightarrow x_{(1)} = \ldots = x_{(n-1)} = 0 < x_{(n)}$

# Kapitel 4

# Mehrdimensionale Daten

# Kapitel 4: Mehrdimensionale Daten

**Mehrdimensional:** mehr als ein Merkmal

Wir betrachten hier meist zwei oder drei Merkmale, die Methodik lässt sich aber allgemein auf $m \geq 2$ Merkmale anwenden.

**Mehrdimensionale Urliste:**

Für jeden der $n$ Merkmalsträger werden $m$ Merkmale gemessen. Insgesamt werden also $n \cdot m$ Daten erhoben.

Die mehrdimensionale Urliste ist dann eine Tabelle mit $n$ Zeilen und $m$ Spalten.

Wir interessieren uns für **Abhängigkeiten** zwischen verschiedenen Merkmalen.

# Kontingenztabelle und Streuungsdiagramm

Hier: zwei Merkmale $X$ und $Y$ bei $n$ Beobachtungen

Urliste:

$$\begin{pmatrix} x_1, x_2, \ldots, x_n \\ y_1, y_2, \ldots, y_n \end{pmatrix}$$

Beobachtung $i$: $(x_i, y_i)$ für $i = 1, \ldots, n$

Zwei Darstellungsformen:

- **Kontingenztabelle** (falls viele der $n$ Wertepaare gleich sind und bei nominalskalierten Merkmalen)
- **Streuungsdiagramm** (falls viele der $n$ Wertepaare von kardinalskalierten Merkmalen unterschiedlich sind)

# Kontingenztabelle

Ausprägungen von Merkmal $X$: $a_1, \ldots, a_k$

Ausprägungen von Merkmal $Y$: $b_1, \ldots, b_l$

Häufigkeit von Beobachtungen $(x, y)$ mit $x = a_i$ und $y = b_j$:

$$h_{ij} = h(a_i, b_j)$$

| | $b_1$ | $b_2$ | $\ldots$ | $b_l$ |
|---|---|---|---|---|
| $a_1$ | $h_{11}$ | $h_{12}$ | $\ldots$ | $h_{1l}$ |
| $a_2$ | $h_{21}$ | $h_{22}$ | $\ldots$ | $h_{2l}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $a_k$ | $h_{k1}$ | $h_{k2}$ | $\ldots$ | $h_{kl}$ |

$h_{ij}$ wird auch als **gemeinsame Häufigkeit** bezeichnet.

# Beispiel: Konsum von Cannabis

Daten in Anlehnung an Moor *et al* (2020) Tabelle 1[1]

$\frac{741}{877} \cdot \frac{1}{n}$

Kontingenztabelle: Cannabiskonsum von 15-jährigen

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 741 | 497 | 1238 |
| Leben | 64 | 81 | 145 |
| Monat | 72 | 64 | 136 |
| $\sum$ | 877 | 642 | 1519 |

Stichprobengröße $n$

„Leben“: Konsum mindestens einmal im Leben, aber nicht innerhalb der letzten 30 Tage.

„Monat“: Konsum mindestens einmal innerhalb der letzten 30 Tage.

[1] Moor *et al* „Alkohol, Tabak- und Cannabiskonsum im Jugendalter – Querschnittergebnisse der HBSC-Studie 2017/18“ Journal of Health Monitoring (2020) Vol 5 (3)

# Randhäufigkeiten, Randverteilung

#Merkmalsträger mit erstem Merkmal $a_i$: $h_{i\bullet} = \sum_{j=1}^{l} h_{ij}$

#Merkmalsträger mit zweitem Merkmal $b_j$: $h_{\bullet j} = \sum_{i=1}^{k} h_{ij}$

| | $b_1$ | $b_2$ | $\dots$ | $b_l$ | $\sum$ |
|---|---|---|---|---|---|
| $a_1$ | $h_{11}$ | $h_{12}$ | $\dots$ | $h_{1l}$ | $h_{1\bullet}$ |
| $a_2$ | $h_{21}$ | $h_{22}$ | $\dots$ | $h_{2l}$ | $h_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $a_k$ | $h_{k1}$ | $h_{k2}$ | $\dots$ | $h_{kl}$ | $h_{k\bullet}$ |
| $\sum$ | $h_{\bullet 1}$ | $h_{\bullet 2}$ | $\dots$ | $h_{\bullet l}$ | $n$ |

Die **Randverteilung** eines Merkmals ergibt sich durch alle Randhäufigkeiten dieses Merkmals.

# Relative Häufigkeiten

Wir erhalten die relativen Häufigkeiten, indem wir alle Zahlen der Tabelle der absoluten Häufigkeiten durch $n$ teilen:

| | $b_1$ | $b_2$ | $\dots$ | $b_l$ | $\sum$ |
|---|---|---|---|---|---|
| $a_1$ | $f_{11}$ | $f_{12}$ | $\dots$ | $f_{1l}$ | $f_{1\bullet}$ |
| $a_2$ | $f_{21}$ | $f_{22}$ | $\dots$ | $f_{2l}$ | $f_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $a_k$ | $f_{k1}$ | $f_{k2}$ | $\dots$ | $f_{kl}$ | $f_{k\bullet}$ |
| $\sum$ | $f_{\bullet 1}$ | $f_{\bullet 2}$ | $\dots$ | $f_{\bullet l}$ | $1$ |

mit $f_{..} = h_{..}/n$

# Relative Häufigkeiten: Konsum von Cannabis

Kontingenztabelle: Cannabiskonsum von 15-jährigen
Gerundete relative Häufigkeiten

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 49% | 33% | 82% |
| Leben | 4% | 5% | 9% |
| Monat | 5% | 4% | 9% |
| $\sum$ | 58% | 42% | 100% |

Interpretationen:

- 58% der Merkmalsträger sind weiblich.
- 4% der Merkmalsträger sind männlich und konsumierten mindestens einmal Cannabis während der letzten 30 Tage.
- 82% der Merkmalsträger haben noch nie Cannabis konsumiert.
- ...

# Bedingte relative Häufigkeiten

**Bedingte relative Häufigkeit von $a_i$ gegeben $b_j$** $f_1(a_i|b_j)$:

Relative Häufigkeit, mit der die Ausprägung $a_i$ bei denjenigen Merkmalsträgern auftritt, die bzgl. des zweiten Merkmals die Ausprägung $b_j$ aufweisen.

Cannabis Konsum → $f_1(\text{noch nie} \mid ♀) = \frac{741}{872}$

$$f_1(a_i|b_j) = \frac{f_{ij}}{f_{\bullet j}} = \frac{\frac{h_{ij}}{n}}{\frac{h_{\bullet j}}{n}} = \frac{h_{ij}}{h_{\bullet j}}$$

**Bedingte relative Häufigkeit von $b_j$ gegeben $a_i$** $f_2(b_j|a_i)$:

Relative Häufigkeit, mit der die Ausprägung $b_j$ bei denjenigen Merkmalsträgern auftritt, die bzgl. des zweiten Merkmals die Ausprägung $a_i$ aufweisen.

$$f_2(b_j|a_i) = \ldots = \frac{h_{ij}}{h_{i\bullet}}$$

Geschlecht

# Bedingte relative Häufigkeiten: Konsum von Cannabis

Kontingenztabelle: Cannabiskonsum von 15-jährigen

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 741 | 497 | 1238 |
| Leben | 64 | 81 | 145 |
| Monat | 72 | 64 | 136 |
| $\sum$ | 877 | 642 | 1519 |

Frage: Wie viel Prozent der Frauen haben noch nie Cannabis konsumiert?

Wir setzen die Anzahl der Frauen, die noch nie Cannabis konsumiert haben, in Relation zu der Anzahl der Frauen:

$$f_1(\text{noch nie}|♀) = \frac{741}{877} = 84\%$$

# Bedingte Verteilungen

**Bedingte Verteilung des ersten Merkmals**
bei gegebener Ausprägung $b_j$ des zweiten Merkmals:

$$f_1(a_1|b_j),\ f_1(a_2|b_j),\ \ldots\ ,\ f_1(a_k|b_j)$$

**Bedingte Verteilung des zweiten Merkmals**
bei gegebener Ausprägung $a_i$ des ersten Merkmals:

$$f_2(b_1|a_i),\ f_2(b_2|a_i),\ \ldots\ ,\ f_2(b_l|a_i)$$

# Bedingte Verteilungen: Konsum von Cannabis

$\frac{145}{1519} = 9{,}545\,\%$

Kontingenztabelle: Absolute Häufigkeiten

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 741 | 497 | 1238 |
| Leben | 64 | 81 | 145 |
| Monat | 72 | 64 | 136 |
| $\sum$ | 877 | 642 | 1519 |

Bedingte Verteilung des Cannabis-Konsums, bedingt auf ♀/ ♂:

$7\% = \frac{64}{877}$

$8\% = \frac{72}{877}$

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 84% | 77% | 82% |
| Leben | 7% | 13% | 10% |
| Monat | 8% | 10% | 9% |
| $\sum$ | 100% | 100% | 100% |

$13\% = \frac{81}{642}$

$9{,}545\,\%$

# Bedingte Verteilungen: Konsum von Cannabis

Kontingenztabelle: Absolute Häufigkeiten

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 741 | 497 | 1238 |
| Leben | 64 | 81 | 145 |
| Monat | 72 | 64 | 136 |
| $\sum$ | 877 | 642 | 1519 |

Bedingte Verteilung des Geschlechts, bedingt auf Cannabis-Konsum:

$60\% = \frac{741}{1238}$

| | ♀ | ♂ | $\sum$ |
|---|---|---|---|
| noch nie | 60% | 40% | 100% |
| Leben | 44% | 56% | 100% |
| Monat | 53% | 47% | 100% |
| $\sum$ | 58% | 42% | 100% |

# Unabhängigkeit: „zwei Merkmale beeinflussen sich nicht“

Eine präzisere Definition folgt in Teil 2 W'keitsrechnung.

Bei zwei unabhängigen Merkmalen sollte die Bedingung keinen Einfluss auf die bedingten Verteilungen haben. Also sollten die bedingten Verteilungen für alle Bedingungen gleich der (unbedingten) Randverteilung sein:

$$\frac{h_{ij}}{h_{\bullet j}} = \frac{h_{i\bullet}}{n} \quad | \cdot h_{\bullet j}$$

$$\left.\begin{array}{rrcl} \text{bedingte Verteilung} & f_1(a_i|b_j) & = & \frac{h_{ij}}{h_{\bullet j}} \\ \text{Randverteilung} & f_1(a_i) & = & \frac{h_{i\bullet}}{n} \end{array}\right\} \Rightarrow h_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$$

Sind Geschlecht und Cannabis-Konsum Deiner Meinung nach Unabhängig?

# Streuungsdiagramm

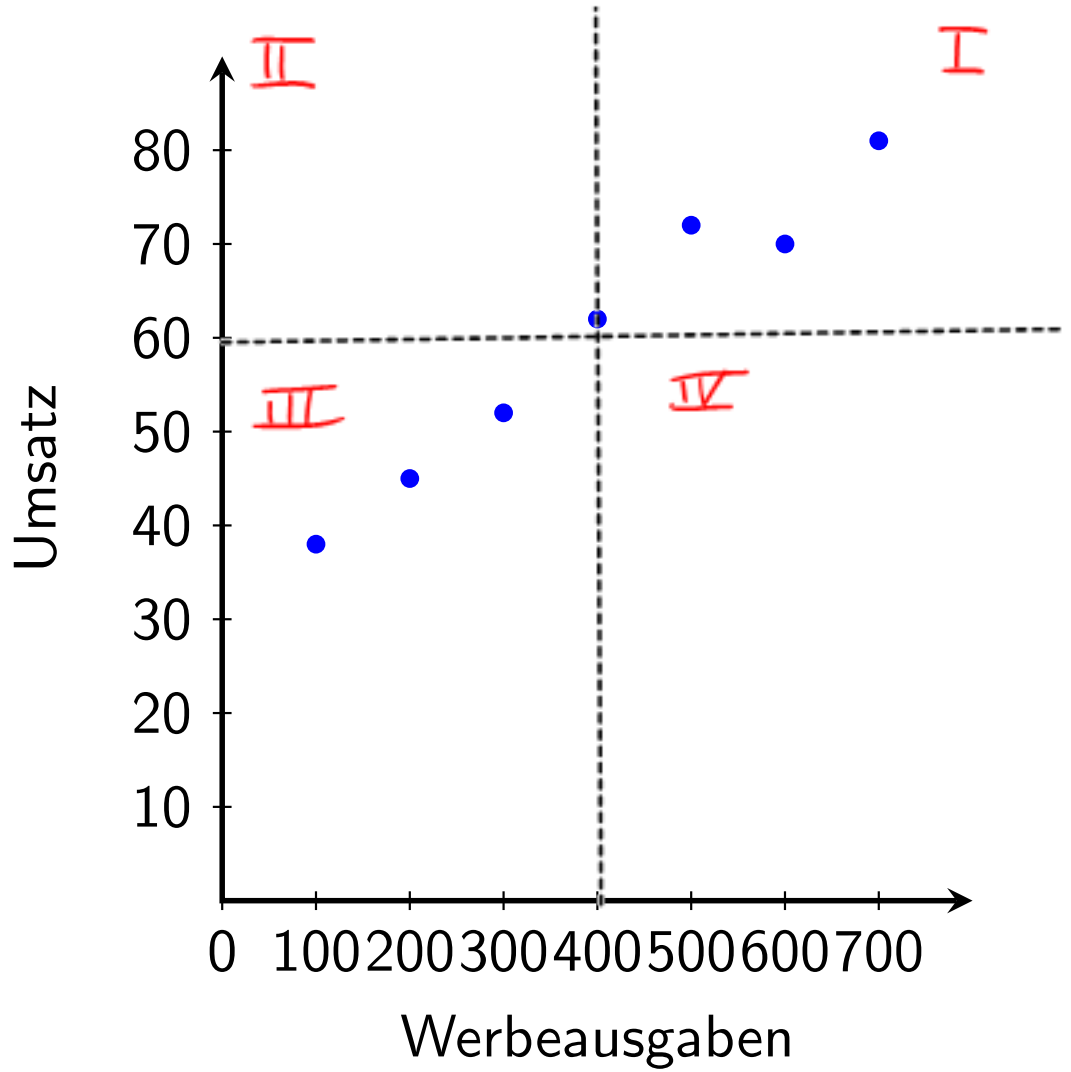
## Beispiel: Werbeausgaben und Umsatz

| Firma $i$ | Werbeausgaben $x_i$ (1.000) | Umsatz $y_i$ (Mio.) |
|---|---|---|
| 1 | 100 | 38 |
| 2 | 200 | 45 |
| 3 | 300 | 52 |
| 4 | 400 | 62 |
| 5 | 500 | 72 |
| 6 | 600 | 70 |
| 7 | 700 | 81 |

# Streuungsdiagramm

## Beispiel: Werbeausgaben und Umsatz

Eintragung der Beobachtungspaare in ein Koordinatensystem:

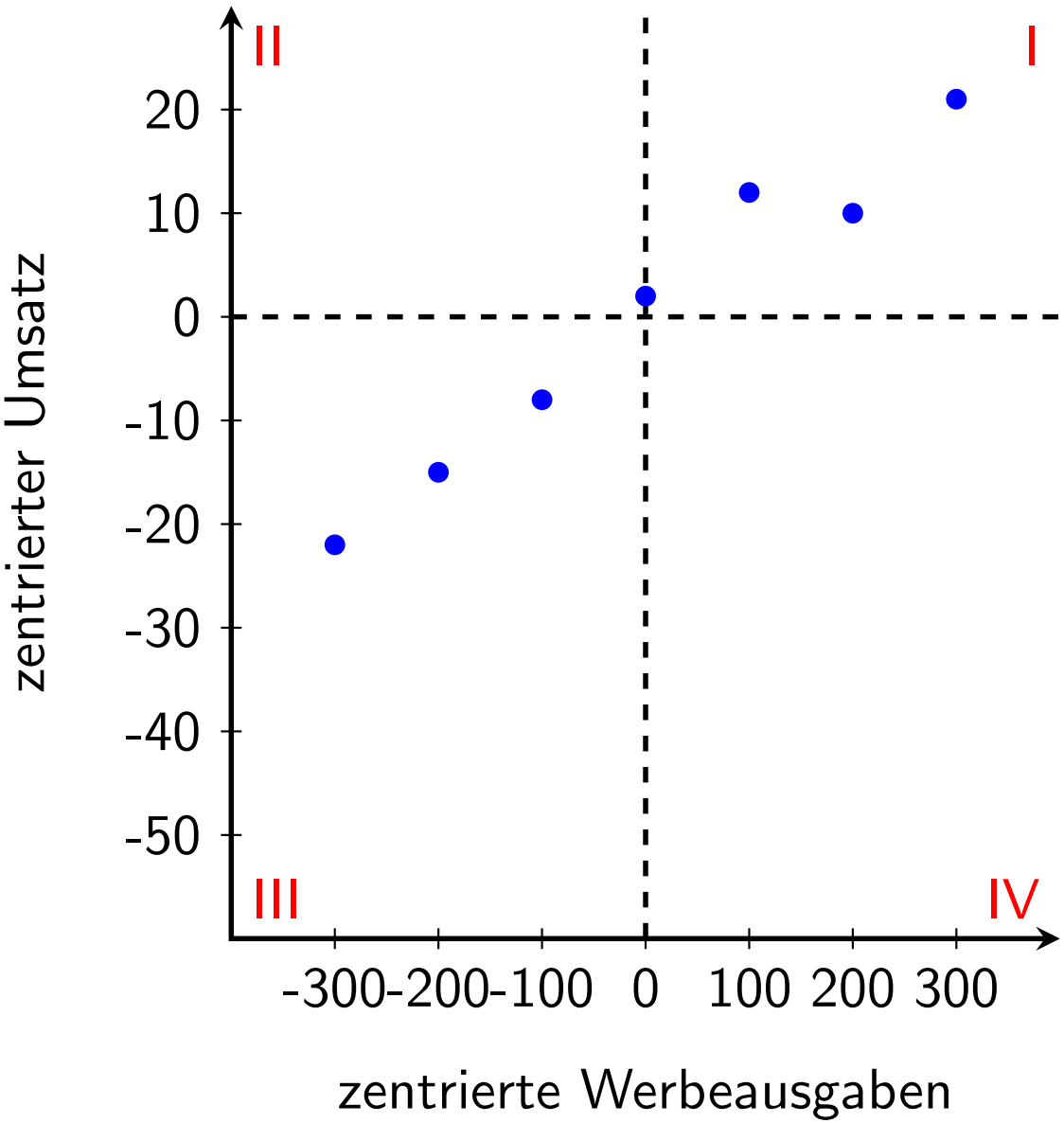


# Zentrierung der Daten

Für jede Beobachtung eines Merkmals wird dessen Differenz zum arithmetischen Mittel des Merkmals berechnet:

| Firma $i$ | Werbeausgaben $x_i$ (1.000) | Umsatz $y_i$ (Mio.) | $x_i - \bar{x}$ | $y_i - \bar{y}$ |
|---|---|---|---|---|
| 1 | 100 | 38 | -300 | -22 |
| 2 | 200 | 45 | -200 | -15 |
| 3 | 300 | 52 | -100 | -8 |
| 4 | 400 | 62 | 0 | 2 |
| 5 | 500 | 72 | 100 | 12 |
| 6 | 600 | 70 | 200 | 10 |
| 7 | 700 | 81 | 300 | 21 |
| arith. Mittel | 400 | 60 | 0 | 0 |

# Streudiagramm der zentrierten Daten



# Geometrischer Zugang zum Zusammenhang

Viele Beobachtungen in I und III: starker positiver Zusammenhang

Viele Beobachtungen in II und IV: starker negativer Zusammenhang

Beobachtungen auf einem Rand werden nicht mitgezählt:

$$\begin{aligned} I + III : 3 + 3 &= 6 \\ II + IV : 0 + 0 &= 0 \end{aligned}$$

Deutliche Evidenz für positiven Zusammenhang: aber sehr grobes Maß – berücksichtigt nicht die „Lage“ der Daten (relativ zum arithmetischen Mittel).

# Korrelationsrechnung

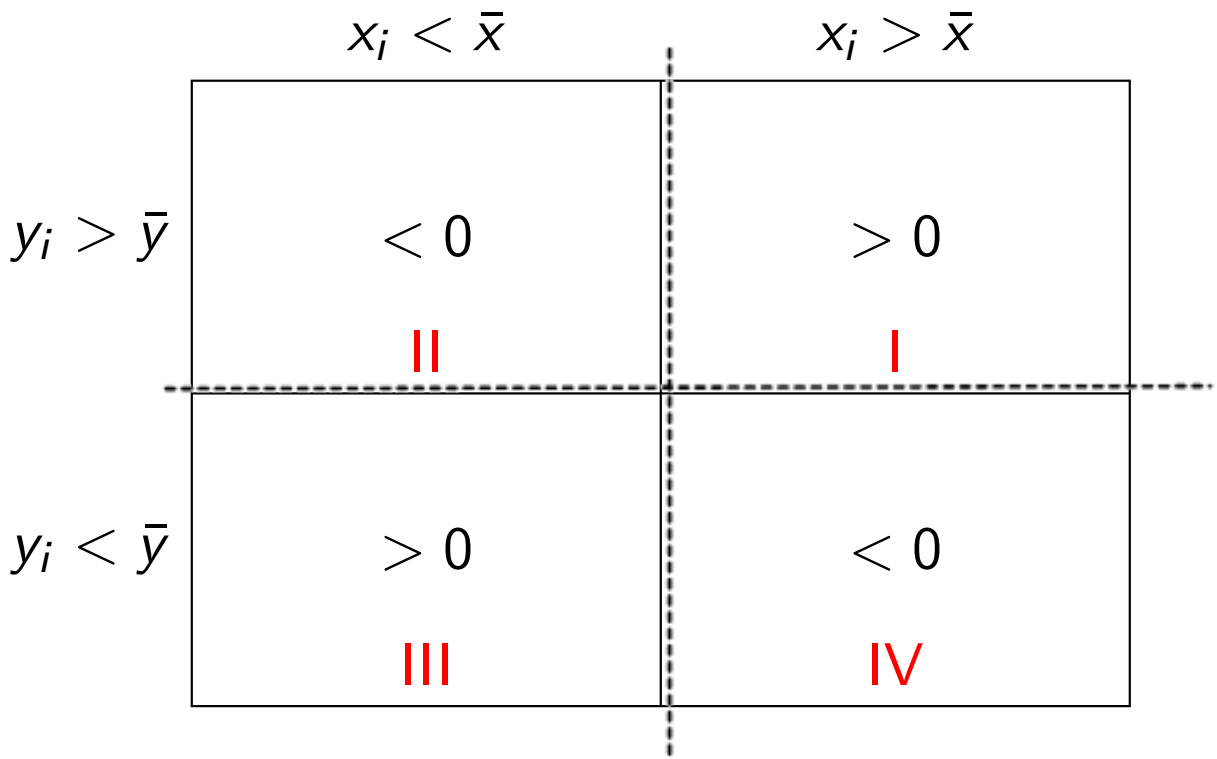
In der Korrelationsrechnung bestimmen wir **Maßzahlen**, welche die **Stärke** und bei ordinal und kardinal skalierten Merkmalen auch die **Richtung des Zusammenhangs** erfassen.

<table>
<tr><th>Skalierung Y<br><br>Skalierung X</th><th>kardinal</th><th>ordinal</th><th>nominal</th></tr>
<tr><td>kardinal</td><td>Bravais-Pearson-Korrelations-koeffizient</td><td rowspan="2">Rangkorrelations-koeffizient von Spearman</td><td rowspan="2"></td></tr>
<tr><td>ordinal</td></tr>
<tr><td>nominal</td><td colspan="2"></td><td>Kontingenz-koeffizient</td></tr>
</table>

# Korrelation bei kardinal skalierten Merkmalen

Im vorigen Beispiel wurden Datenpunkte, welche vollständig in einem Quadranten liegen mit 1 gewichtet.

Stattdessen könnten die Datenpunkte auch entsprechend ihrer „Abweichung zum Mittelwert“ gewichtet werden: $(x_i - \bar{x})(y_i - \bar{y})$



# Kovarianz

Seien $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ Ausprägungen zweier Merkmale $X$ und $Y$ mit kardinaler Skalierung. Dann ist die **(empirische) Kovarianz** von $X$ und $Y$ definiert als:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Bemerkung: Im Fall von $x_i = y_i$ ergibt sich als „Kovarianz von $X$ mit sich selbst“ die schon definierte mittlere quadratische Abweichung.

Mittlere quadratische Abweichung — Steiner:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \overline{xx} - \bar{x}\cdot\bar{x} \quad \text{für } x$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y}) = \overline{yy} - \bar{y}\cdot\bar{y} \quad \text{für } y$$

empirische Kovarianz

kein Quadrat →

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y}$$

$$\overline{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot x_i$$

Korrelationskoeffizient:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \bar{x}\cdot\bar{x} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

$$\rho_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\cdot\sqrt{s_y^2}}, \quad -1 \le \rho_{xy} \le 1$$

# Eigenschaften der Kovarianz

Für die empirische Kovarianz $s_{xy}$ gelten folgende Eigenschaften:

a) $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}$

b) Wenn $z_i = a + b \cdot y_i$ gilt (mit $a, b \in \mathbb{R}$), dann gilt $s_{xz} = b \cdot s_{xy}$

c) Die Kovarianz ist beschränkt durch das Produkt der Standardabweichungen:

$$-s_x \cdot s_y \leq s_{xy} \leq s_x \cdot s_y$$

d) Für $y_i = a + b \cdot x_i$:

$$s_{xy} = \begin{cases} s_x \cdot s_y & \text{falls } b > 0 \\ -s_x \cdot s_y & \text{falls } b < 0 \end{cases}$$

$$z_i = a + b \cdot y_i \quad \Rightarrow \quad \bar{z} = a + b \cdot \bar{y}$$

$$s_{xz} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (z_i - \bar{z})$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (a + b y_i - a - b \bar{y})$$

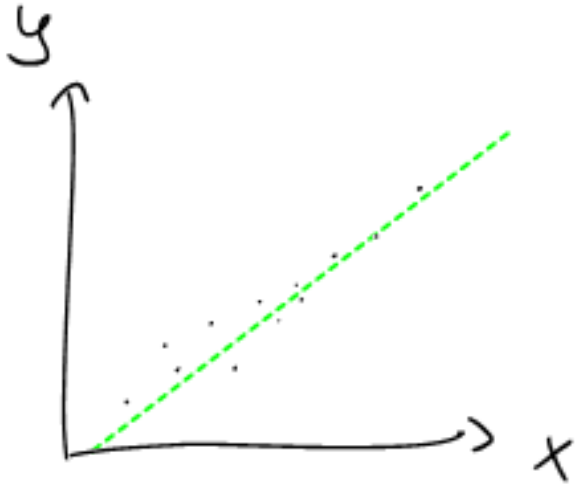
$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (b y_i - b \bar{y})$$

$$= b \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = b \cdot s_{xy}$$

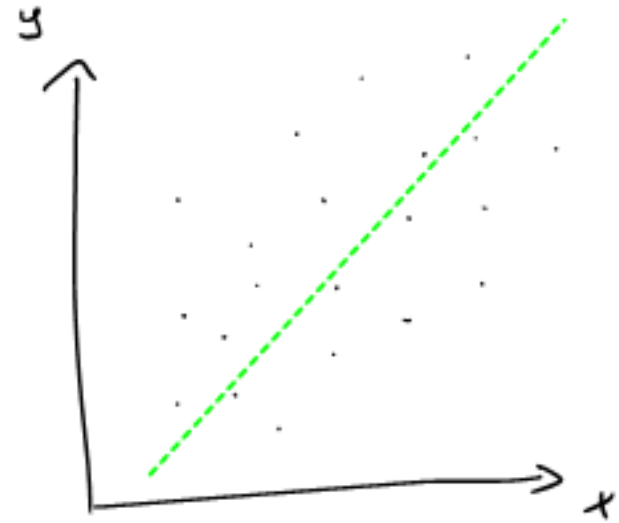
$$s_z^2 = b^2 \cdot s_y^2$$

$$\sqrt{s_z^2} = |b| \cdot s_y$$

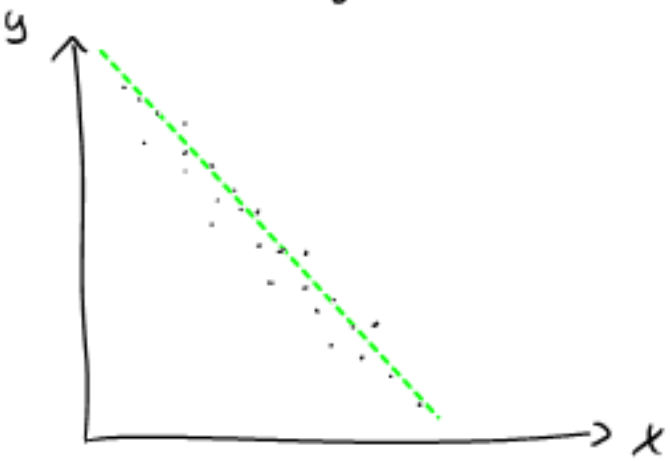
Die empirische Kovarianz misst den linearen Zusammenhang zwischen zwei Merkmalen.

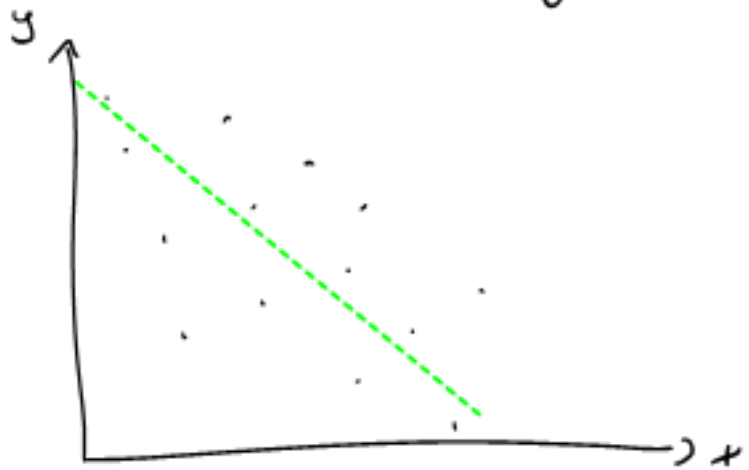


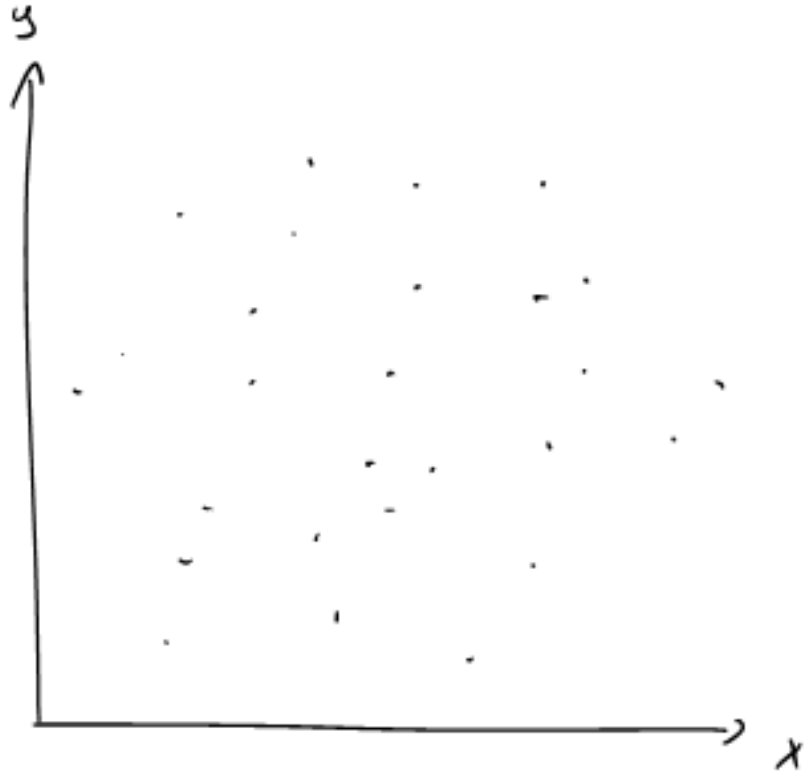

starker linearer Zusammenhang



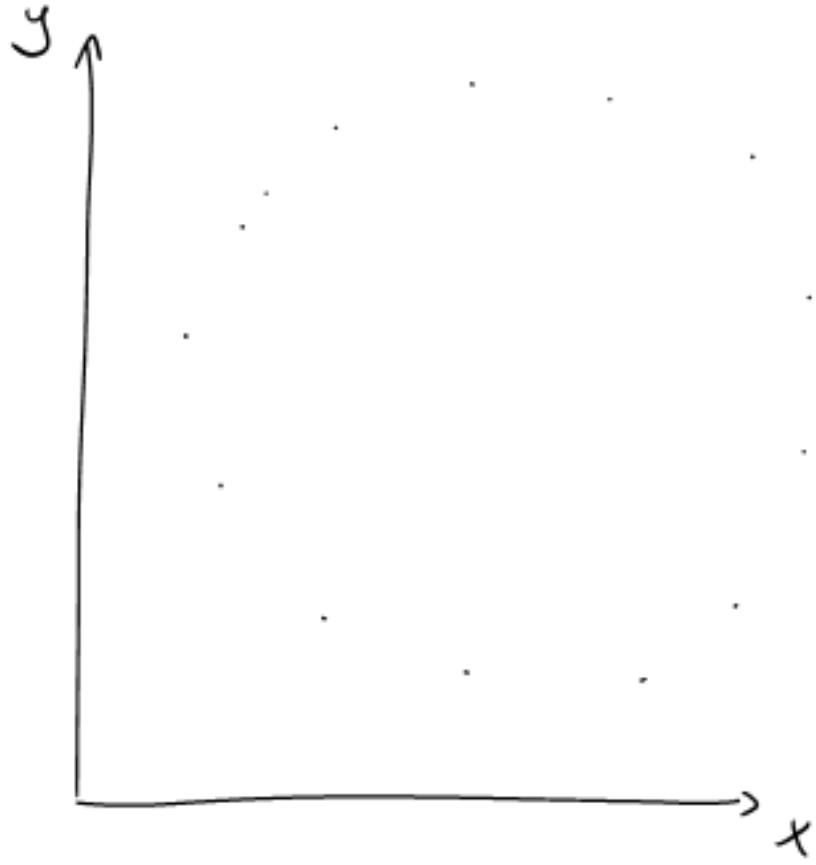

schwacher linearer Zusammenhang






kein Zusammenhang




Zusammenhang vorhanden,
aber nicht linear

empirische Kovarianz:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

mittlere quadratische Abweichung von $y$

Standardabweichung von $y$

Falls $y_i = a + b x_i$, $a, b \in \mathbb{R}$, $b \neq 0$

$\Rightarrow \bar{y} = a + b\bar{x}$

$s_y^2 = b^2 \cdot s_x^2$

$s_y = |b| \cdot s_x$

dann $|s_{xy}|$ maximal

$\Leftrightarrow s_x = \frac{1}{|b|} \cdot s_y$

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})((a + b x_i) - (a + b\bar{x}))$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(b x_i - b\bar{x}) = b \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})$$

$$= b\, s_x^2 = b \cdot s_x \cdot s_x$$

$$\Rightarrow s_{xy} = b \cdot s_x \cdot \frac{1}{|b|} \cdot s_y = \frac{b}{|b|} \cdot s_x \cdot s_y = \begin{cases} s_x \cdot s_y & \text{falls } b > 0 \\ -s_x \cdot s_y & \text{falls } b < 0 \end{cases}$$

Wir können die empirische Kovarianz also eingrenzen. Es gilt immer:

$$-s_x \cdot s_y \leq s_{xy} \leq s_x \cdot s_y$$

# Kovarianz und Maßeinheiten

Die Kovarianz hängt von den Maßeinheiten ab.

Es ist daher von Interesse ein **skalenfreies** Maß zu verwenden.

Ein solches Maß erhalten wir, wenn wir die Kovarianz durch die Standardabweichungen dividieren.

Das entsprechende Maß heißt (Bravais-Pearson) **Korrelationskoeffizient**.

Oder auch einfach (empirische) Korrelation.

# Bravais-Pearson-Korrelationskoeffizient

Seien $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ Ausprägungen zweier Merkmale $X$ und $Y$ mit kardinaler Skalierung und positiven Standardabweichungen $s_x, s_y > 0$.

**Bravais-Pearson-Korrelationskoeffizient von $X$ und $Y$:**

$$s_{xy} = r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

"rho"

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Eigenschaften der Korrelation

a) Für $u_i = a \cdot x_i + b$ und $v_i = c \cdot y_i + d$ gilt:

$$r_{uv} = \begin{cases} r_{xy} & \text{falls } a \cdot c > 0 \\ -r_{xy} & \text{falls } a \cdot c < 0 \end{cases}$$

b) $-1 \leq r_{xy} \leq 1$

c) Für $y_i = a \cdot x_i + b$: $r_{xy} = \begin{cases} 1 & \text{falls } a > 0 \\ -1 & \text{falls } a < 0 \end{cases}$

d) $r_{xy}$ ist eine Maßzahl für den **linearen** Zusammenhang zwischen $X$ und $Y$.

# Korrelation bei ordinal skalierten Merkmalen

Wie wird die Abhängigkeit gemessen, wenn nur Daten mit ordinaler Skalierung zur Verfügung stehen?

Das Berechnen von Mittelwerten der Merkmalsausprägungen ist hier nicht sinnvoll.

Die Rangkorrelation verwendet die Information, die in ordinalen Daten wirklich vorhanden ist: die Rangordnung.

Der Rangkorrelationskoeffizient von Spearman berechnet nun die Korrelation wie zuvor, aber unter Verwendung der Ränge.

ungeordnete Stichprobe: $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$

(Urliste)

geordnete Liste $x_{(1)}, x_{(2)}, \dots, x_{(n)}$

mit $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

Falls Beobachtung $x_i$ die kleinste Beobachtung ist, dann definieren wir $R(x_i) = 1$

—"— die größte Beobachtung: $R(x_i) = n$

Analog für $y$.

Falls z.B. $x_1 = x_2 = x_3$ am kleinsten:

$$\Rightarrow R(x_1) = R(x_2) = R(x_3) = \frac{1}{3}(1 + 2 + 3) = 2$$

Durchschnitt der Ränge

$$\overline{R_x} = \frac{1}{n} \sum_{i=1}^{n} R(x_i) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n}(1 + 2 + \dots + n)$$

Gaußsche Summenformel $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

$$\overline{R_x} = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

# Rangkorrelation von Spearman

Seien $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ Ausprägungen von Merkmalen $X$ und $Y$ mit ordinaler Skalierung.

Mit $R(x_i)$ und $R(y_i)$ bezeichnen wir die Ränge von $x_i$ und $y_i$ für $i = 1, \ldots, n$.[2]

**Rangkorrelation von Spearman:**

$$r_{xy}^{SP} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \frac{1+n}{2})(R(y_i) - \frac{1+n}{2})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \frac{1+n}{2})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \frac{1+n}{2})^2}}$$

[2] Herrscht für Beobachtungen „Gleichstand“, wird diesen das arithmetische Mittel der Ränge zugewiesen welche sie hätten wenn sie leicht unterschiedlich wären. Wenn z.B. 2 Ausprägungen auf Platz 1 stehen, dann wird beiden der Rang $\frac{1+2}{2} = 1,5$ zugewiesen.

# Korrelation bei nominal skalierten Merkmalen

Bei qualitativen Daten sind die zuvor besprochenen Maße nicht anwendbar, da weder Differenzen noch Ränge sinnvoll gebildet werden können.

Die Abhängigkeit für qualitative Daten kann basierend auf den **Häufigkeiten** der Kombinationen gemessen werden.

Die beobachteten Häufigkeiten der Kombinationen werden mit den Häufigkeiten verglichen, die zu erwarten wären, wenn die beiden Merkmale unabhängig voneinander wären.

Die Berechnung erfolgt anhand der bereits vorgestellten **Kontingenztabelle**.

# Kontingenzkoeffizient

$h_{i\bullet} = \sum_{j=1}^{l} h_{ij}$ — # der möglichen Ausprägungen von Y

Beobachtete Häufigkeit von $a_i$ und $b_j$: $h_{ij}$

Hypothetische Häufigkeit von $a_i$ und $b_j$ unter Unabhängigkeit:

$$\tilde{h}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$$

Ein Maß für Abhängigkeit:

Bei Unabhängigkeit $h_{ij} \approx \tilde{h}_{ij}$ $\Rightarrow \chi^2$ klein & positiv

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}}$$

Chi-quadrat

**Kontingenzkoeffizient**

$$0 \leq \quad K = \sqrt{\frac{\chi^2}{n + \chi^2}} \quad < 1$$

$\tilde{h}_{ij}$

| | ♀ | ♂ | |
|---|---|---|---|
| noch nie | 715 | 523 | 1238 |
| Leben | 84 | 61 | 145 |
| Monat | 78 | 57 | 136 |
| | 877 | 642 | 1519 |

$$61 \approx \frac{145 \cdot 642}{1519}$$

$$57 \approx \frac{136 \cdot 642}{1519}$$

$h_{ij} - \tilde{h}_{ij}$

| | ♀ | ♂ |
|---|---|---|
| nie | 26 | -26 |
| Leben | -20 | 20 |
| Monat | -6 | 7 |

# Korrelation ist nicht Kausalität

Der Korrelationskoeffizient ist ausschließlich ein Maß für den „Gleichklang“ von Daten.

Auf die Frage der Kausalität und Kausalitätsrichtung, das heißt welche der beiden Variablen wirkt auf die andere, gibt der Korrelationskoeffizient keine Auskunft.

Dennoch ist die Missinterpretation von Korrelationen im Sinn einer direkten kausalen Beziehung einer der häufigsten Fehler in der angewandten Statistik.

Beispiele dafür sind:

- Läuse senken Fieber
- Störche bringen Kinder
- Krankenhäuser schaden der Gesundheit
- Haarausfall (oder Schuhgröße) erhöht das Einkommen

Weitere interessante Beispiele zu spurious correlations gibt es hier.

# Kapitel 5

# Indexzahlen

# Preisindizes

**Preismesszahl:** Preisveränderungen eines einzelnen Gutes

**Preisindex:** Preisveränderung vieler Güter.

Gegeben seien für $n$ Güter:

| | | | | |
|---|---|---|---|---|
| Preise in Basisperiode 0 | $p_0(1)$ | $p_0(2)$ | $\dots$ | $p_0(n)$ |
| Preise in Berichtsperiode $t$ | $p_t(1)$ | $p_t(2)$ | $\dots$ | $p_t(n)$ |

# Preisindizes

**Preismesszahl:** Preisveränderungen eines einzelnen Gutes

**Preisindex:** Preisveränderung vieler Güter.

Gegeben seien für $n$ Güter:

| | | | | |
|---|---|---|---|---|
| Preise in Basisperiode 0 | $p_0(1)$ | $p_0(2)$ | $\dots$ | $p_0(n)$ |
| Preise in Berichtsperiode $t$ | $p_t(1)$ | $p_t(2)$ | $\dots$ | $p_t(n)$ |
| Mengen in Basisperiode 0 | $q_0(1)$ | $q_0(2)$ | $\dots$ | $q_0(n)$ |
| Mengen in Berichtsperiode $t$ | $q_t(1)$ | $q_t(2)$ | $\dots$ | $q_t(n)$ |

Der Preisindex reduziert die $4 \cdot n$ Beobachtungen auf eine Kennzahl.

# Preisindex nach Laspeyres

$$P_{0t}^{L} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_0(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)},$$

wobei

- $p_0(i)$ den Preis des Gutes i in der Basisperiode 0,
- $p_t(i)$ den Preis des Gutes i in der Periode t und
- $q_0(i)$ die Menge des Gutes i in der Basisperiode 0 bezeichnet.

Der Preisindex nach Laspeyres ist der Quotient aus den hypothetischen Gesamtausgaben der Periode $t$ (Berichtsperiode) bei Verwendung des Warenkorbes aus der Periode 0 (Basisperiode) und den Gesamtausgaben der Basisperiode.

# Preisindex nach Paasche

$$P_{0t}^{P} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_t(i)},$$

wobei

- $p_0(i)$ den Preis des Gutes $i$ in der Basisperiode 0,
- $p_t(i)$ den Preis des Gutes $i$ in der Periode $t$ und
- $q_t(i)$ die Menge des Gutes $i$ in der Berichtsperiode $t$ bezeichnet.

Der Preisindex nach Paasche ist also der Quotient aus den Gesamtausgaben in der Berichtsperiode und den hypothetischen Ausgaben für den Warenkorb der Berichtsperiode zu den Preisen der Basisperiode.

# Vergleich der Indizes von Laspeyres und Paasche

- In der Praxis ist der Laspeyres Index weiter verbreitet.
- Im Laspeyres Index werden **Verhaltensänderungen**, induziert durch sich ändernde Preise, nicht berücksichtigt: Der Warenkorb der Periode 0 findet Verwendung
- Zur Berechnung des Preisindex nach Paasche sind mehr Informationen notwendig: Die Mengen aus allen Berichtsperioden müssen vorliegen;
- zur Berechnung des Laspeyres Index genügen die Mengen aus der Basisperiode.

# Preisindex nach Fisher

$$P_{0t}^F = \sqrt{P_{0t}^L \cdot P_{0t}^P} \ ,$$

d. h. der Preisindex nach Fisher ist das geometrische Mittel des Laspeyres und des Paasche Preisindexes.

Der Preisindex nach Fisher $P_{0t}^F$ liegt immer zwischen den anderen beiden Preisindizes.

# Eigenschaft der Preisindizes

Wenn sich alle Preise von einer Periode zur nächsten um denselben Faktor $a$ ändern, so nehmen die 3 betrachteten Preisindizes $P^L$, $P^P$ und $P^F$ den Wert dieses Faktors an:

$$p_t(i) = a \cdot p_0(i) \text{ für alle i}$$

$$\Rightarrow P_{0t}^L = P_{0t}^P = P_{0t}^F = a$$

# Mengenindizes

**Mengenindex nach Laspeyres:**

$$Q_{0t}^{L} = \frac{\sum_{i=1}^{n} p_0(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)}$$

**Mengenindex nach Paasche:**

$$Q_{0t}^{P} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_t(i) \cdot q_0(i)}$$

**Mengenindex nach Fisher:**

$$Q_{0t}^{F} = \sqrt{Q_{0t}^{L} \cdot Q_{0t}^{P}}$$

## Preisindizes

### Laspeyres

Mengen konstant
→ Basisperiode 0

$$P_{0t}^{L} = \frac{\sum_{i=1}^{n} p_t(i)\, q_0(i)}{\sum_{i=1}^{n} p_0(i)\, q_0(i)}$$

### Paasche

Mengen konstant
→ Berichtsperiode t

$$P_{0t}^{P} = \frac{\sum_{i=1}^{n} p_t(i)\, q_t(i)}{\sum_{i=1}^{n} p_0(i)\, q_t(i)}$$

## Mengenindizes

$$Q_{0t}^{L} = \frac{\sum_{i=1}^{n} p_0(i)\, q_t(i)}{\sum_{i=1}^{n} p_0(i)\, q_0(i)}$$

Preise konstant
→ Basisperiode 0

$$Q_{0t}^{P} = \frac{\sum_{i=1}^{n} p_t(i)\, q_t(i)}{\sum_{i=1}^{n} p_t(i)\, q_0(i)}$$

Preise konstant
→ Berichtsperiode

# Zusammenfassung Teil 1 Deskriptive Statistik

- Datentypen:
  - qualitativ / quantitativ (quantifiziert)
  - nominale, ordinale und kardinale Skalierung
- Darstellung von absoluten und relativen Häufigkeiten
  - eindimensionale / mehrdimensionale Daten
- Lageparameter
  - Modalwert, Median, arithmetisches & geometrisches Mittel
- Streuungsmaße
  - Spannweite, durchschnittliche- , mittlere quadratische- & Standardabweichung
- Konzentrationsmaße
  - Variationskoeffizient, Lorenzkurve, Gini-Koeffizient, Herfindahl-Index
- Mehrdimensionale Daten
  - Darstellung: Kontingenztabelle, Streuungsdiagramm
  - Randverteilung, bedingte relative Häufigkeiten, Unabhängigkeit
  - Korrelationsrechnung: Kovarianz & drei Koeffizienten
- Indexzahlen
  - Preis- und Mengenindizes nach Laspeyres, Paasche & Fisher