

Datensätze und statistische Grundlagen: Begriffe, Definitionen, Konzepte

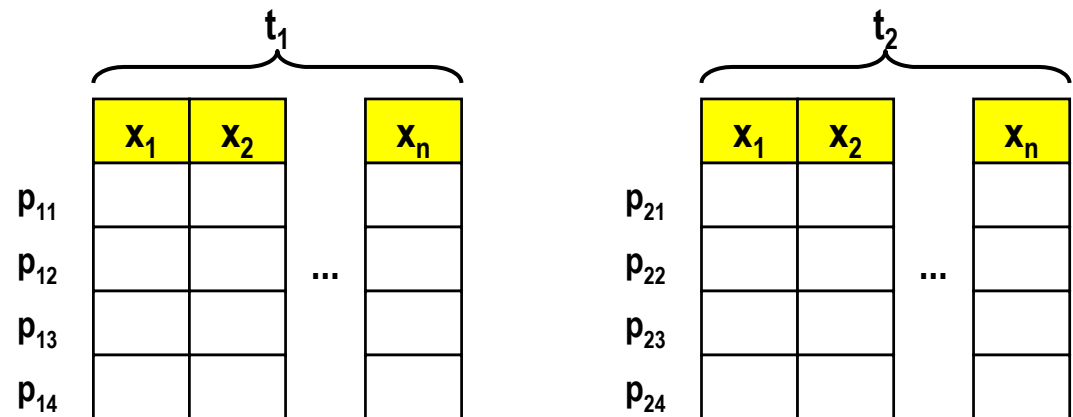
27. Oktober 2003

Datensätze: Querschnitt ↔ Längsschnitt I

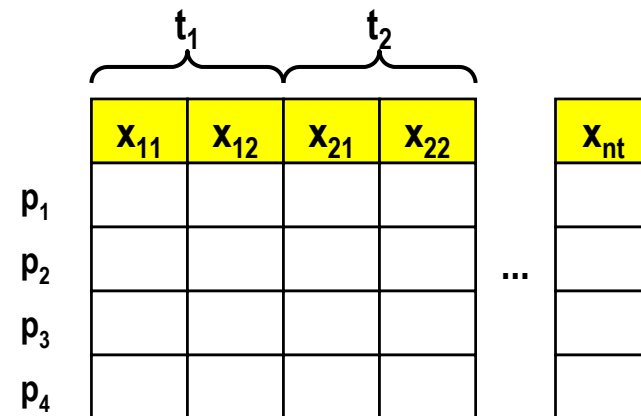
- > Querschnittsdaten = nur zu einem Zeitpunkt erhoben oder keine zeitspezifischen Informationen
- > Längsschnittdaten = zu mehreren Zeitpunkten erhoben oder zeitspezifische Informationen
 - Trendstudie: mind. drei Querschnittserhebungen, gleiche Instrumente, verschiedene Personen
 - Panel: mehrere Erhebungszeitpunkte mit gleichem Abstand, gleiche Instrumente, gleiche Personen
- > kausale Schlussfolgerungen empirisch nur mit Längsschnittdaten zu ziehen (Ursache-Wirkungs-Zusammenhang ist zeitabhängig)
- > theoretische Kausalitäten auch mit Querschnittsdaten möglich, aber nicht empirisch testbar

Datensätze: Querschnitt \leftrightarrow Längsschnitt II

> Trendstudie:



> Panelstudie:



Datensätze: Querschnitt ↔ Längsschnitt III

spezifische Probleme von Panelstudien:

- > Panelmortalität:
 - stetige Verringerung des Panel-Bestands (ca. 5-15% pro Welle)
 - Ursachen: Umzug, Tod, Verweigerung, „drop out“
 - Gegenmaßnahme: Panelpflege
- > Paneleffekte:
 - Anpassung von Antworten an vorhergehende Befragung
- > Beispiele für Panelstudien: SOEP, BHPS, PSID, ECHP, IAB-Betriebspanel
- > Beispiele für Querschnittsbefragungen: Politbarometer, ALLBUS, Eurobarometer

Daten: Aggregatdaten ↔ Individualdaten

- > Aggregatdaten geben Auskunft über Gruppen von einzelnen Objekten (Haushalte, Gemeinden, Staaten, Wahlkreise etc.)
- > Individualdaten geben Auskunft über einzelne Objekte (= statistische Einheiten, i.d.R. Personen)
- > Aggregation = Zusammenfassung von Daten zu einer höheren Ebene (Aggregationsniveau)
 - ist meistens mit Informationsverlust verbunden
- > Fehler bei unterschiedlicher Aussage- und Untersuchungseinheit
 - ökologischer Fehlschluss: Aussageeinheit = Wähler, Untersuchungseinheit = Wahlkreis
 - individualistischer Fehlschluss: Aussageeinheit liegt auf höherem Aggregationsniveau als Untersuchungseinheit

Daten: Skalenniveau

- > Nominalskalenniveau
 - Daten können nur bzgl. „gleich“ bzw. „ungleich“ beurteilt werden
 - Beispiele: Parteipräferenz, Berufsbranche
- > Ordinalskalenniveau
 - Daten können in eine bestimmte Reihenfolge gebracht werden
 - Beispiele: Hausarbeitsnoten, Rankings
- > metrisches Skalenniveau
 - Intervallskalenniveau: Unterschiede zwischen den Ausprägungen könne interpretiert werden
 - Beispiele: Temperatur ($^{\circ}$ C)
 - Ratioskala: es existiert ein absoluter Nullpunkt
 - Beispiele: Temperatur ($^{\circ}$ K), Preis
- > Transformation ist nur von höherem zu niedrigerem Skalenniveau unter Informationsverlust möglich

Statistik: Definitionen

- > neulat. *statisticus* = staatswissenschaftlich (seit dem 17. Jh. gebräuchlich)
- > „Statistik ist die Lehre von Methoden zur Gewinnung, Charakterisierung und Beurteilung von zahlenmäßigen Informationen über die Wirklichkeit.“
- > „Die Methoden der Statistik sind allgemein anwendbar, d.h. sie sind nicht beschränkt auf bestimmte inhaltliche Fragestellungen. Dies heißt aber nicht notwendig, dass sie auch in jedem Fall sinnvoll angewendet werden.“
- > „Statistik ist das methodische Vorgehen bei der Beschaffung von Informationen, die man braucht, um vernünftige Entscheidungen treffen zu können.“

Statistik: Wozu?

Informationsgehalt

**Klassierung,
Gruppenbildung**

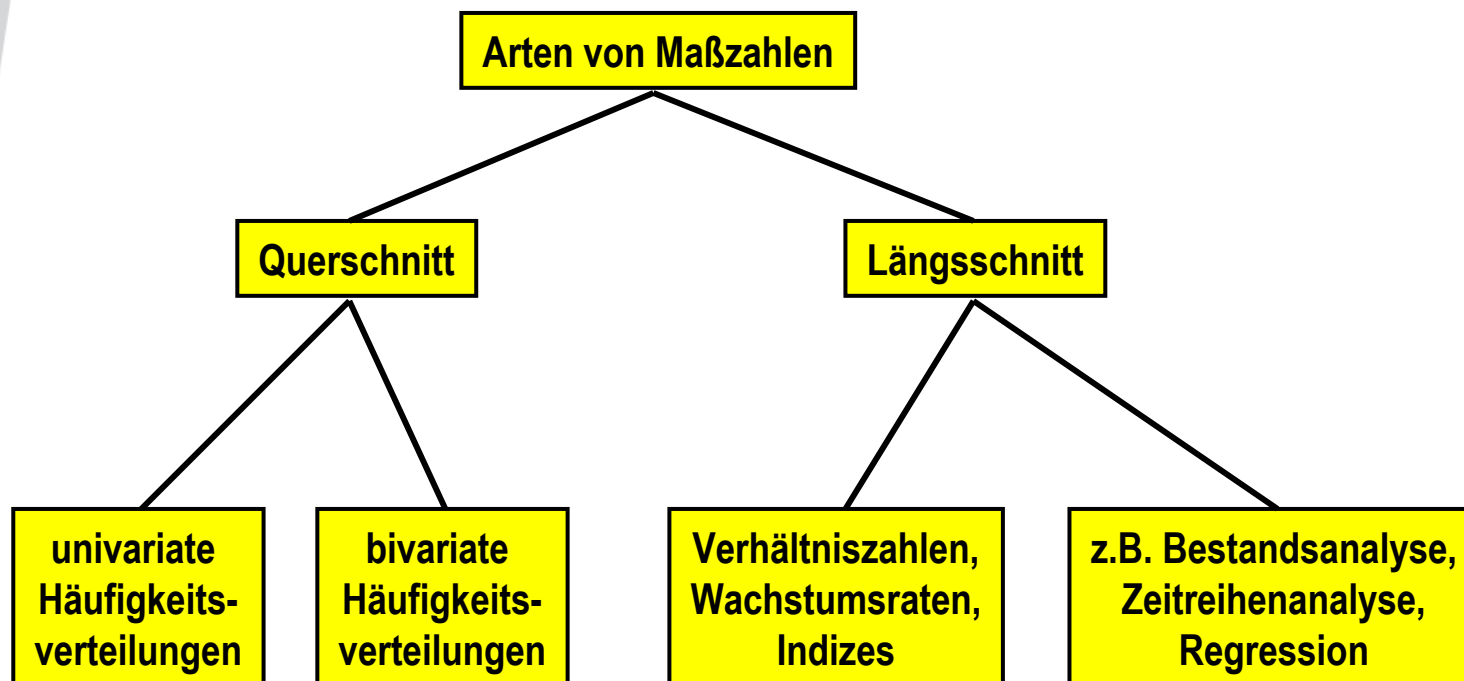
**Daten kennen
lernen + verstehen**

Qualität

**Hypothesen-
bildung**

Statistik: Maßzahlen

- > Definition: Maßzahlen (Kennzahlen) dienen der zusammenfassenden Beschreibung von Daten durch eine Zahl



Statistik: Masse, Einheit

- > statistische Masse (Population) = hinsichtlich sachlicher, räumlicher und zeitlicher Kriterien sinnvoll gebildete Gesamtheit von statistischen Einheiten
 - Grundgesamtheit
 - Teilgesamtheit (Auswahl \leftrightarrow Stichprobe)
 - Bestandsmasse (stock) \leftrightarrow Bewegungsmasse (flow)
 - SOEP: Wohnbevölkerung in Deutschland (Stichprobe)
- > statistische Einheit (Merkmalsträger) = Träger von Informationen bzw. Eigenschaften, die im Rahmen einer empirischen Untersuchung von Interesse sind
 - Individuen, Haushalte, Unternehmen, Wahlberechtigte
 - SOEP: Haushalte + Personen

Statistik: Merkmale, Variablen

Merkmal = Eigenschaften der statistischen Einheiten bzw. Menge an Merkmalsausprägungen

Variablen = Merkmalswerten zugeordnete Zahlen

Arten von Merkmalen:

- > intensive Merkmale \leftrightarrow extensive Merkmale
 - intensiv = Summe ist nicht sinnvoll interpretierbar (z.B. Intelligenz)
- > manifeste Merkmale \leftrightarrow latente Merkmale
 - manifest = direkt beobachtbar (z.B. Körpergröße)
- > diskrete Variable \leftrightarrow stetige Variable
 - diskret = endlich viele Werte im Intervall (Kinderzahl)

Variablen: latent ↔ manifest

- > Abgrenzung oft schwierig
 - Beispiel: Einkommen = „klass.“ manifeste Variable
 - aber: Antworten auf Einkommensfragen sehr ungenau
 - deshalb: immer Fragebogen und Operationalisierung beachten
- > Beispiele:
 - latent: Einstellungen (Ausländerfeindlichkeit, Rechts-Links etc.)
 - manifest: Geschlecht, (Einkommen) etc.
- > Methode zur Analyse latenter Variablen: Lineare Strukturgleichungsmodelle (LISREL)

Statistik: Arithmetisches Mittel, Modus + Median

- > arithmetisches Mittel:

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

- > Median: Wert, der eine nach Größe sortierte Reihe von Messwerten halbiert
- > Modus: Der Messwert, der in einer Verteilung am häufigsten vorkommt.

Statistik: Varianz

> Varianz:

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

> gibt die durchschnittliche Variation aller Merkmale an

Statistik: Standardabweichung

> Standardabweichung:

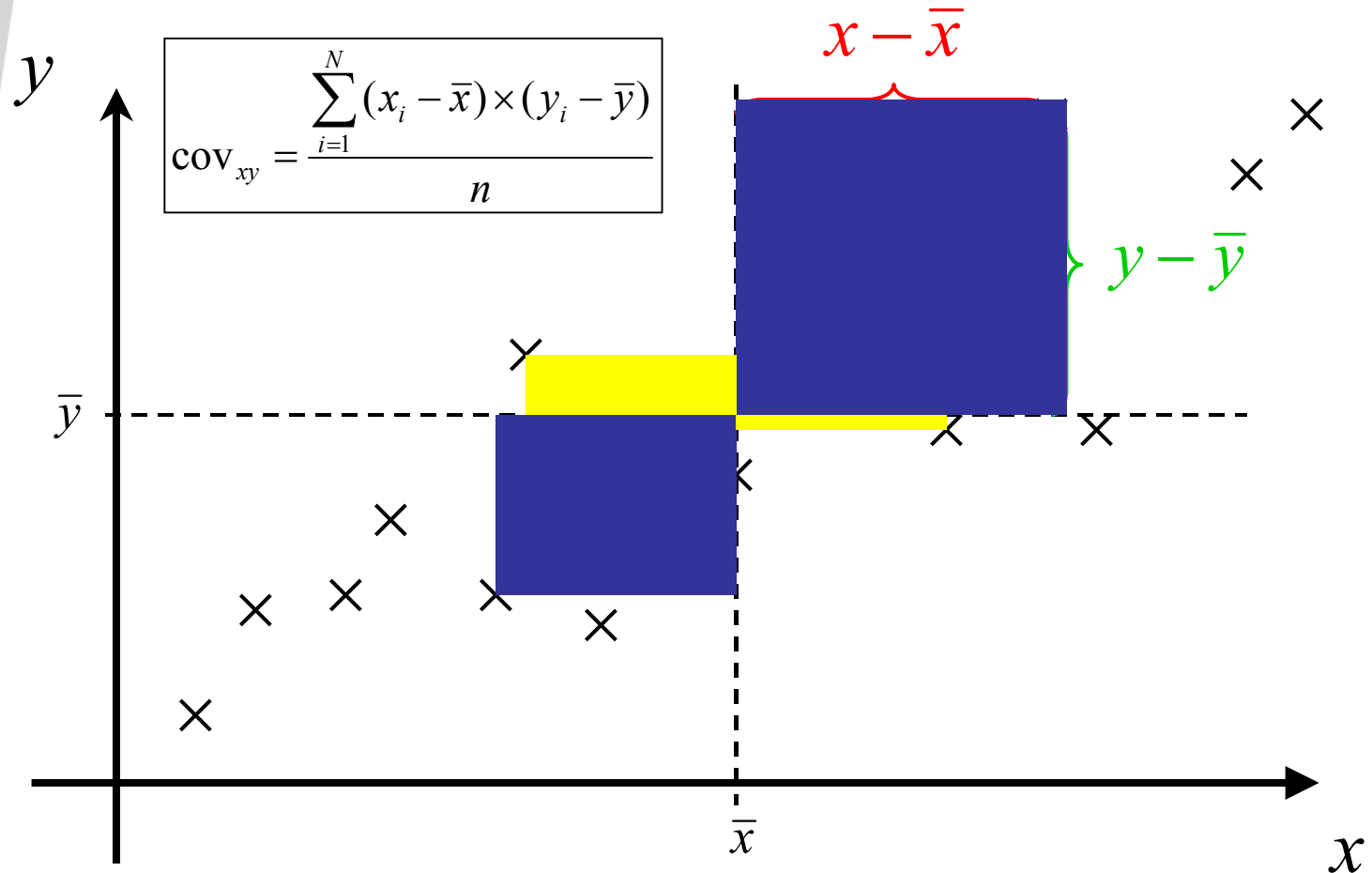
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n}}$$

> korrigiert die Verzerrung durch die Quadrierung der Varianz

Korrelation I

- > Korrelation = Analyse der Stärke der Interdependenz (wechselseitige Abhängigkeit)
- > “Korrelation” umfasst Rangkorrelations-, Kontingenz oder Assoziationsanalyse (je nach Skalenniveau)
- > Beispiele für Korrelationskoeffizienten:
 - χ^2 -Wert (zwei nominalskalierte Variablen)
 - Cramer's V (zwei nominalskalierte Variablen)
 - Pearson's Korrelationskoeffizient r (zwei intervallskalierte Variablen)

Korrelation II



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

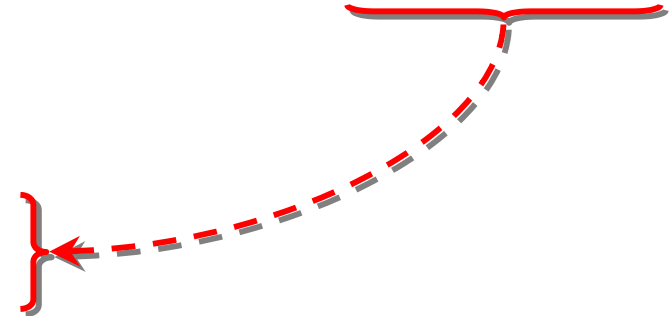
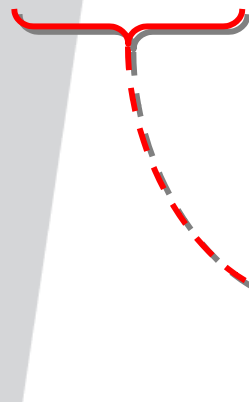
$$r = \frac{\text{COV}_{xy}}{s_x \times s_y}$$

$$\text{COV}_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \times (y_i - \bar{y})}{n}$$

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

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

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



Literatur zum Thema

- Buttler, Günther / Stroh, Reinhold (1992): Einführung in die Statistik; Reinbek: Rowohlt Verlag.
- Gehring, U. / Weins, C. (2002): Statistik für Politologen; 3. Auflage; Opladen: Westdeutscher Verlag.
- DIALEKT-Projekt (2002): Statistik interaktiv!; 2. Auflage; Berlin/Heidelberg/New York: Springer Verlag.
- Knoke, D. / Bohrnstedt, G.W. (2002): Statistics for the Social Data Analysis; 4th Edition; Ithasca: Peacock Publishers.
- v.d. Lippe, P. (1993): Deskriptive Statistik; Stuttgart, Jena: G. Fischer Verlag.
- Wonnacott, Th.H.; Wonnacott, R.J. (1997): Introductory Statistics, 5th Edition; New York, Toronto, Singapore: John Wiley & Sons.