

# Deskriptive Statistik

# Übersicht und Grundbegriffe

## Deskriptive (beschreibende) Statistik

Die deskriptive Statistik verfolgt das Ziel, Daten mithilfe von Tabellen, Grafiken oder auch Kennzahlen darzustellen und zu beschreiben.

## Grundgesamtheit / Population

Die Grundgesamtheit - auch Population genannt - beschreibt die Menge aller Objekte, über die wir eine Aussage treffen wollen. Die Bestimmung von gewissen Eigenschaften der Grundgesamtheit ist das Ziel der schließenden Statistik.

## Stichprobe

Die Stichprobe ist die Menge der beobachteten Objekten. In anderen Worten ist die Stichprobe eine (zufällige) Teilmenge der Grundgesamtheit. Diese Teilmenge wird in der deskriptiven Statistik untersucht.

## Merkmalsträger

Die Objekte dessen Eigenschaften wir beobachten.

## Merkmale

Eine gewisse Eigenschaft, die an den Merkmalsträgern beobachtet wird.

## Merkmalsausprägung / Ausprägung eines Merkmals

Merkmalsausprägungen sind verschiedene Werte, die ein gewisses Merkmal annehmen kann.

# Kategorische Daten (Merkmale)

## Nominale Daten

Die Ausprägungen eines Merkmals können unterschieden werden.

z.B.: Geschlecht, Autokennzeichen, Blutgruppe, Studiengänge an einer Hochschule.

## Ordinale Daten

Ordinalen Daten können zusätzlich zu der Eigenschaft der nominalen Daten sortiert werden. Dabei gilt entweder  $a < b$ ,  $a = b$  oder  $a > b$ . Zudem muss gelten: Wenn  $a > b$  und  $b > c$ , dann muss  $a > c$  sein. Somit können kumulierte Häufigkeiten berechnet werden.

z.B.: Klausurleistung, Kundenzufriedenheit, Lebensdauer eines Bauteils.

# Numerische Daten (Merkmale)

## Diskrete numerische Daten

Die Ausprägungen eines Merkmals können nur bestimmte, abzählbare Werte annehmen. Oft handelt es sich dabei um ganze Zahlen.

z.B.: Anzahl an Bauteilen, Anzahl an Studenten im Kurs.

## Kontinuierliche numerische Daten

Die Ausprägungen eines Merkmals können jeden beliebigen Wert innerhalb eines gegebenen Bereichs annehmen.

z.B.: Verbrauch in MWh pro Einwohner, Lufttemperatur in °C.

## Absolute Häufigkeiten

Die absolute Häufigkeit  $h_i$  beschreibt wie oft die  $i$ -te Ausprägung eines Merkmals vorkommt.

## Relative Häufigkeiten

Die relative Häufigkeit  $r_i$  berechnet sich wie folgt

$$r_i = \frac{h_i}{n},$$

wobei  $n = \sum_i h_i$  die Anzahl an Beobachtungen beschreibt. Zudem gilt  $\sum_i r_i = 1$ .

**Beispiel:**

d	Kleidung
0	T-Shirt
1	Hemd
2	Hemd
3	T-Shirt
4	Hemd
5	Pullover
6	T-Shirt

Kleidung	$h_i$	$r_i$
T-Shirt	3	$3/7$
Hemd	3	$3/7$
Pullover	1	$1/7$

Zudem ist ersichtlich, dass  $n = 7$  und  $\sum_i r_i = 1$  ist.

# Kumulierte Häufigkeiten

## Absolute kumulierte Häufigkeit

Gegeben sind  $n$  Beobachtungen eines ordinalen Merkmals mit  $m$  verschiedenen Ausprägungen  $x_i$  mit der Ordnung  $x_1 < x_2 < \dots < x_m$ . Die absolute kumulierte Häufigkeit wird berechnet als

$$F_{abs}(x_k) = \sum_{i=1}^k h_i \text{ mit } 1 \leq k \leq m,$$

wobei  $h_i$  die absolute Häufigkeit der Ausprägung  $x_i$  beschreibt.

Die absolute kumulierte Häufigkeit ist demnach die Summe der absoluten Häufigkeiten der Ausprägungen von der kleinsten bis zur jeweils gegebenen Schranke.

Beispiel: Anzahl der Studierenden mit einer Note nicht schlechter als eine 3 im Fach Wahrscheinlichkeit und Statistik.



# Kumulierte Häufigkeiten

## Relative kumulierte Häufigkeit

Gegeben sind  $n$  Beobachtungen eines ordinalen Merkmals mit  $m$  verschiedenen Ausprägungen  $x_i$  mit der Ordnung  $x_1 < x_2 < \dots < x_m$ . Die relative kumulierte Häufigkeit wird berechnet als

$$F_{rel}(x_k) = \sum_{i=1}^k r_i \text{ mit } 1 \leq k \leq m,$$

wobei  $r_i$  die relative Häufigkeit der Ausprägung  $x_i$  beschreibt. Die Funktion  $F_{rel}$  wird auch empirische Verteilungsfunktion genannt.

Die relative kumulierte Häufigkeit ist die Summe der relativen Häufigkeiten der Ausprägungen von der kleinsten bis zur jeweils gegebenen Schranke.

## Klasseneinteilung

Die Klasseneinteilung beschreibt die Einteilung in Abhängigkeit der Ausprägung eines Merkmals in Klassen.

Wann ist eine Klasseneinteilung sinnvoll?

- ▶ Bei vielen Ausprägungen eines Merkmals
- ▶ Falls fast kein Unterschied zwischen den Ausprägungen vorhanden ist
- ▶ Falls einzelne Ausprägungen selten oder gar nicht vorkommen

## Was ist bei der Klasseneinteilung zu beachten?

- ▶ Jede Ausprägung wird genau einer Klasse zugeordnet.
- ▶ Die Klassen sind disjunkt und zudem entweder links offen und rechts abgeschlossen oder vice versa.
- ▶ Die erste Klasse kann bis  $-\infty$  und die letzte Klasse kann bis  $+\infty$  gehen.
- ▶ Die Reduktion der Daten geht mit einem Verlust an (wesentlichen) Informationen einher.
- ▶ Faustregeln für die Bestimmung der Anzahl an Klassen sind:

	Anzahl Beobachtungen	Klassen
	$< 50$	5 bis 7
i)	$50 \leq x < 100$	6 bis 10
	$100 \leq x < 250$	7 bis 12
	$\geq 250$	10 bis 20

- ii) Anzahl an Klassen  $\leq \sqrt{\text{Anzahl an Ausprägungen}}$ .

## Lagemaß

Ein Lagemaß ist eine statistische Kennzahl, die durch Zusammenfassen von Informationen zu einem repräsentativen Wert ermittelt werden kann. Dieser repräsentative Wert liefert Positionsinformationen der zugrunde liegenden Daten.

Beispiele für gängige Lagemaße:

- ▶ Arithmetischer Mittelwert
- ▶ Median und Modus
- ▶ Quantile

## Arithmetisches Mittel

Gegeben sind  $n$  Beobachtungen mit Ausprägungen  $x_i$ . Das arithmetische Mittel - auch Mittelwert, Mittel oder Durchschnitt genannt - wird wie folgt berechnet

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Eigenschaften vom arithmetischen Mittel sind:

- ▶ Die Abweichungen  $\epsilon_i$  der Ausprägungen  $x_i$  vom Mittelwert  $\bar{x}$  berechnen sich wie folgt:  $\epsilon_i = x_i - \bar{x}$  für  $i = 1, \dots, n$ .
- ▶ Falls die Summe der Abweichungsquadrate minimiert werden soll ( $\min_z \sum_{i=1}^n (x_i - z)^2$ ), liefert  $z = \bar{x}$  den minimierenden Wert.
- ▶ Das arithmetische Mittel ist eindeutig.

## Median

Gegeben sind  $n$  Beobachtungen eines ordinalen Merkmals mit geordneten Ausprägungen  $x_i$ , d.h.  $x_1 \leq \dots \leq x_n$ . Der Median - auch Zentralwert genannt - wird wie folgt berechnet

$$x_{median} = \begin{cases} x_{m+1} & \text{für ungerades } n = 2m + 1 \\ \frac{1}{2}(x_m + x_{m+1}) & \text{für gerades } n = 2m. \end{cases}$$

- ▶ Der Median ist im Vergleich mit dem arithmetischen Mittel robust gegenüber Ausreißern.
- ▶ Der Median ist eindeutig.

## Modus

Der Modus - auch als Modalwert bekannt - beschreibt jene Ausprägungen eines Merkmals, die am häufigsten in der Stichprobe vorkommen.

- ▶ Der Modus ist nicht immer eindeutig, kann aber immer berechnet werden.
- ▶ Um den Modus berechnen zu können, müssen zumindest nominale Daten vorliegen.
- ▶ Verteilungen werden
  - ▶ mit einem Modus unimodale Verteilung,
  - ▶ mit zwei Modi bimodale Verteilung und
  - ▶ mit zwei oder mehr Modi multimodale Verteilunggenannt.

## Quantil

Gegeben sind  $n$  Beobachtungen eines ordinalen Merkmals mit geordneten Ausprägungen  $x_i$ , d.h.  $x_1 \leq \dots \leq x_n$ . Sei  $p \in (0, 1)$ , dann berechnet sich das empirische  $p$ -Quantil  $x_p$  wie folgt

$$x_p = \begin{cases} \frac{1}{2}(x_{n \cdot p} + x_{n \cdot p + 1}) & \text{falls } n \cdot p \in \mathbb{N} \\ x_{\lfloor n \cdot p + 1 \rfloor} & \text{falls } n \cdot p \notin \mathbb{N}. \end{cases}$$

Hier bezeichnet  $\lfloor \cdot \rfloor$  die Abrundungsfunktion (auch Gaußklammer genannt).

- ▶ Median: Der Median entspricht dem 0.5-Quantil.
- ▶ Terzil: Das untere Terzil entspricht dem  $\frac{1}{3}$ -Quantil und das obere Terzil dem  $\frac{2}{3}$ -Quantil.
- ▶ Quartil: Das erste Quartil ( $Q_1$ ) beschreibt das 0.25-Quantil, das zweite Quartil ( $Q_2$ ) den Median und das dritte Quartil ( $Q_3$ ) das 0.75-Quantil.
- ▶ Dezil:  $p \in \{0.1, 0.2, \dots, 0.9\}$ .
- ▶ Perzentil:  $p \in \{0.01, 0.02, \dots, 0.99\}$ .



## Streumaß

Ein Streumaß - auch Dispersionsmaß genannt - ist eine statistische Kennzahl, die die Streuung der Beobachtungen angibt. Um ein Streumaß definieren zu können, wird ein Abstandsmaß benötigt. Ein kleines (großes) Streumaß bedeutet, dass die Beobachtungen konzentriert (zerstreut) sind.

Streumaße können wie folgt unterteilt werden:

- ▶ Abstand zwischen Lagemaße ( $R$ ,  $IQA$ )
- ▶ Streuung um ein Lagemaß
- ▶ Streuung relativ zu einem Lagemaß

# Abstand zwischen Lagemaße

## Spannweite

Die Spannweite berechnet sich aus der Differenz zwischen der maximalen ( $x_{max}$ ) und der minimalen ( $x_{min}$ ) Ausprägung eines Merkmals in der Stichprobe

$$R = x_{max} - x_{min}.$$

## Interquartilsabstand

Der Interquartilsabstand berechnet sich aus der Differenz zwischen dem dritten Quartil ( $Q_3$ ) und dem ersten Quartil ( $Q_1$ )

$$IQA = Q_3 - Q_1.$$

- ▶  $R$  wird im Vergleich mit  $IQA$  stark durch Ausreißer beeinflusst.
- ▶  $R$  und  $IQA$  existieren immer und sind eindeutig.
- ▶ Länge der Box im Boxplot.

## Ausreißer bestimmen mit IQA

Ein Ausreißer ist eine Beobachtung, die von den meisten anderen Beobachtungen weit entfernt liegt. Eine Möglichkeit Ausreißer zu identifizieren bietet der IQA. Demnach ist eine Beobachtung  $x$  ein Ausreißer, falls:

$$x < Q_1 - 1.5 \cdot IQA \text{ oder}$$

$$x > Q_3 + 1.5 \cdot IQA.$$

# Streuung um ein Lagemaß

## Mittlere absolute Abweichung

Gegeben sind  $n$  Beobachtungen  $x_1, \dots, x_n$ . Die mittlere absolute Abweichung berechnet sich wie folgt

$$MAD_{mean} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

$$MAD_{median} = \frac{1}{n} \sum_{i=1}^n |x_i - x_{median}|.$$

- ▶ Hier wird als Abstandsmaß die Betragsfunktion  $|\cdot|$  verwendet.
- ▶  $MAD_{mean}$  und  $MAD_{median}$  sind nicht robust gegenüber Ausreißer.

# Streuung um ein Lagemaß

## Empirische Varianz

Gegeben sind  $n$  Beobachtungen  $x_1, \dots, x_n$ . Die empirische Varianz  $s^2$  berechnet sich wie folgt

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

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

Hier wird als Abstandsmaß der quadratische Abstand  $(.)^2$  verwendet.

- ▶ Falls die Grundgesamtheit gegeben ist, wird Gleichung 1 verwendet.
- ▶ Falls eine Stichprobe gegeben ist, wird Gleichung 2 verwendet.
- ▶ Das  $s^2$  in Gleichung 2 ist ein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit.
- ▶ Die empirische Standardabweichung  $s$  ist gegeben durch  $s = \sqrt{s^2}$ .

# Streuung relativ zu einem Lagemaß

Um unterschiedliche Datenreihen miteinander vergleichen zu können, wird eine dimensionslose Größe benötigt.

## Variationskoeffizient

Der Variationskoeffizient - auch Abweichungskoeffizient genannt - ist ein relatives Streumaß, das nicht von der Maßeinheit der Beobachtungen abhängt. Dieses Streumaß berechnet sich wie folgt

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

- ▶  $V$  ist nur sinnvoll für Datenreihen mit ausschließlich positiven (oder ausschließlich negativen) Werten.
- ▶ Dieses Streumaß beschreibt die normierte Standardabweichung.
- ▶ Der Variationskoeffizient ermöglicht wegen seiner Dimensionslosigkeit einen Vergleich der Streuung unterschiedlicher Datenreihen.
- ▶  $V$  ist nicht robust gegenüber Ausreißer.

# Streuung relativ zu einem Lagemaß

## Quartilsdispersionskoeffizient

Der Quartilsdispersionskoeffizient ist ein relatives Streumaß, das nicht von der Maßeinheit der Beobachtungen abhängt. Dieses Streumaß berechnet sich wie folgt

$$V_r = \frac{IQA}{x_{median}}.$$

- ▶  $V_r$  ist eine robuste Version von  $V$ .
- ▶ Dieses Streumaß beschreibt den normierten Interquartilsabstand.

# Zusammenhangsmaße

## Übersicht

- ▶ Nominale sowie ordinale Daten:
  - ▶ Zusammenhangsmaß: Es wird der Chi-Quadrat Koeffizient und daraus der Kontingenzkoeffizient bestimmt.
- ▶ Metrische Daten: Die Ausprägung eines Merkmals besteht aus einer Zahl, einer Dimension und einem Nullpunkt. Zum Beispiel: Einkommen, Alter, Geschwindigkeit.
  - ▶ Zusammenhangsmaß: Die Kovarianz und der Korrelationskoeffizient werden bestimmt.



# Randhäufigkeiten

Seien  $x_1, \dots, x_m$  Merkmalsausprägungen der Variable X und  $y_1, \dots, y_k$  Merkmalsausprägungen der Variable Y. Wie oben bereits definiert, beschreibt  $h_{ij}$  die absolute Häufigkeit der Merkmalsausprägung  $x_i$  mit  $y_j$ .

	$y_1$	$\dots$	$y_k$	<b>Randhäufigkeiten von X</b>
$x_1$	$h_{11}$	$\dots$	$h_{1k}$	$h_{1\bullet}$
$x_2$	$h_{21}$	$\dots$	$h_{2k}$	$h_{2\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_m$	$h_{m1}$	$\dots$	$h_{mk}$	$h_{m\bullet}$
<b>Randhäufigkeiten von Y</b>	$h_{\bullet 1}$	$\dots$	$h_{\bullet k}$	$h_{\bullet\bullet}$

Die Randhäufigkeiten werden wie folgt berechnet:  $h_{\bullet j} = \sum_i h_{ij}$  oder  $h_{i\bullet} = \sum_j h_{ij}$ .  
Die Summe der Randhäufigkeiten ergibt sich aus  $n = h_{\bullet\bullet} = \sum_j \sum_i h_{ij}$ .

## Von absoluten zu relativen Häufigkeiten

Wir können die Tabelle von der vorherigen Folie auch in relative Häufigkeiten umwandeln. Dafür erinnern wir uns, dass  $r_i = \frac{h_i}{n}$  die relative Häufigkeit der  $i$ -ten Ausprägung und  $n$  die Anzahl an Beobachtungen beschreibt. Demnach gilt:

	$y_1$	$\dots$	$y_k$	<b>Randhäufigkeiten von X</b>
$x_1$	$u_{11}$	$\dots$	$u_{1k}$	$u_{1\bullet}$
$x_2$	$u_{21}$	$\dots$	$u_{2k}$	$u_{2\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_m$	$u_{m1}$	$\dots$	$u_{mk}$	$u_{m\bullet}$
<b>Randhäufigkeiten von Y</b>	$u_{\bullet 1}$	$\dots$	$u_{\bullet k}$	$u_{\bullet\bullet}$

Auch hier gilt:  $u_{\bullet j} = \sum_i u_{ij}$ ,  $u_{i\bullet} = \sum_j u_{ij}$  und  $u_{\bullet\bullet} = \sum_j \sum_i u_{ij} = 1$ .

## Beispiel 1: Kontingenztafel

	BB	VZ	Summe (Zeile)
Männlich	12	20	32
Weiblich	3	18	21
Summe (Spalte)	15	38	$n = 53$

- ▶ Wie hoch war der Anteil der Studenten im Studiendesign BB insgesamt?
- ▶ Wie hoch war der Anteil der weiblichen Studenten im Studiendesign VZ?
- ▶ Wie hoch war der Anteil vom Studiendesign VZ an den männlichen Studierenden?

# Statistische Abhängigkeit und Unabhängigkeit

Unabhängig ist die Tabelle, wenn das Geschlecht nicht mit dem Studiendesign zusammenhängt. Dafür muss folgende Gleichung erfüllt sein:

$$\tilde{h}_{ij} * n = h_{i\bullet} \cdot h_{\bullet j}.$$

$\tilde{h}_{ij}$  beschreibt den erwarteten Wert der absoluten Häufigkeit der Merkmalskombination in der Zeile  $i$  und der Spalte  $j$ .

Falls die Gleichung für eine Zelle nicht erfüllt ist, gibt es eine Abhängigkeit der Merkmale.

Ergebnis:

$$12 \cdot 53 \neq 15 \cdot 32$$

$$3 \cdot 53 \neq 15 \cdot 21$$

$$20 \cdot 53 \neq 38 \cdot 32$$

$$18 \cdot 53 \neq 38 \cdot 21$$

In diesem Fall liegt eine Abhängigkeit der Merkmale vor.

## Beispiel 2 Kontingenztafel

	$y_1$	$y_2$	<b>Summe (Zeile)</b>
$x_1$	20	20	40
$x_2$	6	6	12
<b>Summe (Spalte)</b>	26	26	$n = 52$

In diesem Fall liegt eine Unabhängigkeit der Merkmale vor.

## Chi-Quadrat Koeffizient

Chi-Quadrat wird wie folgt berechnet

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

- ▶  $\chi^2$  beschreibt Chi-Quadrat.
- ▶  $m$  beschreibt die Anzahl an Zeilen.
- ▶  $k$  beschreibt die Anzahl an Spalten.
- ▶  $h_{ij}$  beschreibt die absolute Häufigkeit der Merkmalskombination in der Zeile  $i$  und der Spalte  $j$ .
- ▶  $\tilde{h}_{ij}$  beschreibt den erwarteten Wert der absoluten Häufigkeit der Merkmalskombination in der Zeile  $i$  und der Spalte  $j$ .

# Ergebnis für Beispiel 1

- ▶ Schritt 1: Berechnung  $\tilde{h}_{ij}$ .
  - ▶  $\tilde{h}_{11} : \frac{15 \cdot 32}{53} \approx 9$
  - ▶  $\tilde{h}_{21} : \frac{15 \cdot 21}{53} \approx 6$
  - ▶  $\tilde{h}_{12} : \frac{38 \cdot 32}{53} \approx 23$
  - ▶  $\tilde{h}_{22} : \frac{38 \cdot 21}{53} \approx 15$
- ▶ Schritt 2: Berechnung  $(h_{ij} - \tilde{h}_{ij})^2$ .
  - ▶  $(12 - 9)^2 = 9$
  - ▶  $(3 - 6)^2 = 9$
  - ▶  $(20 - 23)^2 = 9$
  - ▶  $(18 - 15)^2 = 9$
- ▶ Schritt 3: Berechnung  $\frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$ .
  - ▶  $\frac{9}{9} = 1$
  - ▶  $\frac{9}{6} = 1.5$
  - ▶  $\frac{9}{23} \approx 0.39$
  - ▶  $\frac{9}{15} = 0.6$
- ▶ Schritt 4:  $X^2 = 1 + 1.5 + 0.39 + 0.6 = 3.49$

## Aussagekraft von $\chi^2$

- ▶ Aussagekraft des Koeffizienten ist gering.
- ▶  $\chi^2 \in [0, n \cdot (\min\{m, k\} - 1)]$ .
- ▶  $\chi^2 = 0$ : Völlige Unabhängigkeit der Merkmale.
- ▶ Wert bei vollkommener Abhängigkeit hängt ab von:
  - ▶ Anzahl an Ausprägungen.
  - ▶ Untersuchter Gesamtheit  $n$ .
- ▶  $\chi^2$  ist demnach nicht standardisiert und daher nur begrenzt vergleichbar.

Aus diesem Grund können wir den Chi-Quadrat Koeffizienten in den Kontingenzkoeffizient umwandeln.



## Korrigierter Kontingenzkoeffizient

Der korrigierte Kontingenzkoeffizient wird wie folgt berechnet

$$K^p = \sqrt{\frac{X^2}{n + X^2} \cdot \frac{\min\{m, k\}}{\min\{m, k\} - 1}}.$$

- ▶  $K^p$  beschreibt den korrigierten Kontingenzkoeffizient.
- ▶  $X^2$  beschreibt den Chi-Quadrat Koeffizient.
- ▶  $n$  beschreibt die Größe der Stichprobe.
- ▶  $\min\{m, k\}$  nimmt das Minimum der Anzahl an Zeilen und Spalten.

## Beispiel und Interpretation

- ▶  $K^p = \sqrt{\frac{3.49}{53+3.49} \cdot \frac{2}{1}} \approx 0.35$
- ▶ Der Kontingenzkoeffizient nimmt Werte zwischen 0 und 1 an.
- ▶ Ein Wert von Null bedeutet, dass kein Zusammenhang zwischen den Merkmalen vorliegt.
- ▶ Ein Wert von Eins bedeutet, dass ein vollständiger Zusammenhang zwischen den Merkmalen existiert.
- ▶ In unserem Beispiel liegt ein schwacher statistischer Zusammenhang zwischen den Merkmalen Geschlecht und Studiendesign vor.
- ▶ Der korrigierte Kontingenzkoeffizient ermöglicht den Vergleich von Kontingenztabellen unterschiedlicher Größen.

# Kovarianz

## Korrigierte empirische Kovarianz

Die empirische korrigierte Kovarianz einer Stichprobe mit  $n$  Beobachtungen berechnet sich wie folgt:

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

- ▶ Die korrigierte empirische Kovarianz ist ein erwartungstreuer Schätzer der Kovarianz der gesamten Population mithilfe einer Stichprobe.
- ▶ Die empirische Kovarianz beschreibt den linearen Zusammenhang zweier Merkmale.
- ▶ Die Kovarianz ist keine dimensionslose Größe.
- ▶ Falls  $s_{xy} > 0$  ist, liegt ein positiver Zusammenhang vor. Vice versa für  $s_{xy} < 0$ .

# Korrelationskoeffizient

## Empirischer Korrelationskoeffizient

Der empirische Korrelationskoeffizient für eine zweidimensionale Stichprobe  $(x_i, y_i)_{i=1}^n$  berechnet sich wie folgt:

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

- ▶ Dieses Zusammenhangsmaß beschreibt den Grad des linearen Zusammenhangs zwischen zwei Merkmalen und ist eine dimensionslose Größe.
- ▶ Die Korrelation bedeutet nicht, dass die Änderung einer Variable eine Änderung der anderen Variable zur Folge hat.
- ▶ Falls  $r_{xy} \approx 0$  kann trotzdem ein nicht-linearer Zusammenhang vorliegen.

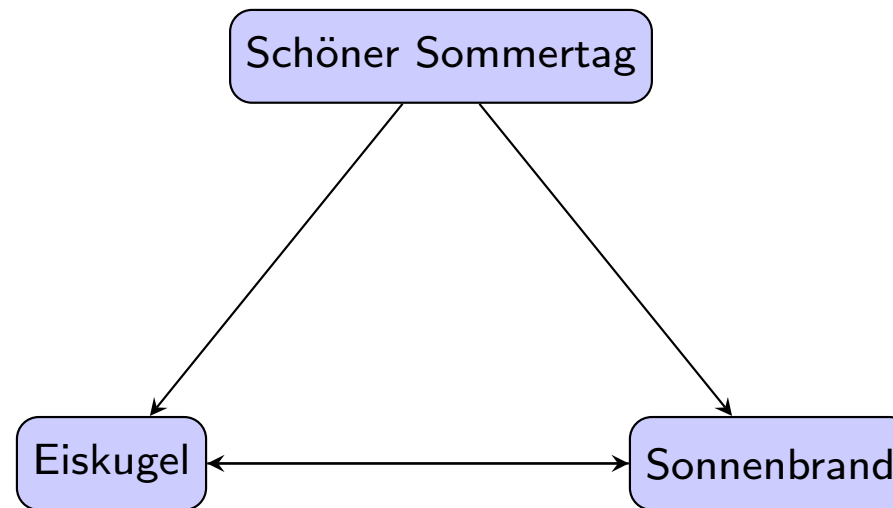
# Interpretation

- ▶  $|r_{xy}| = 1$ : Die zwei Merkmale liegen im 2-dimensionalen Koordinatensystem auf einer Linie.
- ▶ Ab  $|r_{xy}| > 0.5$  sprechen wir davon, dass der lineare Zusammenhang groß ist.
- ▶ Bei  $|r_{xy}| \approx 0.3$  sprechen wir von einer moderaten Korrelation.
- ▶ Bei  $|r_{xy}| \approx 0.1$  sprechen wir von einer geringen Korrelation.
- ▶ Sprechweise bei z.B.  $r_{xy} = 0.8$ : Ein hohes  $x$  geht mit einem hohen  $y$  einher.

# Kausalität

- ▶ Unter Kausalität wird verstanden, dass die Änderung einer Variable  $x$  eine Änderung der Variable  $y$  zur Folge hat.
- ▶ Korrelation impliziert keine Kausalität.

## Beispiel Korrelation vs. Kausalität



Es liegt eine positive Korrelation zwischen Eiskugel und Sonnenbrand vor. Kausalität: Ein schöner Sommertag hat zur Folge, dass mehr Eis konsumiert wird und die Individuen einen Sonnenbrand bekommen können.

# Scheinkorrelation

Falls ein drittes Merkmal, das nicht berücksichtigt wurde, die Ursache des Zusammenhangs von zwei Merkmalen ist, so nennt man dies eine Scheinkorrelation.

- ▶ Störche und Geburten

- ▶ Urbanisierung und Industrialisierung: Den Störchen wird ihr Lebensraum entzogen.
- ▶ Urbanisierung und Industrialisierung: Zunahme der Erwerbstätigkeit von Frauen und somit sinkende Geburtenrate.

- ▶ Schuhgröße und Einkommen

- ▶ Drittes Merkmal Geschlecht entscheidend.
- ▶ Frauen haben im Allgemeinen kleinere Schuhnummern und Einkommen.
- ▶ Vice Versa für Männer.