Hochschule Kempten
University of Applied Sciences

Fakultät Betriebswirtschaft

# Statistik

Prof. Dr. Roland Jeske

# Literatur

**Lehrbücher:**

- **Jeske, R. (2017): Kochbuch der Quantitativen Methoden, Band 3: Statistik, 2. Auflage, Lulu (3. Auflage derzeit nur beim Autor erhältlich)**
- Bamberg et. al. (2009) Statistik 11. Aufl., Oldenbourg
- Fahrmair et. al. (2014): Statistik 12. Aufl., Springer
- Hartung et. al. (2013): Statistik 17. Aufl., Oldenbourg

**Übungsbuch:**

- **Jeske, R. (2018): Aufgabenbuch Statistik, Lulu** (2. erweiterte Auflage nur bei mir erhältlich)

Die Notationen in diesem Skript **(N:\\Skripten\BW\Jeske\Betriebswirtschaft)** entsprechen den Bezeichnungsweisen der fett gedruckten Literaturhinweise, also der Eigenpublikationen. Die weiteren Literaturhinweise stellen ergänzende Empfehlungen dar, die im Wesentlichen ähnliche Schwerpunkte setzen.

Hochschule Kempten
University of Applied Sciences

Fakultät Betriebswirtschaft

# Statistik

## Willkommen im neuen Sommersemester 2022

**Einige Anmerkungen:**

- Ein Skriptum ist nicht zitierfähig für wissenschaftliche Arbeiten!
- Dieses Skriptum unterliegt in Gänze dem Urheberrechtsgesetz. Das volle Copyright liegt bei Prof. Dr. Roland Jeske. Jegliche Verwendung, auch auszugsweise, bedarf der Zustimmung des Verfassers.
- Das vorliegende Skriptum besteht nahezu vollständig aus Auszügen der vorgenannten Eigenveröffentlichungen des Dozenten, eine jeweilige Quellenangabe wäre damit obsolet.
  Der Autor und Rechteinhaber Roland Jeske hat dazu dem Dozenten Roland Jeske das Recht übertragen, die Inhalte seiner Bücher frei in seinen Vorlesungen zu verwenden.

Ein Rechtsanspruch Dritter ergibt sich daraus nicht. Im Gegenteil muss jegliche anderweitige Verwendung Dritter durch den Autor der Lehrbücher bzw. des Skripts autorisiert werden.

Hochschule Kempten
University of Applied Sciences

Fakultät Betriebswirtschaft

# Hinweise zur Klausur
## „Open Book Klausur“

**Zugelassene Hilfsmittel:**

- Taschenrechner der Taschenrechner**serien**
  Casio FX-82, FX-85,FX-86,FX-87,
  Texas Instruments TI 30,
  SHARP EL-520, EL 531.
  Die Aufzählung ist **abschließend**!
- Gebundene Bücher (auch mit Anmerkungen) zum Thema. Eigene Skripten sind nur zulässig, soweit sie **gebunden** sind (Ringbücher, Schnellhefter, Leitzordner oder geklammerte Seiten sind **nicht** zugelassen)

**Weitere Hinweise:**

- Die Klausur besteht aus 6 Aufgaben sowie einer Zusatzaufgabe.
- Mit Bleistift oder roter Tinte bearbeitete Aufgaben können nicht in die Bewertung einbezogen werden.
- Die Klammerung der Klausur darf nicht entfernt werden.
- Lösungswege sind hinreichend zu dokumentieren.
- Die Klausur gilt als bestanden, wenn mindestens 50 von 100 Punkten erreicht wurden.

# Inhalt

# 1 Einführung

**Populäre Aussagen über Statistik:**

**“Ich traue keiner Statistik, die ich nicht selbst gefälscht habe”**

**(Winston Churchill?)**

**“Es gibt drei Arten von Lügen: Notlüge, gemeine Lüge und Statistik”**

**(Benjamin Disraeli)**

# 1 Einführung

**Meinung Ihres Dozenten:**

**“Statistik sollte wie die effiziente Zubereitung eines Glases frischgepressten Orangensaftes sein: Werfen Sie nicht das halbe Fruchtfleisch weg, d.h. verschwenden Sie keine Informationen Ihrer Daten. Aber vermeiden Sie auch Überinterpretationen, denn wenn Sie zu dem Fruchtfleisch auch noch die Schale auspressen, wird das Ergebnis im Saftladen wie auch in der Statistik zuweilen bitter…”**

**Roland Jeske**

# 1.1 Statistische “Vokabeln”

**Statistische Einheiten**
(auch: **Untersuchungseinheiten, Merkmalsträger**) Personen oder Objekte, an denen interessierende Größen gemessen/erhoben werden.

**Grundgesamtheit**
(auch: **Population**) Menge aller in Frage kommenden statistischen Einheiten

**Teilgesamtheit**
(auch: **Subpopulation)** Teilmenge der Grundgesamtheit

**Stichprobe**
Teilmenge der Grundgesamtheit (im Regelfall deutlich kleiner), Menge der statistischen Einheiten, die erfasst wurden.

**Merkmal(e)**
(auch: **Variablen**) Größe(n), die erfasst wurde(n)

**Merkmalsausprägung**
Wert eines Merkmals

# 1.1 Quantitativ versus Qualitativ

**Quantitative Merkmale** unterscheiden sich durch ihre **Größe,** es wird **gezählt** oder **gemessen.**
(z.B. Alter, Körpergröße, Einkommen, Umsatz,…)

**Qualitative Merkmale** unterscheiden sich nur durch ihre **Art und Weise** (z. B. Geschlecht, Nationalität, …)

# 1.1 Stetig versus Diskret

**Stetige Merkmale** können jeden **beliebigen Wert annehmen** (z.B. beliebig exakt messbare Größen)

**Diskrete Merkmale** haben **endlich viele** Ausprägungen oder auf der Achse mit Ausprägungen **Lücken** (z. B. Alter in Jahren)

**Quasi-stetige Merkmale:** sind eigentlich diskret, werden aber als stetig aufgefasst (z. B. Geldbeträge wie Umsatz oder Einkommen)

# 1.1 Skalierungen

**Nominalskalierte Merkmale** unterscheiden sich **nur** durch ihre **Art und Weise** und können in **keine sinnvolle Reihenfolge** gebracht werden (z.B. Geschlecht, Nationalität,…)

**Ordinalskalierte Merkmale** weisen hinsichtlich ihrer Ausprägungen eine **Reihenfolge** auf, die **Abstände** sind aber **nicht interpretierbar**
(z. B. Güteklassen, Leistungsklassen, Noten, Ratings,…)

**Kardinalskalierte (metrisch skalierte) Merkmale** unterliegen in ihren Merkmalsausprägungen einer **Reihenfolge und** die **Abstände der Merkmalsausprägungen sind interpretierbar.**
(z. B. Alter, Größe, Geldbeträge,…)

**Kardinalskala** kann noch unterteilt werden in **Verhältnisskala** und **Intervallskala,** Unterteilung ist aber für betriebswirtschaftliche Anwendungen nicht so relevant.

# 1.2 Daten

**1.2.1 Datenerhebung**

**1.2.2 Datenherkunft**

**1.2.3 Datenarten**

**1.2.4 Datenformate**

# 1.2.1 Datenerhebung

- **Totalerhebung** (Vollerhebung)
  Aus der Grundgesamtheit werden **alle** Untersuchungseinheiten erhoben.

- **Stichprobe**
  Aus der Grundgesamtheit wird nur ein Teil (häufig: ein sehr geringer Anteil) der Untersuchungseinheiten erhoben.

**Beispiele für Totalerhebungen:**
- Volkszählung (bis 1987)
- Inventuren (gesetzlich vorgeschrieben bis 1979)

**Beispiele für Stichproben:**
- Zensus (ab 2011)
- Inventuren (möglich ab 1980)

# 1.2.1 Datenerhebung

**Anwendung von Stichproben:**
In allen Bereichen:

- Sonntagsfrage (-> Kosten!)
- Pharmazeutische Forschung (-> ethische Aspekte)
- …

**Aber:**
Rückschluss von Stichprobe auf Grundgesamtheit muss zulässig sein!

# 1.2.2 Datenherkunft

- **Primärdaten:**
  Selbst erhobene Daten
- **Sekundärdaten**
  Daten aus fremden Bezugsquellen, die weiteren Anwendern zur Verfügung stehen

**Bekannte Sekundärdatenlieferanten:**

- Wirtschaftsforschungsinstitute
- Markt- und Meinungsforschungsinstitute
- Verbände
- Gewerkschaften
- Kommerzielle Anbieter

# 1.2.2 Datenherkunft

**Amtliche Statistik:**

- Statistisches Bundesamt (Destatis)
- Statistische Landesämter
- EUROSTAT
- Bundesagentur für Arbeit
- Bundesbank
- Kraftfahrtbundesamt

In Deutschland gehört die Kommunalstatistik **nicht** zur amtlichen Statistik, liefert aber wichtige Beiträge

# 1.2.3 Datenarten

**Querschnittsdaten:**
Zu einem bestimmten Zeitpunkt wird an Objekten der Grundgesamt ein (oder mehrere) Merkmale erhoben.
Schreibweise: $x_1, x_2, \ldots, x_n$

**Beispiele für Querschnittsdaten:**
Liegedauern von Patienten, Jahresumsätze 2018 von Unternehmen einer Branche, …

**Zeitreihendaten**
An einem einzigen Objekt wird in regelmäßigen Abständen ein (oder mehrere) Merkmal(e) erhoben.
Schreibweise: $y_1, y_2, \ldots, y_T$

**Beispiele für Zeitreihendaten**
DAX-Verlauf, Zeitreihe der Arbeitslosenzahlen, Herzfrequenzkurve,…

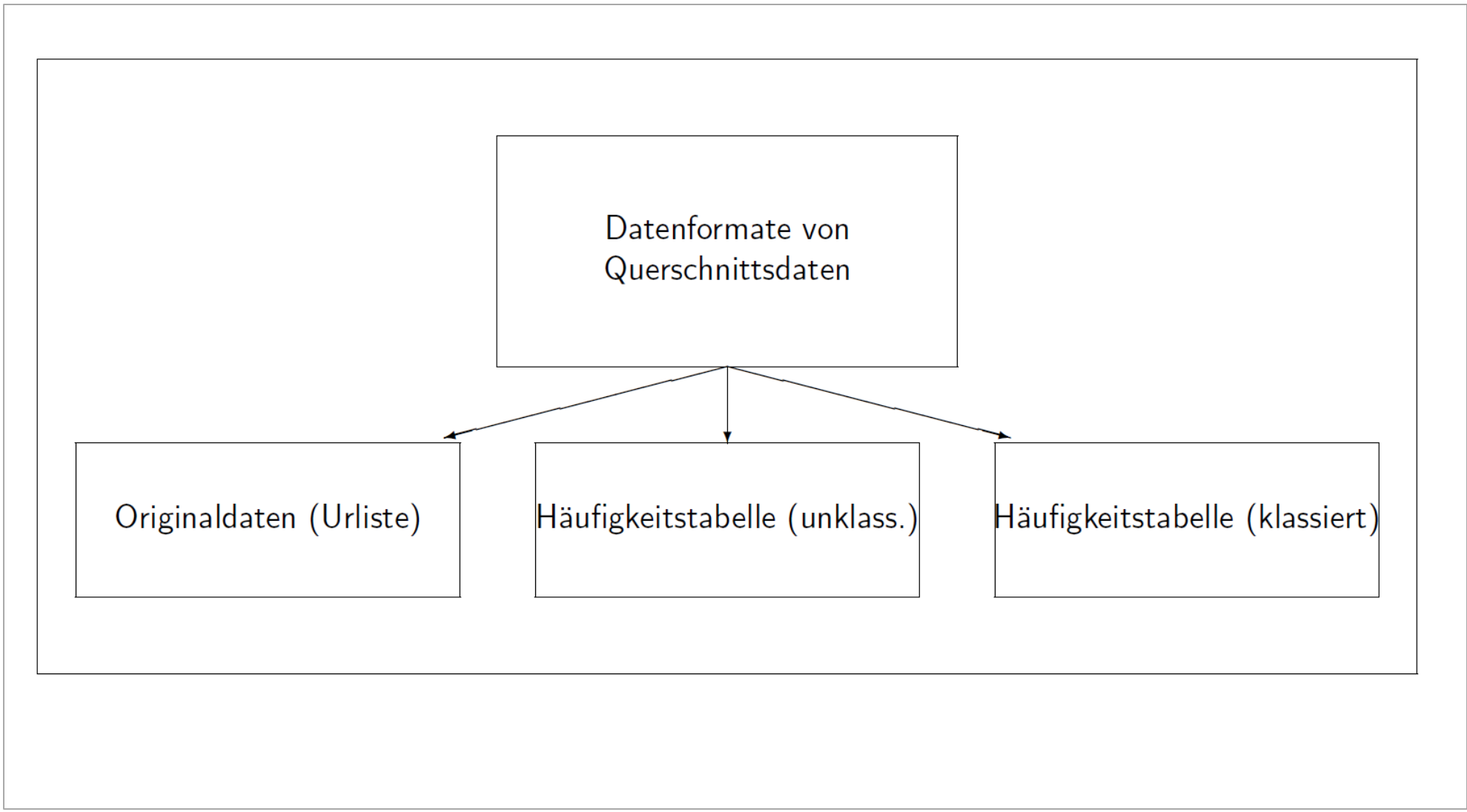
# 1.2.3 Datenarten

**Paneldaten (Kombination aus Querschnitts- und Zeitreihendaten)**
Zu bestimmten Zeitpunkt**en** wird an **mehreren** Objekten der Grundgesamt ein (oder mehrere) Merkmal(e) erhoben.

**Vorteil** von Panels: Auswertung in beide Richtungen möglich
**Nachteil** von Panels: Panelsterblichkeit

**Beispiel für Paneldaten :** Sozioökonomisches Panel (SOEP)

# 1.2.4 Datenformate



# 1.2.4 Datenformate

**Originaldaten (Urliste):**

Originaldaten (auch: Rohdaten, Einzelbeobachtungen, Urliste) liegen in Form einer Beobachtungsreihe

$$x_1, x_2, \ldots, x_n$$

vor.
Die Zahl $n$ heißt dabei Datenumfang, Stichprobenumfang oder Stichprobengröße.

# 1.2.4 Datenformate

**Beispiel 1.1 (Originaldaten):**

*Von zehn Patienten einer Notaufnahme wurde das Alter ermittelt:*

$$25,\ 21,\ 18,\ 37,\ 56,\ 89,\ 46,\ 23,\ 21,\ 34.$$

# 1.2.4 Datenformate

Ein erster Schritt der Aufarbeitung (auch zur Anwendung bestimmter Methoden), besteht darin, den Datensatz der Reihe nach zu **sortieren:**

**Geordneter Datensatz, geordnete Stichprobe**

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

# 1.2.4 Datenformate

**Fortsetzung Beispiel 1.1 (zur geordneten Stichprobe):**

Der geordnete Datensatz für das Patientenalter sieht wie folgt aus:

| $x_{(1)}$ | $x_{(2)}$ | $x_{(3)}$ | $x_{(4)}$ | $x_{(5)}$ | $x_{(6)}$ | $x_{(7)}$ | $x_{(8)}$ | $x_{(9)}$ | $x_{(10)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 18 | 21 | 21 | 23 | 25 | 34 | 37 | 46 | 56 | 89 |

# 1.2.4 Datenformate

## Häufigkeitstabelle ohne Klassierung (auch “gruppierte Daten”):

Gruppierte Daten liegen in Form einer **Häufigkeitstabelle** vor, in der zu den jeweiligen Merkmalsausprägungen $a_i$ die zugehörigen (absoluten) Häufigkeiten $n_i$ notiert werden:

| Merkmalsausprägung | Häufigkeit |
|---|---|
| $a_1$ | $n_1$ |
| $a_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $a_k$ | $n_k$ |
| Summe: | $n = n_1 + \ldots + n_k$ |

# 1.2.4 Datenformate

**Beispiel 1.2 (gruppierte Daten bzw. Häufigkeitstabelle ohne Klassierung):**

*Die zufällige Befragung von 128 Oktoberfestbesuchern zu ihrem „maßvollen“ Bierkonsum während des Festbesuchs ergab folgende Werte:*

| *Anzahl konsumierter Maß Bier* | *Häufigkeit* |
|---:|---:|
| *1* | *2* |
| *2* | *30* |
| *3* | *37* |
| *4* | *28* |
| *5* | *23* |
| *6* | *8* |

# 1.2.4 Datenformate

**Häufigkeitstabelle mit Klassierung (auch “klassierte Daten”):**

Klassierte Daten liegen in Form einer Häufigkeitstabelle vor, in der zur jeweiligen Merkmalsklasse $[x_i^u; x_i^o)$ die zugehörigen (absoluten) Häufigkeiten $n_i$ notiert werden:

| Merkmalsklasse | Häufigkeit |
|---|---|
| $[x_1^u; x_1^o)$ | $n_1$ |
| $[x_2^u; x_2^o)$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $[x_k^u; x_k^o)$ | $n_k$ |
| Summe: | $n = n_1 + \ldots + n_k$ |

# 1.2.4 Datenformate

**Beispiel 1.3 (klassierte Daten, Häufigkeitstabelle mit Klassierung):**

*Die Umsatzverteilung (in Mio. € ohne MWSt.) deutscher Apotheken im Jahr 2012 ist wie folgt gegeben:*

| *Klasse* | $r_i$ | *Klasse* | $r_i$ | *Klasse* | $r_i$ | *Klasse* | $r_i$ |
|---|---|---|---|---|---|---|---|
| $[0; 0,75)$ | *0,070* | $[1,75; 2)$ | *0,112* | $[3; 3,25)$ | *0,022* | $[4,25; 4,5)$ | *0,005* |
| $[0,75; 1)$ | *0,096* | $[2; 2,25)$ | *0,087* | $[3,25; 3,5)$ | *0,016* | $[4,5; 4,75)$ | *0,003* |
| $[1; 1,25)$ | *0,136* | $[2,25; 2,5)$ | *0,051* | $[3,5; 3,75)$ | *0,014* | $[4,75; 5)$ | *0,004* |
| $[1,25; 1,5)$ | *0,151* | $[2,5; 2,75)$ | *0,049* | $[3,75; 4)$ | *0,010* | $[5; \infty)$ | *0,014* |
| $[1,5; 1,75)$ | *0,120* | $[2,75; 3)$ | *0,033* | $[4; 4,25)$ | *0,007* | | |

*Quelle: ABDA – Bundesvereinigung Deutscher Apothekerverbände*

# 1.2.4 Datenformate

**Klassierungsregeln:**

- Vellemann: $[2\sqrt{n}]$
- Dixon/Kronmal: $[10 \log_{10} n]$
- Vielfach ist „persönliches“ Augenmaß gefragt!

**Sinnvoll:**

- Gleiche Klassenbreiten
- Keine offenen Endklassen

**Aber:** manchmal unumgänglich!

# 1.2.4 Datenformate

Wenn Merkmale mindestens ordinal skaliert sind, ergänzt man in der Häufigkeitstabelle noch die kumulierten Häufigkeiten:

- Kumulierte absolute Häufigkeit: $\sum_{j=1}^{i} n_j$
- Kumulierte relative Häufigkeit: $\sum_{j=1}^{i} r_j = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{j=1}^{i} n_j$

# 1.2.4 Datenformate

**Beispiel 1.4 (klassierte Daten, Häufigkeitstabelle mit Klassierung):**

*Studentin Witta Miene bedient die Salattheke eines Restaurants, bei dem gramm-genau der Preis ermittelt wird. Dies führt zu sehr ungeraden Trinkgeldern, da die Zahlbeträge von den Kunden vielfach aufgerundet werden. Zwanzig zufällig ausgewählte Trinkgeldbeträge (in €) ergaben folgende Werte:*

| | | | | |
|---|---|---|---|---|
| 3,20 | 2,79 | 4,07 | 4,13 | 2,63 |
| 1,34 | 2,51 | 0,07 | 0,36 | 2,69 |
| 1,43 | 4,78 | 0,82 | 4,09 | 1,22 |
| 4,04 | 4,69 | 4,47 | 3,11 | 1,73 |

*Klassieren Sie die Daten, indem Sie Klassen mit 1 € Breite wählen.*

**Lösung: siehe Vorlesung**

# 2 Einfache Grafiken

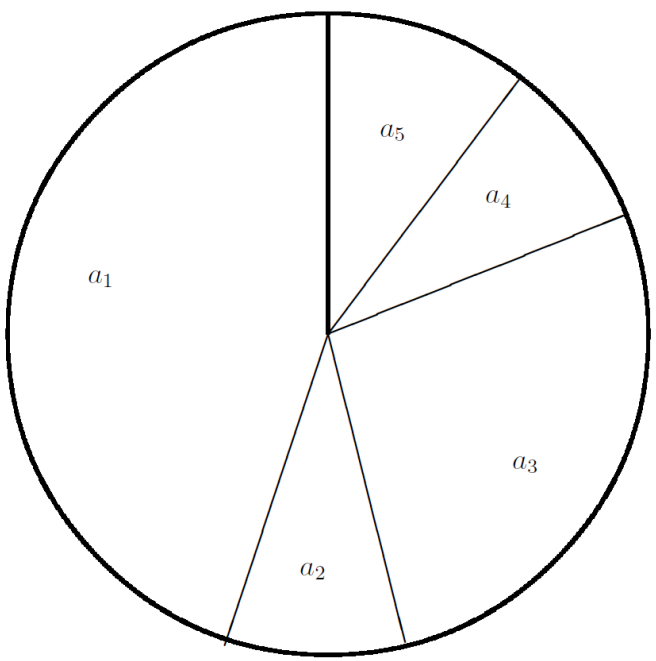
Eine **erste Aufbereitung** der Daten erfolgt vielfach durch eine **Grafik:**

- Kreisdiagramm
- Blockdiagramm
- Stabdiagramm
- Polygonzug (Liniendiagramm)
- Histogramm

# 2 Einfache Grafiken

**Kreisdiagramm (Tortendiagramm):**

**Gemäß der Häufigkeit werden Kreissektoren (Tortenstücke) eingezeichnet:**



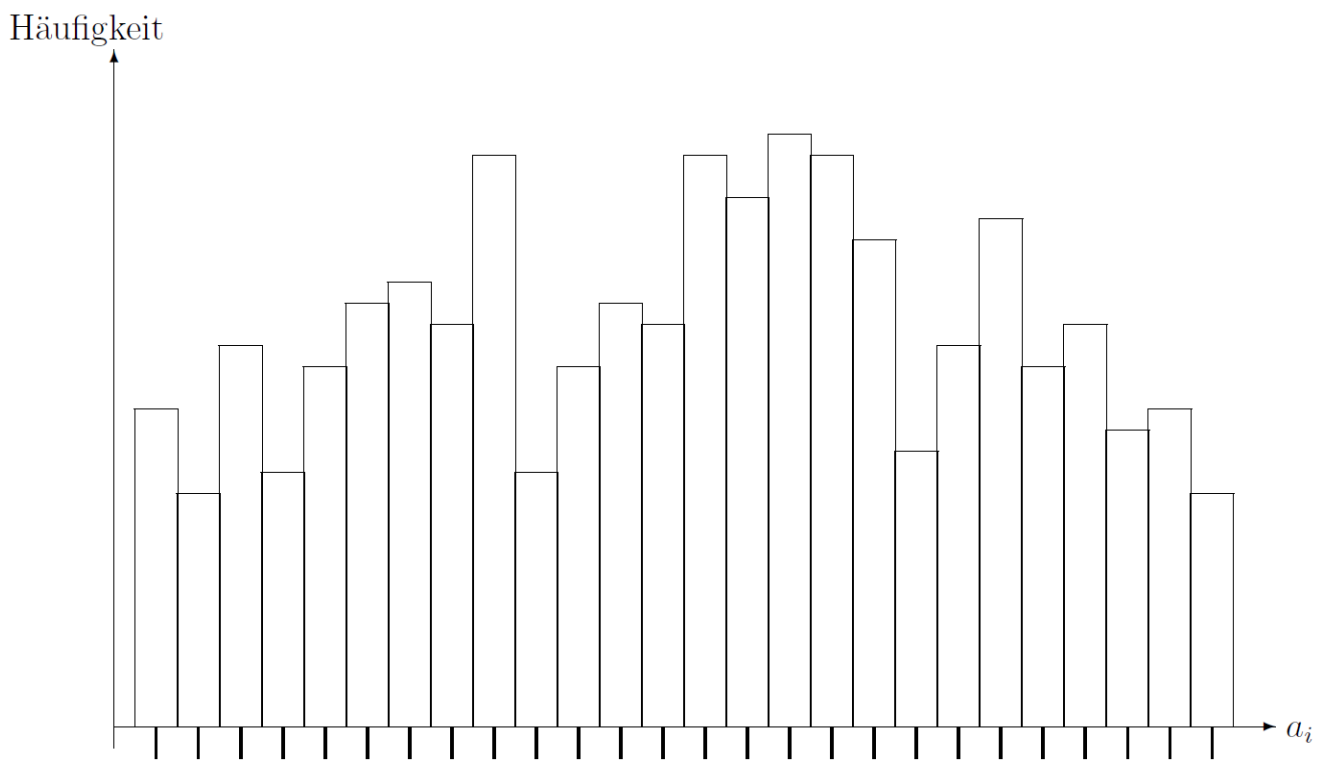

**Die Winkel der Kreissegmente**

**berechnen sich über** $\alpha_i = r_i \cdot 360°$

# 2.1 Blockdiagramm

**Blockdiagramm (Säulen- oder Balkendiagramm):**

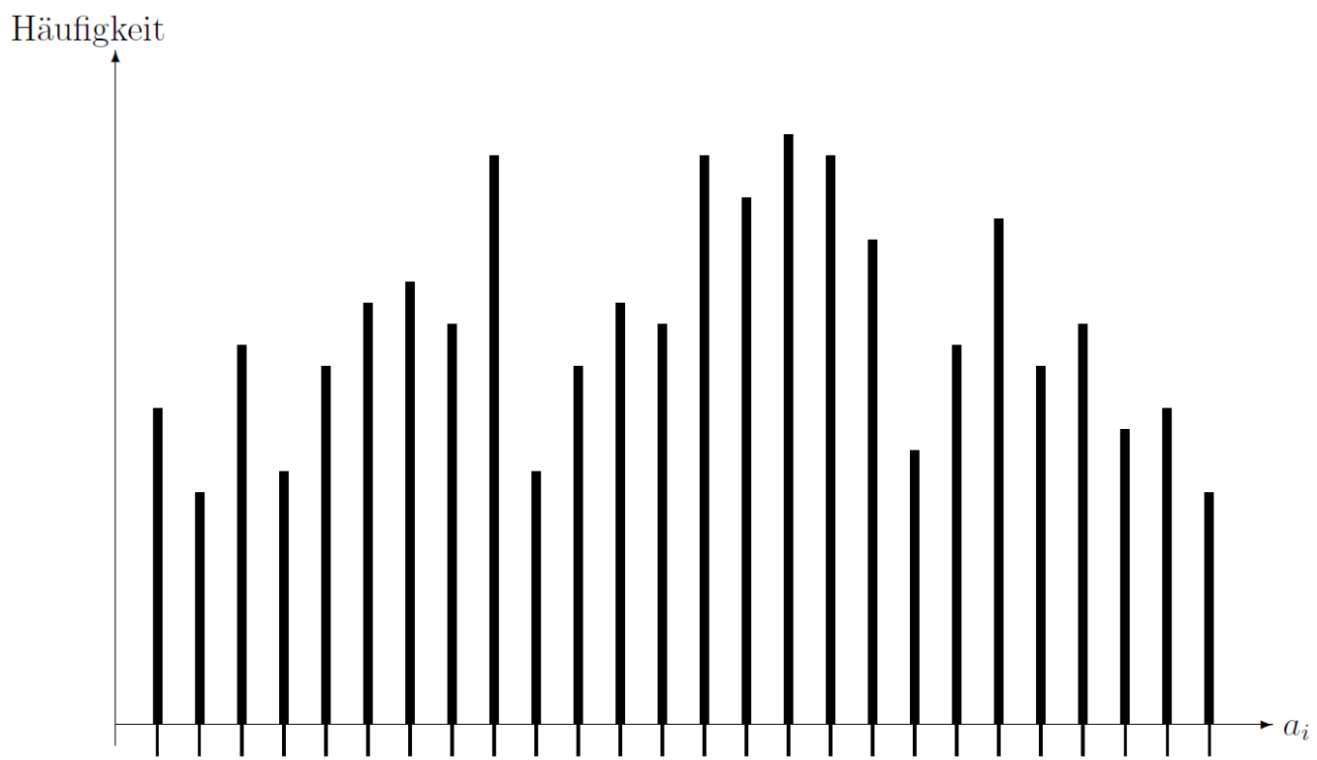
**Gemäß der relativen oder absoluten Häufigkeit werden die Blöcke gezeichnet:**



# 2.2 Stabdiagramm

**Stabdiagramm:**

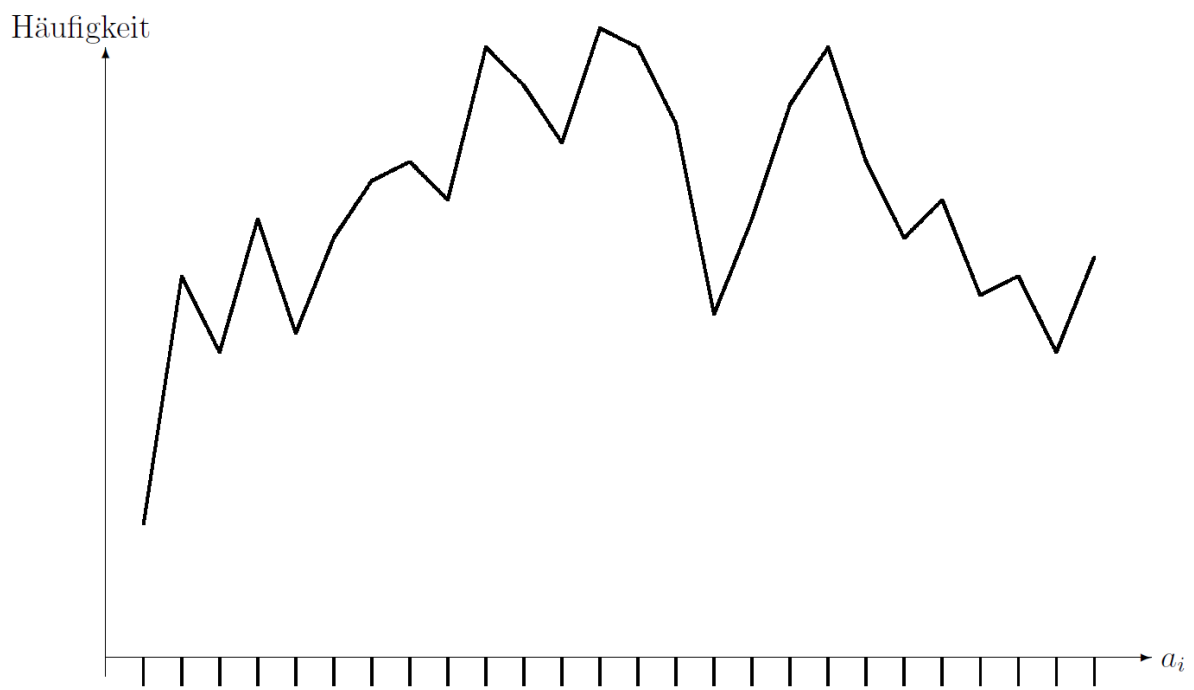
**Gemäß der relativen oder absoluten Häufigkeit werden die Stäbe gezeichnet:**



# 2.3 Liniendiagramm

**Polygonzug (Liniendiagramm):**

**Gemäß der relativen oder absoluten Häufigkeit werden aufeinanderfolgende Endpunkte mit Verbindungslinien versehen:**



# 2.4 Histogramm

**Histogramm:**

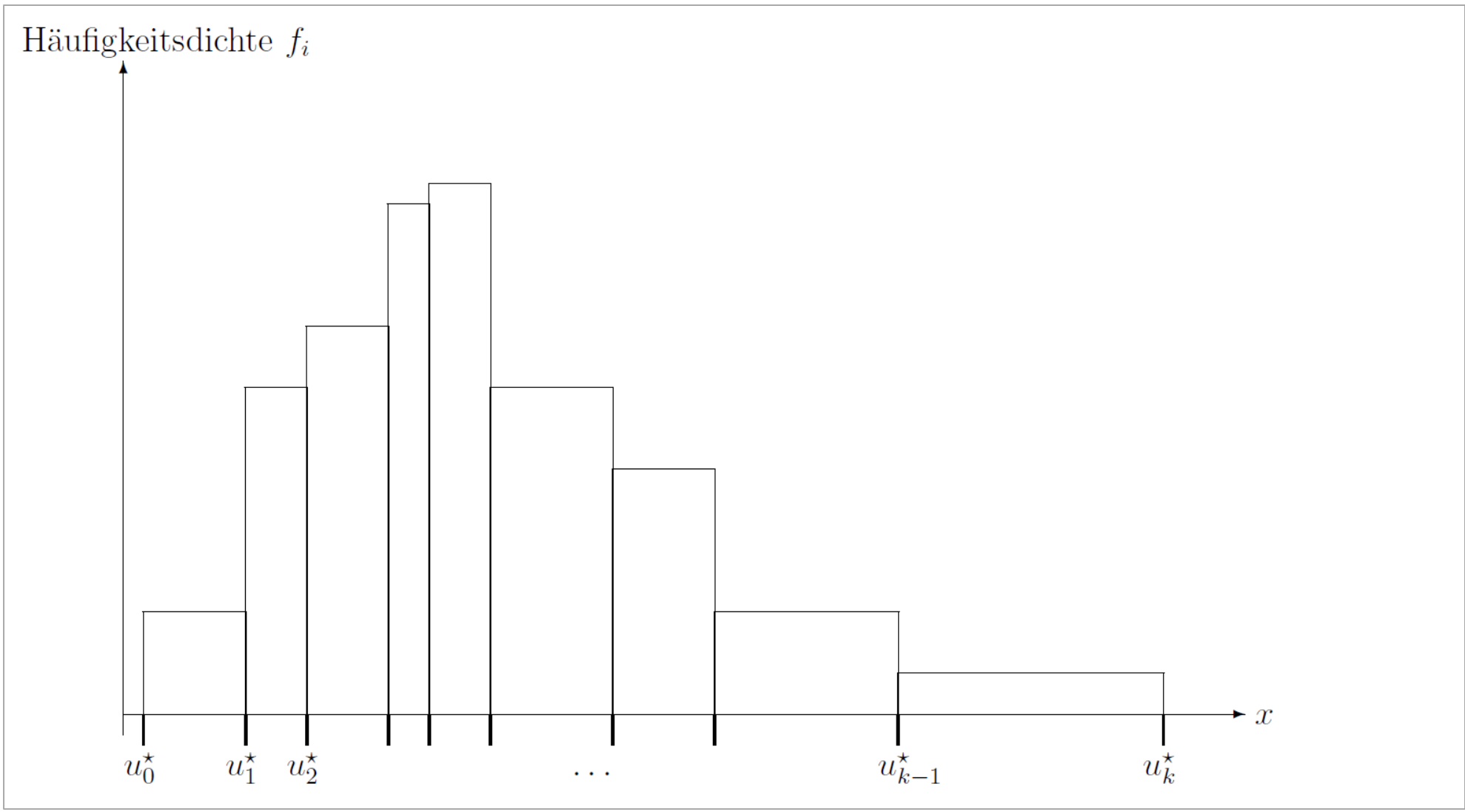
- **Wird insbesondere bei klassierten Daten verwendet.**
- **Achtung!** Im Gegensatz zum Blockdiagramm wird die **Fläche, nicht die Höhe proportional zur Häufigkeit gezeichnet.** Damit gilt für die Fläche des Blockes:

**Die Höhe wird auch Häufigkeitsdichte $f_i$ genannt:**

$$\text{Blockfläche} = \text{Grundseite} \cdot \text{Höhe} = \Delta_i \cdot \frac{r_i}{\Delta_i} = r_i$$

$$f_i = \frac{r_i}{\Delta_i}$$

# 2.4 Histogramm



# 3 Lagemaße

**Sinn eines Lagemaßes:**

- beschreibt das **Zentrum** eines Datensatzes
- **Achtung!** Nur sinnvoll, wenn es sich um eine **unimodale (eingipflige) Verteilung** handelt. Bei mehrgipfligen Verteilungen ist eher eine Grafik angebracht als die Daten auf einen einzigen Wert zu reduzieren!

# 3.1 Arithmetisches Mittel

**Arithmetisches Mittel:**

**Idee: berechne Durchschnittswert**

- **Einfach zu berechnen**
- **populär**

**Aber**

- **Ausreißerempfindlich**
- **Verlangt kardinales Skalenniveau**

# 3.1 Arithmetisches Mittel

**Arithmetisches Mittel (Originaldaten, Urliste):**

$$\overline{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

**Arithmetisches Mittel (Häufigkeitstabelle ohne Klassierung):**

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i a_i = \sum_{i=1}^{k} \underbrace{\frac{n_i}{n}}_{=r_i} a_i = \sum_{i=1}^{k} r_i a_i$$

**Arithmetisches Mittel (Häufigkeitstabelle mit Klassierung):**

$$\overline{x} = \frac{n_1 \cdot m_1 + n_2 \cdot m_2 + \ldots n_k \cdot m_k}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i m_i = \sum_{i=1}^{k} \underbrace{\frac{n_i}{n}}_{=r_i} m_i$$

wobei $\boldsymbol{m_i}$ **Klassenmitte** der $i$-ten Klasse ist.

# 3.1 Arithmetisches Mittel

**Beispiele:**

Berechnen Sie das arithmetische Mittel

1. für Beispiel 1.1 (“Patientendaten”),
2. für Beispiel 1.2 (“Oktoberfestdaten”),
3. für Beispiel 1.4 (“Trinkgelddaten”).
4. Welche Schwierigkeiten treten auf, wenn Sie das arithmetische Mittel für Beispiel 1.3 (“Apothekendaten”) berechnen möchten?

**Lösung: siehe Vorlesung**

# 3.2 Median

**Idee:** berechne **Mitte** (nicht Mitte**l**) eines Datensatzes, d.h. der Median teilt die Daten in **zwei Hälften** ein, die jeweils 50% einnehmen.

- Mitunter schwieriger zu berechnen als arithmetisches Mittel
- Ausreißerunempfindlich
- Bereits ab ordinalem Skalenniveau einsetzbar

**Aber**

- Bei kardinalen Daten geht Information verloren!

# 3.2 Median

**Median (Originaldaten):**

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n \text{ gerade} \end{cases}$$

**Median (gruppierte Daten):**

$$\tilde{x} = \begin{cases} a_i & \text{falls } \sum_{j=1}^{i-1} r_j < 0,5 \text{ und } \sum_{j=1}^{i} r_j > 0,5 \\ \frac{1}{2}\left(a_i + a_{i+1}\right) & \text{falls } \sum_{j=1}^{i-1} r_j < 0,5 \text{ und } \sum_{j=1}^{i} r_j = 0,5 \end{cases}$$

**D.h., suche im Wesentlichen die Merkmalsausprägung, bei der die kumulierte relative Häufigkeit erstmals den Wert 0,5 überschreitet.**

# 3.2 Median

**Median (Klassierte Daten):**

$$\tilde{x} = x_i^u + \frac{0,5 - \sum_{j=1}^{i-1} r_j}{r_i} (x_i^o - x_i^u)$$

falls $i$ die **erste Klasse** ist, in der

$$\sum_{j=1}^{i} r_j \geq 0,5$$

gilt.

# 3.2 Median

**Beispiele:**

Berechnen Sie den Median

1. für Beispiel 1.1 (“Patientendaten”)
2. für Beispiel 1.2 (“Oktoberfestdaten”)
3. für Beispiel 1.3 (“Apothekendaten”)
4. für Beispiel 1.4 (“Trinkgelddaten”)

**Lösung:** siehe Vorlesung

# 3.3 Quantile

Die Idee des Medians lässt sich **verallgemeinern**:

- Der Median teilt den Datensatz in zwei Hälften, die jeweils 50% der Daten umfassen.
- Nun: Suche Wert, der Datensatz in beliebige Anteile teilt, etwa im Verhältnis 90% zu 10%.

Gezählt wird immer der Anteil, der vor dem Quantil liegt, dieser wird als Index notiert:

- $x_p$ ist derjenige Wert, den der Anteil $p$ der Daten nicht überschreitet und der Anteil $(1-p)$ der Daten nicht unterschreitet.

# 3.3 Quantile

**Quantil (Originaldaten):**

$$x_p = \begin{cases} x_{(\lceil np \rceil)} & np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & np \in \mathbb{N}, \end{cases}$$

wobei $\lceil x \rceil$ heißt, dass auf die nächst höhere ganze Zahl aufzurunden ist.

**Quantil (gruppierte Daten):**

$$x_p = \begin{cases} a_i & \text{falls } \sum_{j=1}^{i-1} r_j < p \text{ und } \sum_{j=1}^{i} r_j > p \\ \frac{1}{2}\left(a_i + a_{i+1}\right) & \text{falls } \sum_{j=1}^{i-1} r_j < p \text{ und } \sum_{j=1}^{i} r_j = p \end{cases}$$

# 3.3 Quantile

**Quantil (klassierte Daten):**

$$x_p = x_i^u + \frac{p - \sum_{j=1}^{i-1} r_j}{r_i} (x_i^o - x_i^u)$$

falls $i$ die **erste Klasse** ist, in der

$$\sum_{j=1}^{i} r_j \geq p$$

gilt.

# 3.3 Quantile

**Besondere Quantile:**

- **Quartile:** In diesem Fall wird der Datensatz **geviertelt.** $x_{0,25}$ heißt **unteres Quartil.** $x_{0,75}$ heißt **oberes Quartil.** $x_{0,5}$, also das **“mittlere” Quartil,** stellt gerade den **Median** dar.
- **Perzentile:** Bei Quantilen, die beliebige **Prozentanteile** beschreiben, spricht man auch von Perzentilen, etwa das 5%-Perzentil $x_{0,05}$

In der Statistik werden häufig, auch für die Zusammensetzung weiterer Maßzahlen, **Quartile** verwendet.

# 3.3 Quantile

**Beispiele:**

Berechnen Sie die Quartile

1. für Beispiel 1.1 (“Patientendaten”)
2. für Beispiel 1.2 (“Oktoberfestdaten”)
3. für Beispiel 1.3 (“Apothekendaten”)
4. für Beispiel 1.4 (“Trinkgelddaten”)

**Lösung:** siehe Vorlesung

# 3.4 Modus

**Modus:**

**= häufigste Merkmalsausprägung (falls eindeutig)**

- Einfach zu berechnen
- Bereits ab nominalem Skalenniveau einsetzbar

**Aber**

- Vergleichsweise primitives Lagemaß, sollte bei höherem Skalenniveau nicht alleiniges Kriterium sein!

# 3.4 Modus

**Modus (Originaldaten):** Daten gruppieren, dann siehe dort…

**Modus (gruppierte Daten):**

$$\overline{x}_M = a_i \quad \text{mit } r_i = \max_j \{r_j\}$$

Der Modus ist demnach die Merkmalsausprägung, die am häufigsten vorkommt. Ist dieser Wert nicht eindeutig, verzichtet man auf die Angabe des Modus'.

**Modus (Häufigkeitstabelle mit Klassierung):**

$$\overline{x}_M = m_i \quad \text{mit } \frac{r_i}{\Delta_i} = \max_j \left\{ \frac{r_j}{\Delta_j} \right\}$$

Der Modus ist demnach die Klassenmitte derjenigen Klasse mit der größten Häufigkeitsdichte.

# 3.4 Modus

**Beispiele:**

Berechnen Sie den Modus

1. für Beispiel 1.1 (“Patientendaten”)
2. für Beispiel 1.2 (“Oktoberfestdaten”)
3. für Beispiel 1.3 (“Apothekendaten”)
4. für Beispiel 1.4 (“Trinkgelddaten”)

**Lösung: siehe Vorlesung**

# 3.5 Gewogenes arithmetisches Mittel

**Bislang:** gleiche Gewichtung aller Beobachtungen (mit **1/n**)

**Idee: gezielte unterschiedliche Gewichtung** von Beobachtungen:

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i$$

mit

$$w_i \geq 0 \qquad \text{und} \qquad \sum_{i=1}^{n} w_i = 1$$

# 3.5 Gewogenes arithmetisches Mittel

**Beispiel:**

*An der Finalrunde eines Tanzturniers nehmen sechs Paare, teil, deren Leistungen von fünf Wertungsrichtern $R_1, \ldots, R_5$ bewertet werden. Jede Note zwischen 1 und 6 wird von jedem Wertungsrichter genau einmal vergeben, dabei steht die 1 für die beste und die 6 für die schlechteste Bewertung:*

| *Paar* | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 1 | 6 |
| 2 | 5 | 4 | 5 | 5 | 5 |
| 3 | 4 | 5 | 3 | 4 | 4 |
| 4 | 2 | 1 | 2 | 3 | 2 |
| 5 | 6 | 6 | 6 | 6 | 1 |
| 6 | 3 | 3 | 4 | 2 | 3 |

*Zur Gesamtbeurteilung sollen Mittelwerte berechnet werden.*
*Erstellen Sie eine Rangfolge für die Paare, indem Sie folgende Mittelwerte berechnen:*

*a) das arithmetische Mittel,*

*b) das gewogene arithmetische Mittel berechnen, bei dem jeweils die größte und die kleinste Beobachtung gestrichen werden und die verbleibenden Beobachtungen jeweils das Gewicht $\frac{1}{3}$ erhalten.*

# 3.5 Gewogenes arithmetisches Mittel

**Weitere wichtige Anwendung des gewogenen arithmetischen Mittels:**

- **Mittelung von Verhältniszahlen (Quotienten), wenn die Nennerverteilung bekannt ist.**

**Beispiel:**

Der chinesische Radprofi Do Ping fährt in einem Rennen 2,5 Stunden lang mit einer konstanten Geschwindigkeit von 42 km/h. Anschließend fährt er 3 Stunden lang mit einer Geschwindigkeit von 31 km/h.

Mit welcher Durchschnittsgeschwindigkeit befährt er die Gesamtstrecke?

# 3.5 Gewogenes arithmetisches Mittel

**Lösung 1:**

**Offenbar gilt**

- In den ersten **2,5 Stunden** legt Do Ping insgesamt **105 km** zurück.
- In der zweiten Etappe fährt er **3 Stunden** lang und legt **93 km** zurück.

⇒ Er fährt **insgesamt 5,5 Stunden** und legt dabei **198 km** zurück.

Die **Durchschnittsgeschwindigkeit** beträgt somit $\frac{198\ km}{5{,}5\ h} = 36\frac{km}{h}$

**Lösung 2 (über gewogenes arithmetisches Mittel):**

$$\overline{x} = \frac{2{,}5}{3+2{,}5} \cdot 42 + \frac{3}{3+2{,}5} \cdot 31 = 36$$

# 3.5 Gewogenes arithmetisches Mittel

**Allgemein:**

**Sind Verhältniszahlen** $V_i = \frac{Z_i}{N_i}$ **für Teilgesamtheiten** $i = 1, \ldots k$ **bekannt, so ergibt sich für die Verhältniszahl der Grundgesamtheit:**

$$V = \frac{Z_1 + \ldots + Z_k}{N_1 + \ldots + N_k}$$

**bzw., wenn nur die Nennerverteilung bekannt ist:**

$$V = \sum_{i=1}^{k} \frac{N_i}{N} V_i$$

D.h., bei bekannter Verteilung der Nennergrößen ergibt sich das Aggregat von Verhältniszahlen als gewogenes arithmetisches Mittel der einzelnen Verhältniszahlen.

# 3.6 Geometrisches Mittel

**Geometrisches Mittel (Originaldaten):**

$$\overline{x}_g = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

**Geometrisches Mittel (gruppierte Daten):**

$$\overline{x}_g = \sqrt[n]{a_1^{n_1} \cdot \ldots \cdot a_k^{n_k}}$$

**Geometrisches Mittel für klassierte Daten eher unüblich, da zu viel Information durch die Klassierung verloren geht.**

# 3.6 Geometrisches Mittel

**Anwendung des geometrischen Mittels: Mittelung von Wachstumsraten:**

**Beispiel:** Der Wert einer Aktie steigt im ersten Jahr um 30%, im zweiten Jahr fällt er um 20%.

→ **Populärer Irrtum:** Das durchschnittliche jährliche Wachstum beträgt 5%.

**Tatsächlich liegt das durchschnittliche Wachstum unterhalb von 2%!**

# 3.6 Geometrisches Mittel

**Lösung mit Hilfe der Wachstumsfaktoren:**

Ausgehend vom Anfangskapital $K_0$ erhält man das Endkapital nach zwei Jahren über

$$K_2 = K_0(1 + 0{,}3)(1 - 0{,}2) = K_0 \cdot 1{,}3 \cdot 0{,}8$$

Hätte sich das Kapital gleichmäßig mit dem Zinssatz $i$ verzinst, so müsste gelten:

$$K_2 = K_0(1 + i)^2$$

Damit erhält man $i$ durch das geometrische Mittel wie folgt:

$i = \sqrt[2]{1{,}3 \cdot 0{,}8}$-1=0,0198=1,98%.

**Achtung!** Sie müssen das geometrische Mittel auf die **Wachstumsfaktoren**, nicht auf die Wachstumsraten anwenden!

# 3.7 Harmonisches Mittel

Anblick des **harmonischen Mittels** ist zunächst gewöhnungsbedürftig…

**Harmonisches Mittel (Originaldaten):**

$$\overline{x}_h = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

**Harmonisches Mittel (gruppierte Daten):**

$$\overline{x}_h = \frac{1}{\frac{1}{n}\sum_{i=1}^{k}\frac{n_i}{a_i}} = \left(\frac{1}{n}\sum_{i=1}^{k}\frac{n_i}{a_i}\right)^{-1} = \left(\sum_{i=1}^{k}\frac{n_i}{n}\frac{1}{a_i}\right)^{-1}$$

**Harmonisches Mittel für klassierte Daten eher unüblich.**

# 3.7 Harmonisches Mittel

**Anwendung: Mittelung von Verhältniszahlen (Quotienten),** wenn die **Zählerverteilung** bekannt ist

**Beispiel:**

Die Radsportlerin Anna Bolika fährt 90 km lang mit einer konstanten Geschwindigkeit von 36km/h. Anschließend fährt sie 40 km lang mit einer konstanten Geschwindigkeit von 32 km/h.

Mit welcher **Durchschnittgeschwindigkeit** befährt sie die Gesamtstrecke?

# 3.7 Harmonisches Mittel

**Lösung 1: Offenbar gilt:**

- Für die ersten **90 km** benötigt sie $\frac{90}{36}$ =**2,5** Stunden.
- Für die weiteren **40 km** benötigt sie $\frac{40}{32}$ =**1,25** Stunden.

⇒ Sie fährt insgesamt **3,75 Stunden** und legt dabei **130 km** zurück.

Die **Durchschnittsgeschwindigkeit** beträgt somit $\frac{130\ km}{3{,}75\ h} = 34,\overline{6}\ \frac{km}{h}$

**Lösung 2 (über gewogenes harmonisches Mittel)**

$$\overline{x}_h = \left(\frac{90}{90+40} \cdot \frac{1}{36} + \frac{40}{90+40} \cdot \frac{1}{32}\right)^{-1} = 34,\overline{6}$$

# 3.7 Gewogenes harmonisches Mittel

**Allgemein:**

**Sind Verhältniszahlen** $V_i = \frac{Z_i}{N_i}$ **für Teilgesamtheiten** $i = 1, \ldots k$

**bekannt, so ergibt sich für die Verhältniszahl der Grundgesamtheit:**

$$V = \frac{Z_1 + \ldots + Z_k}{N_1 + \ldots + N_k}$$

**oder**

$$V = \left( \sum_{i=1}^{k} \frac{Z_i}{Z} V_i^{-1} \right)^{-1}$$

D.h., bei bekannter Verteilung der Zählergrößen ergibt sich das Aggregat von Verhältniszahlen als gewogenes harmonisches Mittel der einzelnen Verhältniszahlen.

# 3.8 Abschließende Bemerkungen

Die nachfolgende Tabelle gibt Aufschluss darüber, welches Maß bei welchem Skalenniveau eingesetzt ($\surd$) bzw. nicht eingesetzt (—) werden darf:

| Maßzahl | Skalenniveau | | |
|---|---|---|---|
| | kardinal | ordinal | nominal |
| arithmetisches Mittel | $\surd$ | — | — |
| gewogenes arithm. Mittel | $\surd$ | — | — |
| geometrisches Mittel | $\surd$ | — | — |
| harmonisches Mittel | $\surd$ | — | — |
| Median | $\surd$ | $\surd$ | — |
| Quantile | $\surd$ | $\surd$ | — |
| Modus | $\surd$ | $\surd$ | $\surd$ |

# 3.8 Abschließende Bemerkungen

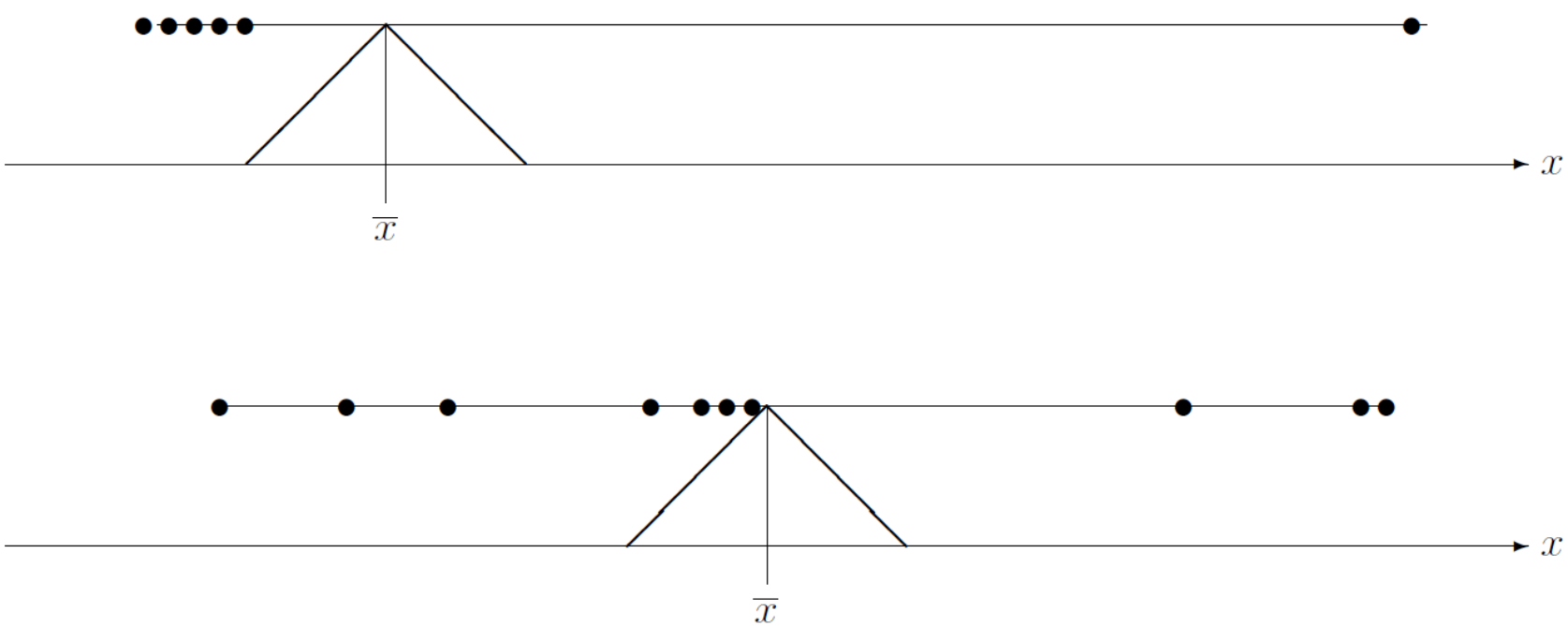
**Konsequenzen:**

- **Geometrisches** und **harmonisches Mittel** haben **besondere Anwendungsgebiete.**
- Bei **nominalen** Daten kann nur der **Modus** verwendet werden.
- Bei höherwertigen Skalen ist der Modus primitiv und sollte nicht als alleiniges Lagemaß berechnet werden.
- Bei kardinalen Daten stellt sich vielfach die Frage, ob arithmetisches Mittel oder Median eingesetzt werden sollten…

# 3.8 Arithmetisches Mittel oder Median?

**Einerseits:**

**Das arithmetische Mittel ist aufgrund seiner Schwerpunkteigenschaft ausreißerempfindlich, der Median nicht:**



# 3.8 Arithmetisches Mittel oder Median?

**Beispiel:**

Man betrachte die nachfolgenden **3 Teilpopulationen:**

| | |
|---|---|
| Pop. I: | 1, 2, 5, 5, 6, 7, 11, 16, 20, 33, 37 |
| Pop. II: | 3, 3, 16, 25, 33 |
| Pop. III: | 4, 4, 7, 11, 15, 19, 25, 33, 35 |

Welche Schlüse können Sie für die Lagemaße ziehen, wenn Sie jeweils arithmetisches Mittel, Median und Modus für die Teilpopulationen und die Gesamtpopulation berechnen?

# 3.8 Arithmetisches Mittel oder Median?

**Andererseits:**

Der Median einer Gesamtpopulation lässt sich nicht aus den Medianen der Teilpopulationen berechnen, wie ein Mittelwert. Grund ist, dass der Median auf die Abstandsmessung (kardinale Eigenschaft) verzichtet (es geht Information verloren!).

Diese Nichtaggregierbarkeit kann zu eigentümlichen Ergebnissen in der Praxis führen, etwa bei der Berechnung der Armutsgefährdungsgrenze (siehe Kapital 12)

**Fazit:**

In der Praxis ist die Frage nach dem **sinnvollsten** Lagemaß gar nicht so einfach zu beantworten…

# 3.9 Lineare Transformationen

Wie verhalten sich Lagemaße, wenn Daten **linear transformiert** werden, d.h.

$$y_i = a + bx_i$$

Dann gilt für …

- das arithmetische Mittel: $\overline{y} = a + b\overline{x}$
- den Median: $\tilde{y} = a + b\tilde{x}$
- den Modus: $\overline{y}_M = a + b\overline{x}_M$
- ein Quantil:

$$y_p = \begin{cases} a + bx_p & \text{falls } b > 0 \\ a + bx_{1-p} & \text{falls } b < 0 \end{cases}$$

# 3.9 Lineare Transformationen

**Speziell für**

$$y_i = bx_i$$

gilt für das

- **geometrisches Mittel:** $\overline{y}_g = b\overline{x}_g$
- **harmonisches Mittel:** $\overline{y}_h = b\overline{x}_h$

# 3.9 Lineare Transformationen

**Beispiel: (Oktoberfestdaten)**

Der Bierkonsum soll nicht in Maßkrügen sondern in Litern gemessen werden.

Mithin hat sich herausgestellt, dass ein Maßkrug auf der Wies'n nur mit 0,9 l gefüllt wird.

Berechnen Sie arithmetisches Mittel, Median und Modus für den Literkonsum.

**Lösung:**

Für die Transformation $y_i = a + bx_i = 0 + 0{,}9x_i$ erhält man mit den zuvor berechneten Lagemaßen:

$$\bar{y} = 0{,}9\bar{x} = 0{,}9 \cdot 3{,}5 = 3{,}15,$$
$$\tilde{y} = 0{,}9 \cdot \tilde{x} = 0{,}9 \cdot 3 = 2{,}7,$$
$$\overline{y_M} = 0{,}9 \cdot \overline{x_M} = 0{,}9 \cdot 3 = 2{,}7.$$

# 4 Streuungsmaße

Wie misst man Streuung?

**Erste Idee:** $$d_x = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|$$

**= mittlere absolute Abweichung vom Mittelwert**

**Problem:**

- Vielfach möchte man Streuung minimieren (siehe etwa Regressionsmodell)
- Minimieren heißt meistens **ableiten**
- **Aber:** Betragsfunktion nicht differenzierbar!

**Ausweg:** Betrachte **quadratische Abweichungen!**

# 4.1 Varianz

**Varianz:**

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

**Stichprobenvarianz:**

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

**In der Praxis:** Verwende $\sigma^2$, wenn Varianz einer **Totalerhebung** berechnet wird und $s^2$, wenn Varianz einer **Stichprobe** berechnet wird.

# 4.1 Varianz

**Für die unterschiedlichen Datenformate ergibt sich damit:**

**Varianz (Originaldaten)**

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

**Varianz (gruppierte Daten)**

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (a_i - \overline{x})^2$$

**Varianz (klassierte Daten)**

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (m_i - \overline{x})^2$$

# 4.1 Varianz

**Entsprechend gilt für die Stichprobenvarianz:**

**Stichprobenvarianz (Originaldaten)**

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

**Stichprobenvarianz (gruppierte Daten)**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} n_i (a_i - \overline{x})^2$$

**Stichprobenvarianz (klassierte Daten)**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} n_i (m_i - \overline{x})^2$$

# 4.1 Varianz

Für $\sigma^2$ gilt der sogenannte **Verschiebungssatz,** der die **manuelle Berechnung** der Varianz **vereinfacht:**

**Varianz (Originaldaten)**

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

**Varianz (gruppierte Daten)**

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i a_i^2 - \overline{x}^2$$

**Varianz (klassierte Daten)**

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i m_i^2 - \overline{x}^2$$

# 4.1 Varianz

**Will man den Verschiebungssatz für $s^2$ nutzen, so kann man zunächst $\sigma^2$ berechnen und dann in $s^2$ umrechnnen:**

**Umrechnung** $\sigma^2 \leftrightarrow s^2$

$$s^2 = \frac{n}{n-1}\sigma^2 \qquad \text{bzw.} \qquad \sigma^2 = \frac{n-1}{n}s^2$$

**Unterschiede sind nur bei kleinen Stichprobenumfängen gravierend, bei großen Stichprobenumfängen vernachlässigbar.**

# 4.1 Varianz

## Beispiel: (Patientendaten)

| $i$ | $x_i$ | $x_i^2$ |
|---|---|---|
| 1 | 25 | 625 |
| 2 | 21 | 441 |
| 3 | 18 | 324 |
| 4 | 37 | 1.369 |
| 5 | 56 | 3.136 |
| 6 | 89 | 7.921 |
| 7 | 46 | 2.116 |
| 8 | 23 | 529 |
| 9 | 21 | 441 |
| 10 | 34 | 1.156 |
| Summe: | 370 | 18.058 |

$$\sigma^2 = \frac{1}{10} \cdot 18.058 - 37^2 = 436,8.$$

$$s^2 = \frac{10}{9} \cdot 436,8 = 485,\overline{3}.$$

# 4.1 Varianz

## Beispiel: (Oktoberfestdaten)

| $a_i$ | $n_i$ | $n_i a_i^2$ |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 30 | 120 |
| 3 | 37 | 333 |
| 4 | 28 | 448 |
| 5 | 23 | 575 |
| 6 | 8 | 288 |
| Summe: | 128 | 1766 |

$$\sigma^2 = \frac{1766}{128} - 3,5^2 = 1,55.$$

$$s^2 = \frac{128}{127} \cdot 1,55 = 1,56.$$

# 4.1 Varianz

**Beispiel (Trinkgelddaten):**

| Klasse | $m_i$ | $n_i$ | $n_i m_i^2$ |
|---|---|---|---|
| $[0;1)$ | 0,5 | 3 | 0,75 |
| $[1;2)$ | 1,5 | 4 | 9,00 |
| $[2;3)$ | 2,5 | 4 | 25,00 |
| $[3;4)$ | 3,5 | 2 | 24,50 |
| $[4;5)$ | 4,5 | 7 | 141,75 |
| Summe: | x | 20 | 201,00 |

$$\sigma^2 = \frac{1}{20} \cdot 201 - 2,8^2 = 2,21.$$

$$s^2 = \frac{20}{19} \cdot 2,21 = 2,33.$$

# 4.1 Varianz

**Bemerkungen:**

- Starke Bedeutung ausgehend von der induktiven Statistik

**Aber:**

- **Ausreißerempfindlich**
- Verwendet die **quadrierte Maßeinheit** des Ausgangsdatensatzes und daher **schwierig zu interpretieren.**

**Ausweg: Wurzelziehen → Standardabweichung**

# 4.1 Standardabweichung

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

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

**Bemerkungen:**

- Vorteil gegenüber Varianz: **gleiche Einheit wie Ausgangsdatensatz**
- **Populärstes** Streuungsmaß

**Aber:**

- Nach wie vor **ausreißerempfindlich** (Eigenschaft überträgt sich von Varianz)

# 4.1 Standardabweichung

**Beispiele: Berechnen Sie die Standardabweichung zu den zuvor betrachteten Datensätzen.**

**Lösungen:**

- **Patientendaten:** $\sigma = \sqrt{436{,}8} = 20{,}9 \quad s = \sqrt{485} = 22{,}03$
- **Oktoberfestdaten:** $\sigma = \sqrt{1{,}55} = 1{,}24 \quad s = \sqrt{1{,}56} = 1{,}25$
- **Klassierte Trinkgelddaten:** $\sigma = \sqrt{2{,}21} = 1{,}49, \quad s = \sqrt{2{,}33} = 1{,}53$

# 4.2 Spannweite

**Spannweite (Originaldaten):** $R = x_{(n)} - x_{(1)}$

**Spannweite (gruppierte Daten):** $R = a_k - a_1$

**Spannweite (klassierte Daten): unüblich**

**Die Spannweite ist rein axiomatisch ein Streuungsmaß, aber als solches eher ungeeignet, da extrem ausreißerempfindlich. Sie dient eher als Orientierungshilfe, etwa bei der Klassenbildung.**

**Beispiel:**

**Patientendaten:** $R = 89 - 18 = 71$

**Oktoberfestdaten:** $R = 6 - 1 = 5$

# 4.3 Quartilsabstand

**(auch: Interquartilsabstand)**

$$IQR = QD_{0,25} = x_{0,75} - x_{0,25}$$

**Bemerkungen:**

- **Ausreißerunempfindlich**

**Aber:**

- **nicht unmittelbar vergleichbar mit der Standardabweichung!**

**Beispiele:**

**Patientendaten:** $IQR = 46 - 21 = 25$

**Oktoberfestdaten:** $IQR = 4 - 2{,}5 = 1{,}5$

**Apothekendaten:** $IQR = 2{,}19 - 1{,}15 = 1{,}04$

# 4.4 Lineare Transformationen

**Was passiert mit Streuungsmaßen, wenn eine lineare Transformation der Art**

$$y_i = a + bx_i$$

**durchgeführt wird?**

**Varianz:** $\sigma_y^2 = b^2\sigma_x^2 \qquad s_y^2 = b^2 s_x^2$

**Standardabweichung:** $\sigma_y = |b|\sigma_x \qquad s_y = |b|s_x$

**Spannweite:** $R_y = |b|R_x$

**(Inter-)quartilsabstand:** $IQR_y = |b|IQR_x$

# 4.4 Lineare Transformationen

**Beispiel: Oktoberfestdaten**

**Die Standardabweichung für die Anzahl der Maßkrüge lautete**

$$s_x = 1{,}25.$$

**Betrachtet man erneut den Literkonsum und unterstellt jedem Maßkrug eine Füllung von 0,9 l, so erhält man die Standardabweichung für den transformierten Datensatz als**

$$s_y = 0{,}9 s_x = 0{,}9 \cdot 1{,}25 = 1{,}125.$$

# 4.4 Variationskoeffizient

**Bislang:** Maße der **absoluten** Streuung

Betrachte erneut Oktoberfestdaten:

Ist die Streuung des Konsumverhaltens eine andere, wenn man Maßkrüge oder Liter misst?

**Ja und Nein…**

- Natürlich ist die **absolute** Streuung **skalenabhängig** und ändert sich damit, je nachdem, ob man die Einheit in Litern, Millilitern, etc. misst.
- Die den Daten eigene Streuung sollte aber nicht skalenabhängig sein. Deshalb sucht man zuweilen ein **skalenunabhängiges Maß**, ein sog. Maß der **relativen** Streuung.

# 4.4 Variationskoeffizient

**Der Variationskoeffizient setzt Standardabweichung und arithmetisches Mittel ins Verhältnis:**

$$V = \frac{\sigma}{\overline{x}} \qquad \text{bzw.} \qquad V = \frac{s}{\overline{x}}$$

**Für lineare Transformationen der Art** $y_i = bx_i$

**mit** $b > 0$ **gilt daher:**

$$V_y = V_x.$$

**Fortsetzung Oktoberfestdaten: siehe Vorlesung.**

# 5 Schiefe und Wölbung

**Welche weiteren Maßzahlen sind informativ zur Beschreibung des Datensatzes?**

- **Schiefe** versus **Symmetrie** eines Datensatzes
- Verhalten in den **Außenbereichen** einer Verteilung **(Wölbung)**

→ Beurteilung über sogenannte **Momente.**

# 5 Schiefe und Wölbung

Bislang:

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

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

→ Einheitliches Muster!

# 5 Schiefe und Wölbung

## Momente:

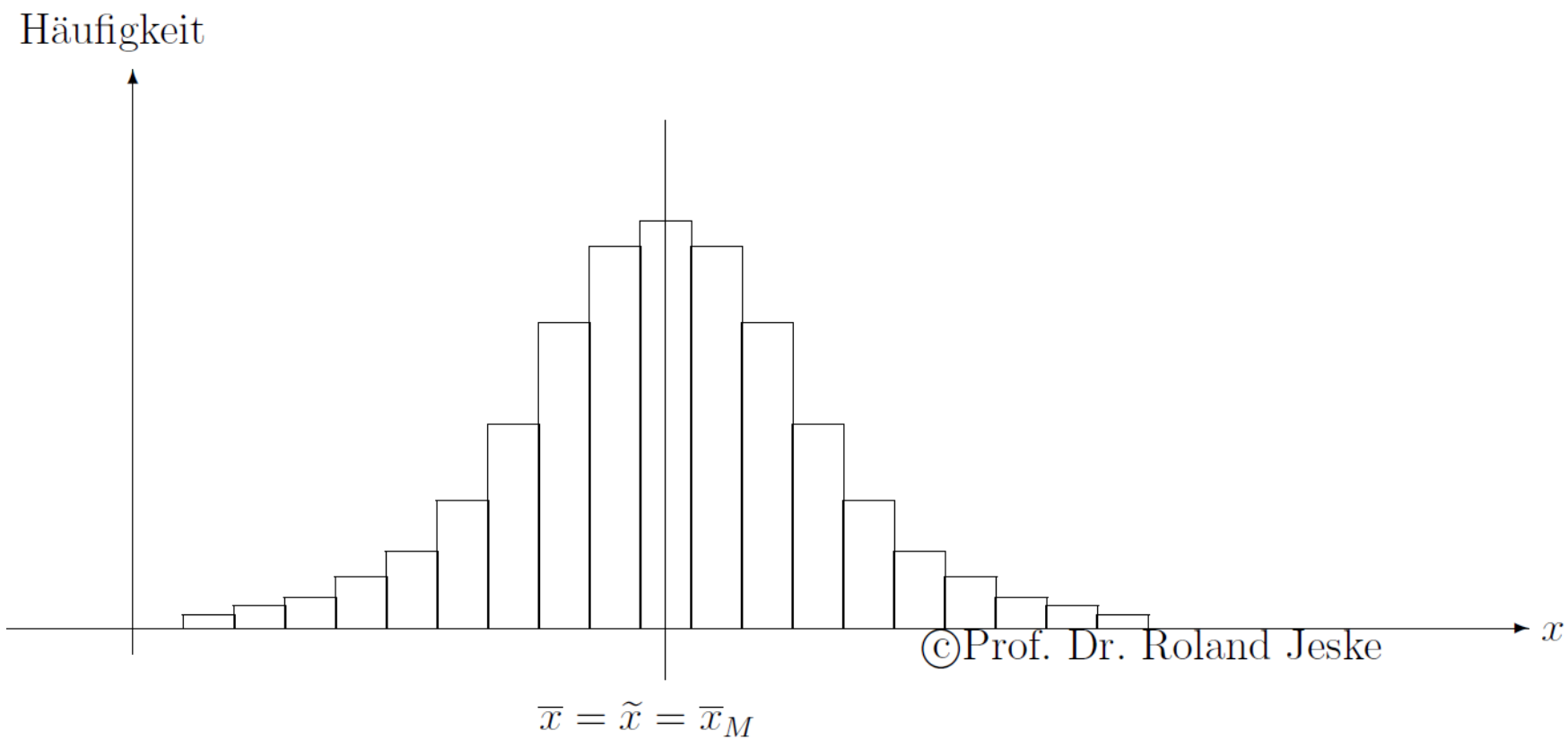
Allgemein:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k$$ k-tes (nichtzentriertes) Moment

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k$$ k-tes zentriertes Moment

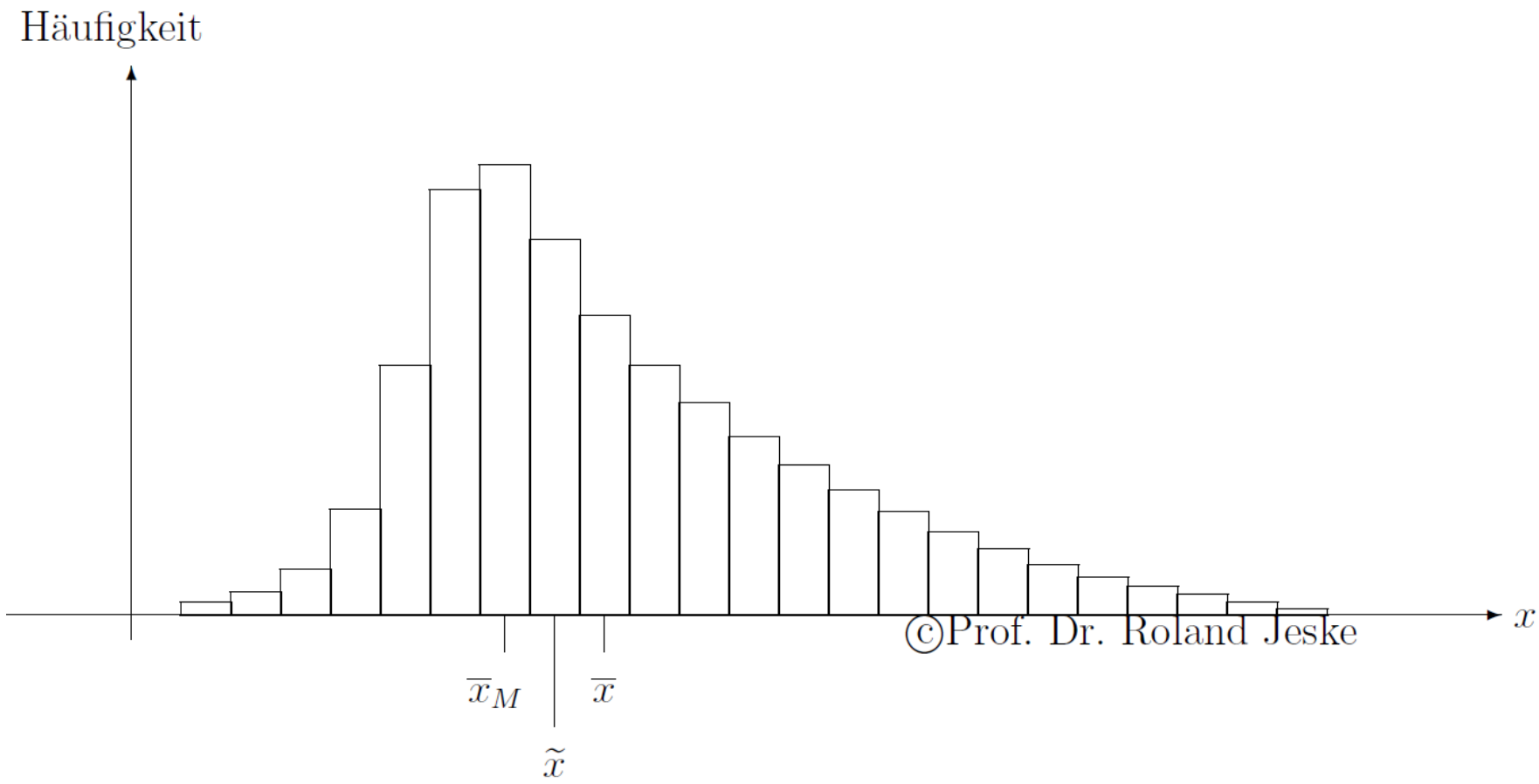
# 5.1 Schiefe

## Histogramm einer symmetrischen Verteilung:



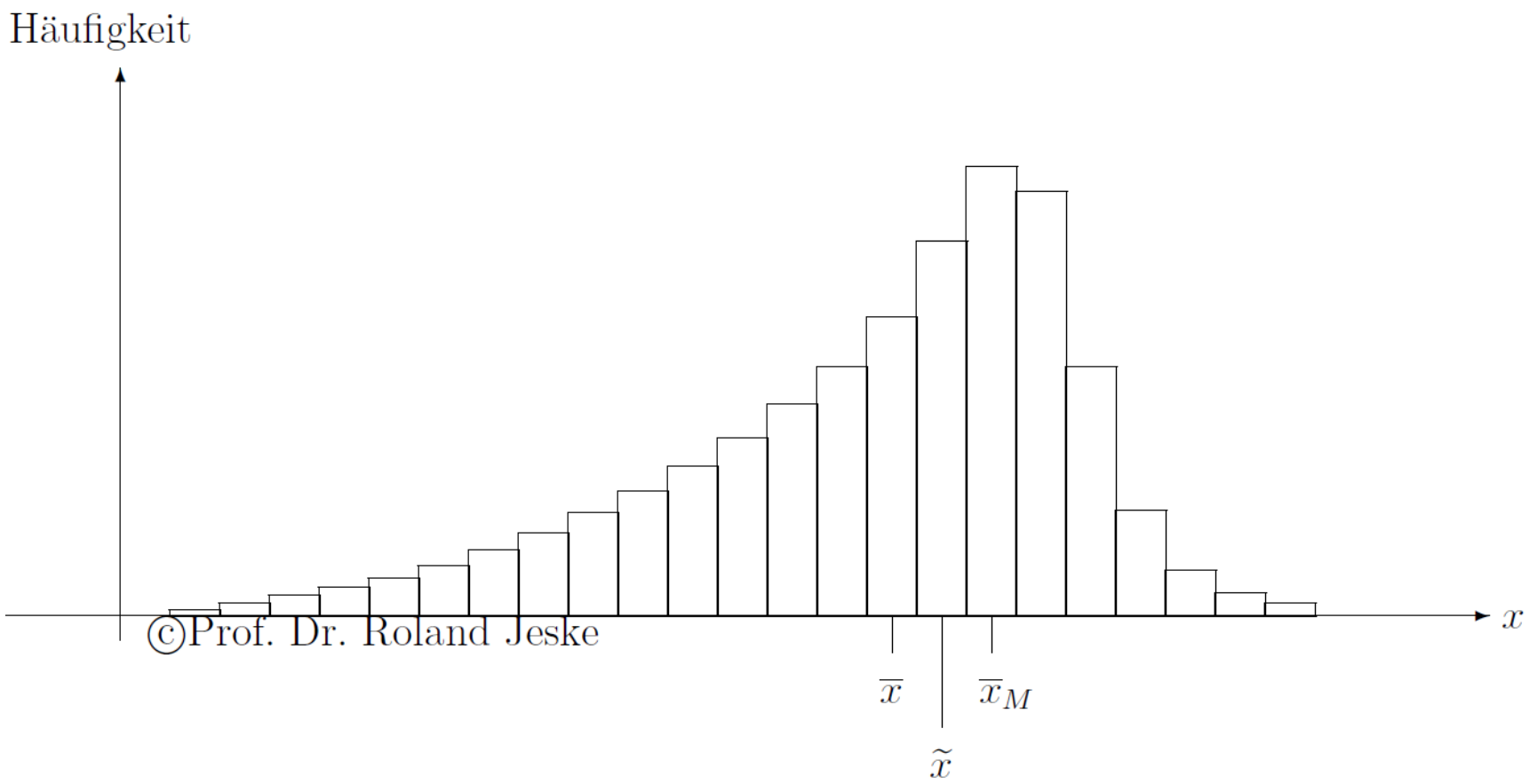
# 5.1 Schiefe

## Histogramm einer rechtsschiefen Verteilung:



# 5.1 Schiefe

## Histogramm einer linksschiefen Verteilung:



# 5.1 Schiefe

**Fisher'scher Schiefekoeffizient für Originaldaten**

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^3} \in \mathbb{R}$$

**Schieferegel zum Schiefekoeffizienten nach Fisher**

$$\begin{aligned} \text{Symmetrische Verteilung} &\Rightarrow \gamma_1 = 0 \\ \text{Rechtsschiefe Verteilung} &\Rightarrow \gamma_1 > 0 \\ \text{Linksschiefe Verteilung} &\Rightarrow \gamma_1 < 0 \end{aligned}$$

- **Problem: aufwändig in der manuellen Berechnung (kein Analogon zum Verschiebungssatz der Varianz)**
- **extrem ausreißerempfindlich, aber populärstes Maß für Schiefe!**

# 5.1 Schiefe

**Beispiel:**
Liegedauern (in Tagen) von Patient/innen nach einer bestimmten OP:
7, 4, 10, 5, 6, 8, 8, 9, 7, 8, 18, 8, 7, 9, 6

Berechnen Sie die Schiefe (nach Fisher) der Daten.

**Lösung:** siehe Vorlesung

# 5.1 Schiefe

**Einfache** Beurteilung der **Schiefe** anhand der **Lageregel** nach **Fechner:**

- Verteilung ist symmetrisch $\Rightarrow \bar{x} = \tilde{x} = x_M$
- Verteilung ist rechtsschief $\Rightarrow \bar{x} > \tilde{x} > x_M$
- Verteilung ist linksschief $\Rightarrow \bar{x} < \tilde{x} < x_M$

**Beispiel:**
Wenden Sie die Fechnerscher Regel für die Liegedauern an.

**Lösung:** siehe Vorlesung

# 5.1 Schiefe

## Quartilskoeffizient zur Schiefe

$$QS_{0,25} = \frac{(x_{0,75} - \tilde{x}) - (\tilde{x} - x_{0,25})}{x_{0,75} - x_{0,25}} \in [-1; 1]$$

## Schieferegel für den Quartilskoeffizienten zur Schiefe

$$\begin{array}{lcl} \text{Symmetrische Verteilung} & \Rightarrow & QS_{0,25} = 0 \\ \text{Rechtsschiefe Verteilung} & \Rightarrow & QS_{0,25} > 0 \\ \text{Linksschiefe Verteilung} & \Rightarrow & QS_{0,25} < 0 \end{array}$$

- **Einfach zu berechnen**
- **ausreißerunempfindlich**

# 5.1 Schiefe

**Beispiel:**

Berechnen Sie den Quartilskoeffizienten der Schiefe für die Liegedauern.

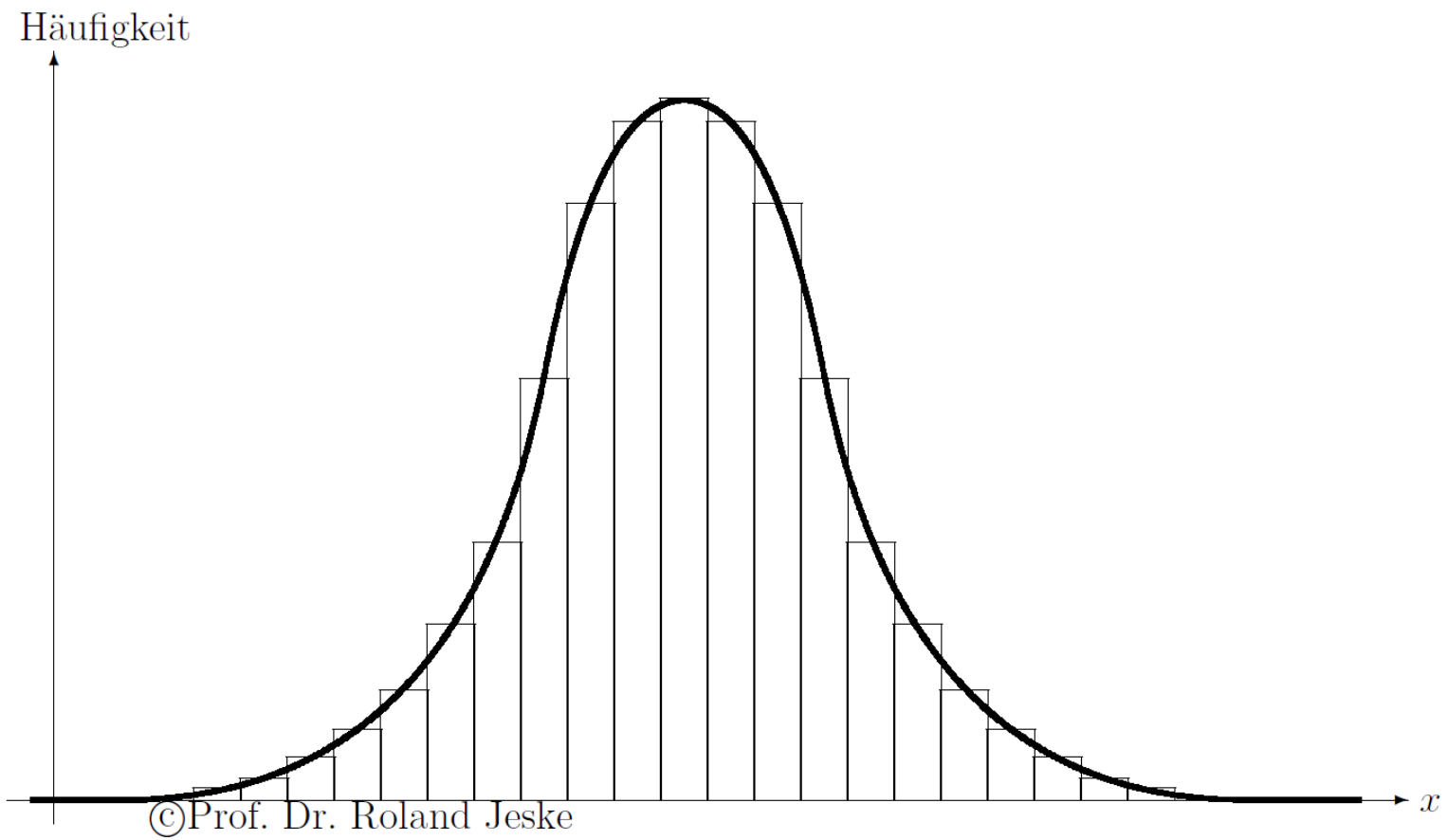
**Lösung:** siehe Vorlesung

**Fazit?**

- **Implikationen für die Praxis in die falsche Richtung (es gibt mehr als Symmetrie und Schiefe!)**
- **Schiefe macht keinen Sinn für kleine Datensätze, sondern ist eine Eigenschaft für große Datensätze (ab n=100)**

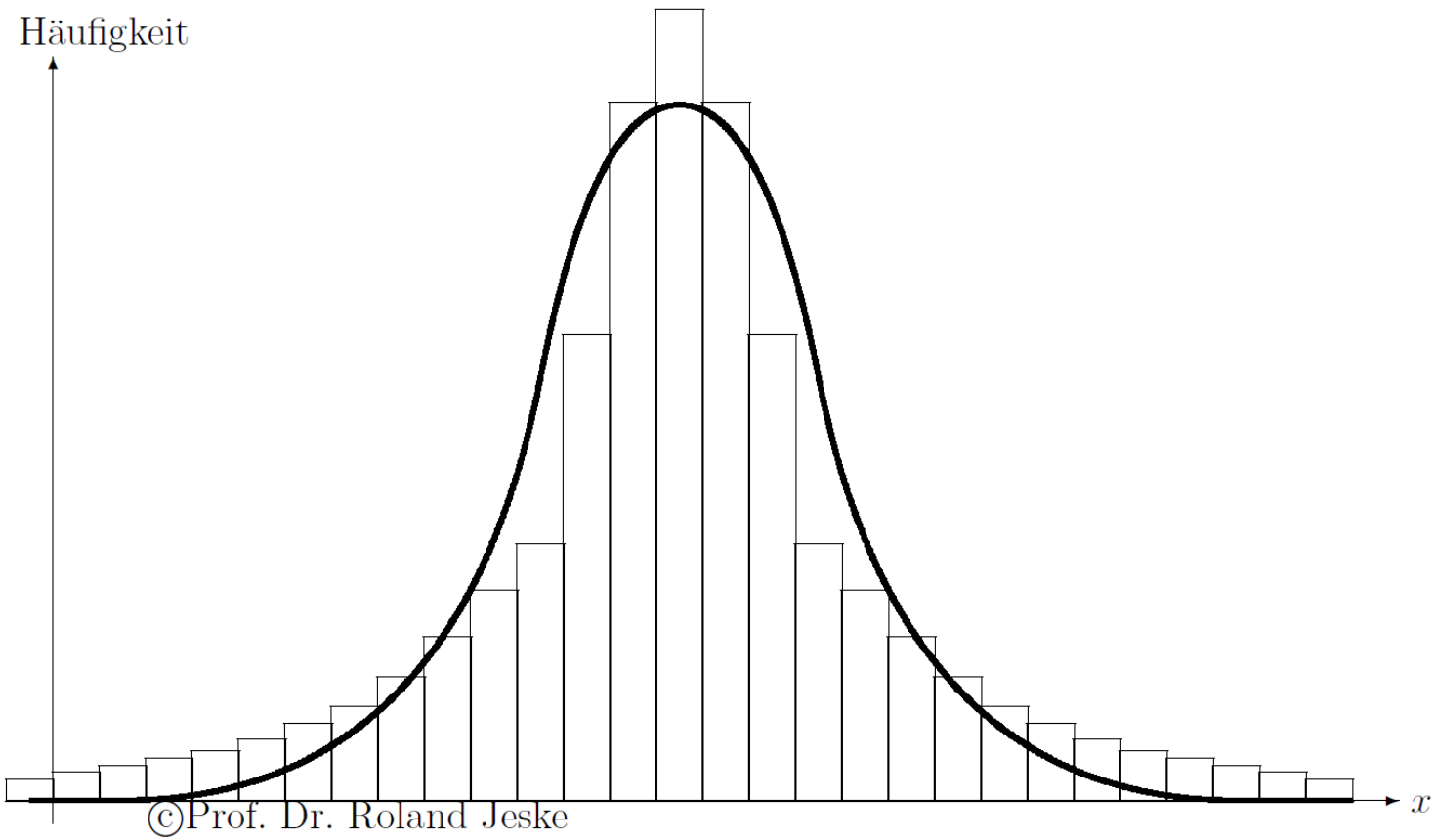
# 5.2 Wölbung

**Histogramm einer normalgewölbten (mesokurtischen) Verteilung:**



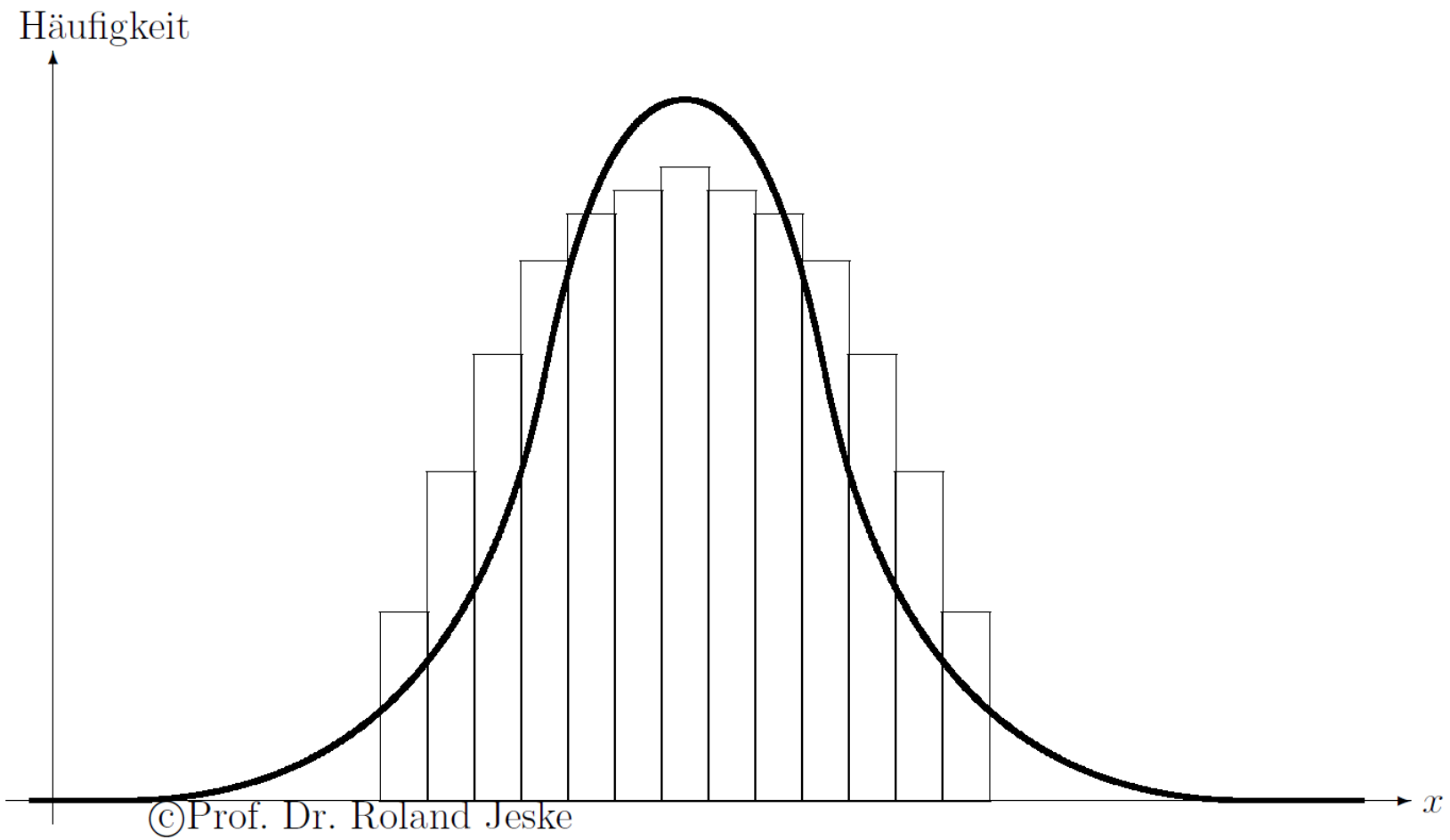
# 5.2 Wölbung

## Histogramm einer spitzen (leptokurtischen) Verteilung:



# 5.2 Wölbung

**Histogramm einer flachen (platykurtischen) Verteilung:**



# 5.2 Wölbung

## Fisher'scher Wölbungskoeffizient für Originaldaten

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} \in \mathbb{R}$$

### Wölbungsregel zur Fisher'schen Wölbung

| | | |
|---|---|---|
| Mesokurtische (normalgewölbte) Verteilung | $\Rightarrow$ | $\gamma_2 = 3$ |
| Leptokurtische (spitze) Verteilung | $\Rightarrow$ | $\gamma_2 > 3$ |
| Platykurtische (abgeflachte) Verteilung | $\Rightarrow$ | $\gamma_2 < 3$ |

# 5.2 Wölbung

**Fisher'scher Exzess für Originaldaten**

$$\gamma_2^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3 \in \mathbb{R}$$

$$\begin{array}{lcl} \text{Mesokurtische (normalgewölbte) Verteilung} & \Rightarrow & \gamma_2^* = 0 \\ \text{Leptokurtische (spitze) Verteilung} & \Rightarrow & \gamma_2^* > 0 \\ \text{Platykurtische (abgeflachte) Verteilung} & \Rightarrow & \gamma_2^* < 0 \end{array}$$

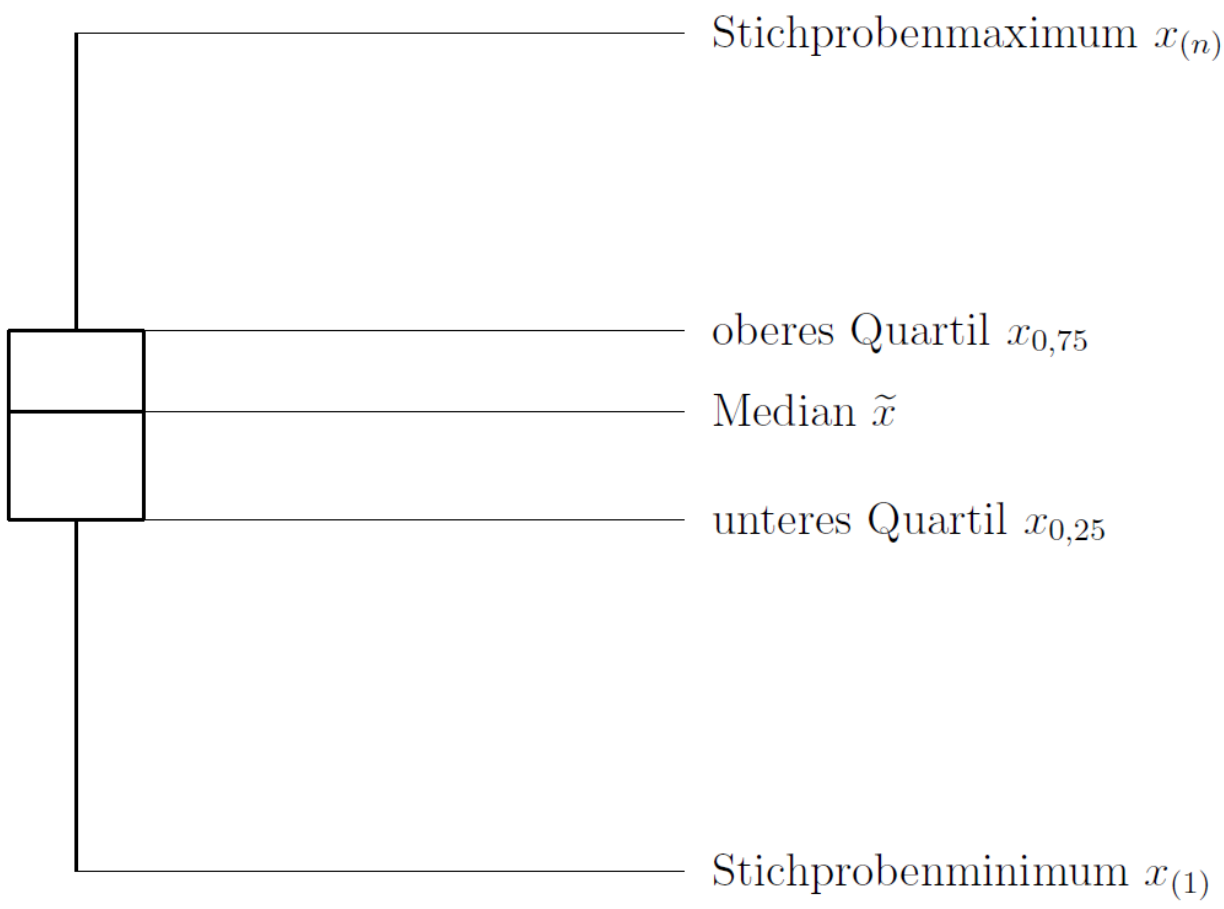
In der deutschen Literatur wird der Exzess auch häufig als Wölbung bezeichnet (**aufpassen, ob 3 bereits abgezogen wurde!**), die englische Literatur unterscheidet eindeutig zwischen Wölbung und Exzess.

**Welche Informationen sind für die Beschreibung eines Datensatzes wichtig?**

**5-Number-Summary**

| | |
|---|---|
| $x_{(1)}$ | Stichprobenminimum |
| $x_{0,25}$ | unteres Quartil |
| $\tilde{x}$ | Median |
| $x_{0,75}$ | oberes Quartil |
| $x_{(n)}$ | Stichprobenmaximum |

## Grafische Darstellung der 5 Number Summary: Der Box-Plot



**Interpretationen des Boxplots:**

- **Am Boxplot kann mit dem Median (Strich in der Boxmitte) ein Lagemaß abgelesen werden. Anhänger des Boxplots legen Wert darauf, dass dieses Lagemaß ein ausreißerunempfindliches Maß ist.**
- **Mögliche Ausreißer hingegen kann man an den Strichen zum Stichprobenmaximum oder Minimum erkennen. Vielfach werden diese auch in einem verfeinerten Boxplot, dem sog. punktierten Boxplot individuell dargestellt (siehe etwa Buch, Abschnitt 6.1).**
- **Weiterhin kann mit der Breite der Box ein Streuungsmaß abgelesen werden, nämlich der Quartilsabstand. Auch dieses Maß wurde bewusst ausreißerunemfindlich gewählt.**

**Interpretationen des Boxplots (Fortsetzung):**

- **Auch zur Schiefe bzw. Symmetrie eines Datensatzes kann mittels eines Boxplots eine Aussage getroffen werden (Viasualisierung des Quartilskoeffizienten der Schiefe):**
  **Liegt der Median *(etwa)* mittig in der Box, so deutet dies auf eine *(annähernd)* symmetrische Verteilung hin.**
  **Liegt der Median mehr zum unteren Quartil hin, so deutet dies auf eine rechtsschiefe Verteilung hin.**
  **Liegt der Median hingegen mehr zum oberen Quartil hin, so spricht dies für eine linksschiefe Verteilung.**
  **Es handelt sich bei dieser Schiefebetrachtung wiederum um ein robustes, d.h. ausreißerunempfindliches Maß.**

# 6 Hochwertige Grafiken

**Beispiel:**

Zeichnen Sie den Boxplot für

- die Patientendaten
- die Oktoberfestdaten

**Lösung: siehe Vorlesung**

# 7 Bivariate Daten

**Bislang: univariate Daten,** d.h., es wurde jeweils nur ein **einzelnes Merkmal** betrachtet. Zukünftig wird das **Zusammenwirken zweier Merkmale** (bivariate Untersuchung) betrachtet.

**Mögliche Datenformate:**

- **Originaldaten:**

Bivariate Originaldaten liegen als Punkte vor:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

# 7 Bivariate Daten

## Gruppierte Daten:

Bivariat gruppierte Daten liegen in Form einer zweidimensionalen Häufigkeitstabelle, einer sogenannten Kontingenztabelle oder Kontingenztafel vor:

| | $b_1$ | $b_2$ | $\dots$ | $b_\ell$ | |
|---|---|---|---|---|---|
| $a_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1\ell}$ | $n_{1\bullet}$ |
| $a_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2\ell}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $a_k$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{k\ell}$ | $n_{k\bullet}$ |
| | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet\ell}$ | $n$ |

# 7 Bivariate Daten

## Klassierte Daten:

Bivariat klassierte Daten liegen in Form einer zweidimensionalen Kontingenztabelle vor:

| | $[b_0^\star; b_1^\star)$ | $[b_1^\star; b_2^\star)$ | $\dots$ | $[b_{\ell-1}^\star; b_\ell^\star)$ | |
|---|---|---|---|---|---|
| $[a_0^\star; a_1^\star)$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1\ell}$ | $n_{1\bullet}$ |
| $[a_1^\star; a_2^\star)$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2\ell}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $[a_{k-1}^\star; a_k^\star)$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{k\ell}$ | $n_{k\bullet}$ |
| | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet\ell}$ | $n$ |

# 7 Bivariate Daten

**Auch gemischte Datenformen (z. B. gruppiert/klassiert) sind möglich:**

| | $b_1$ | $b_2$ | $\dots$ | $b_\ell$ | |
|---|---|---|---|---|---|
| $[a_0^\star; a_1^\star)$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1\ell}$ | $n_{1\bullet}$ |
| $[a_1^\star; a_2^\star)$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2\ell}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $[a_{k-1}^\star; a_k^\star)$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{k\ell}$ | $n_{k\bullet}$ |
| | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet\ell}$ | $n$ |

# 7 Bivariate Daten

## Beispiel:

| Altersklassen | Geschlecht | | |
|---|---|---|---|
| | männlich | weiblich | Insgesamt |
| unter 3 Jahre | 1.018.505 | 966.018 | 1.984.523 |
| 3 bis unter 6 Jahre | 1.041.011 | 984.172 | 2.025.183 |
| 6 bis unter 15 Jahre | 3.485.685 | 3.309.900 | 6.795.585 |
| 15 bis unter 18 Jahre | 1.195.380 | 1.133.681 | 2.329.061 |
| 18 bis unter 25 Jahre | 3.325.707 | 3.194.751 | 6.520.458 |
| 25 bis unter 30 Jahre | 2.455.885 | 2.416.648 | 4.872.533 |
| 30 bis unter 40 Jahre | 4.763.360 | 4.731.444 | 9.494.804 |
| 40 bis unter 50 Jahre | 6.756.735 | 6.594.133 | 13.350.868 |
| 50 bis unter 65 Jahre | 8.081.342 | 8.247.217 | 16.328.559 |
| 65 bis unter 75 Jahre | 4.246.483 | 4.788.107 | 9.034.590 |
| 75 Jahre und mehr | 2.775.848 | 4.707.683 | 7.483.531 |
| Insgesamt | 39.145.941 | 41.073.754 | 80.219.695 |

Tabelle 8.1: Bevölkerung Deutschlands am 09.05.2011 (Zensusstichtag) nach Geschlecht und Altersgruppen (Quelle: Destatis)

## Randverteilung bivariater Daten

Absolute Randhäufigkeiten

$$n_{i\bullet} = \sum_{j=1}^{\ell} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$$

Relative Randhäufigkeiten

$$r_{i\bullet} = \sum_{j=1}^{\ell} r_{ij}$$

$$r_{\bullet j} = \sum_{i=1}^{k} r_{ij}$$

## Bedingte Verteilung bivariater Daten

Bedingte absolute Häufigkeiten

$$n_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$$

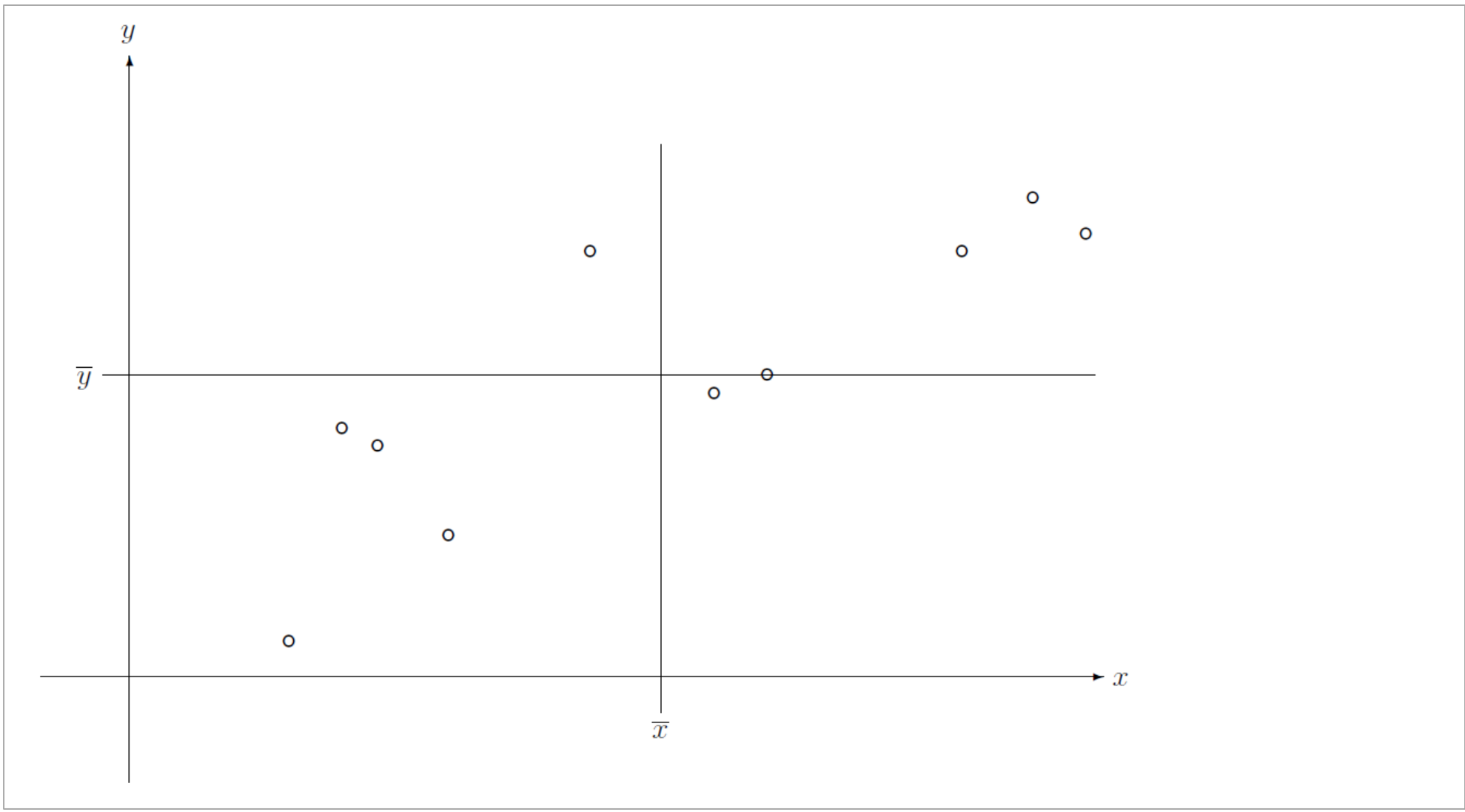
$$n_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$$

Bedingte relative Häufigkeiten

$$r_{i|j} = \frac{r_{ij}}{r_{\bullet j}}$$

$$r_{j|i} = \frac{r_{ij}}{r_{i\bullet}}$$

# 7 Grafik bivariater Daten: Scatterplot



# 8 Zusammenhangsmaße

**Allgemeine Unterscheidung:**

- **Korrelation:** Zusammenhangsmaß zwischen **kardinal** oder **ordinal** skalierten Merkmalen
- **Kontingenz:** Zusammenhangsmaß bei **nominal** skalierten Merkmalen

# 8.1 Kovarianz

**Kovarianz**

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

**Trivialerweise:**

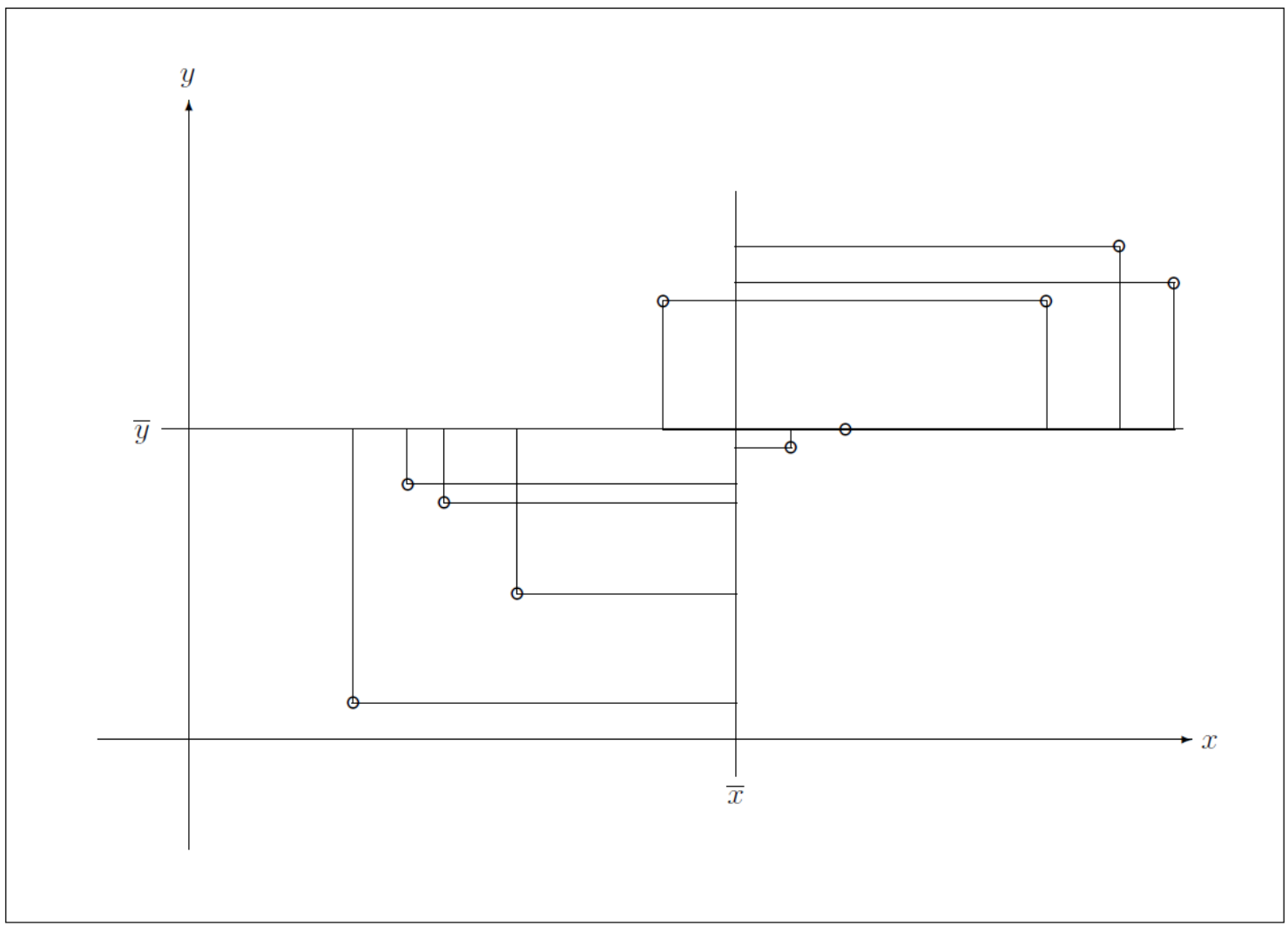
$$\sigma_{xx} = \sigma_x^2$$

**Speziell:**

**Kovarianz (Verschiebungsformel)**

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} \qquad (9.1)$$

# 8.1 Kovarianz



# 8.1 Kovarianz

- Quadrant I:

$$x_i > \overline{x} \text{ und } y_i > \overline{y} \quad \Leftrightarrow \quad x_i - \overline{x} > 0, y_i - \overline{y} > 0 \quad \Rightarrow \quad (x_i - \overline{x})(y_i - \overline{y}) > 0$$

- Quadrant II:

$$x_i < \overline{x} \text{ und } y_i > \overline{y} \quad \Leftrightarrow \quad x_i - \overline{x} < 0, y_i - \overline{y} > 0 \quad \Rightarrow \quad (x_i - \overline{x})(y_i - \overline{y}) < 0$$

- Quadrant III:

$$x_i < \overline{x} \text{ und } y_i < \overline{y} \quad \Leftrightarrow \quad x_i - \overline{x} < 0, y_i - \overline{y} < 0 \quad \Rightarrow \quad (x_i - \overline{x})(y_i - \overline{y}) > 0$$

- Quadrant IV:

$$x_i > \overline{x} \text{ und } y_i < \overline{y} \quad \Leftrightarrow \quad x_i - \overline{x} > 0, y_i - \overline{y} < 0 \quad \Rightarrow \quad (x_i - \overline{x})(y_i - \overline{y}) < 0$$

# 8.1 Kovarianz

**Folglich:**

- **Viele** und **große Flächen** im **ersten** und **dritten Quadranten** führen zu positiven Werten **(“positiver Zusammenhang”)**
- **Viele** und **große Flächen** im **zweiten** und **vierten Quadranten** führen zu negativen Werten **(“negativer Zusammenhang”)**

**Problem:**

**Kovarianz** ist **unbeschränkt** und kann **jeden reellen Wert** annehmen!

# 8.1 Korrelation nach Bravais-Pearson

**Ausweg: Standardisierung:**

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1; 1]$$

**Korrelationskoeffizient nach Bravais-Pearson für Originaldaten**

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

**Korrelationskoeffizient nach Bravais-Pearson für Originaldaten (Verschiebungsformel)**

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

# 8.1 Korrelation nach Bravais-Pearson

**Was misst der Korrelationskoeffizient nach Bravais-Pearson genau?**

**Interpretation des Korrelationskoeffizienten nach Bravais-Pearson**

Der Korrelationskoeffizient nach Bravais-Pearson misst den **linearen** Zusammenhang zweier Merkmale:

$r_{xy} = 1 \quad \Leftrightarrow$ Alle Beobachtungen liegen auf einer Geraden mit positiver Steigung.

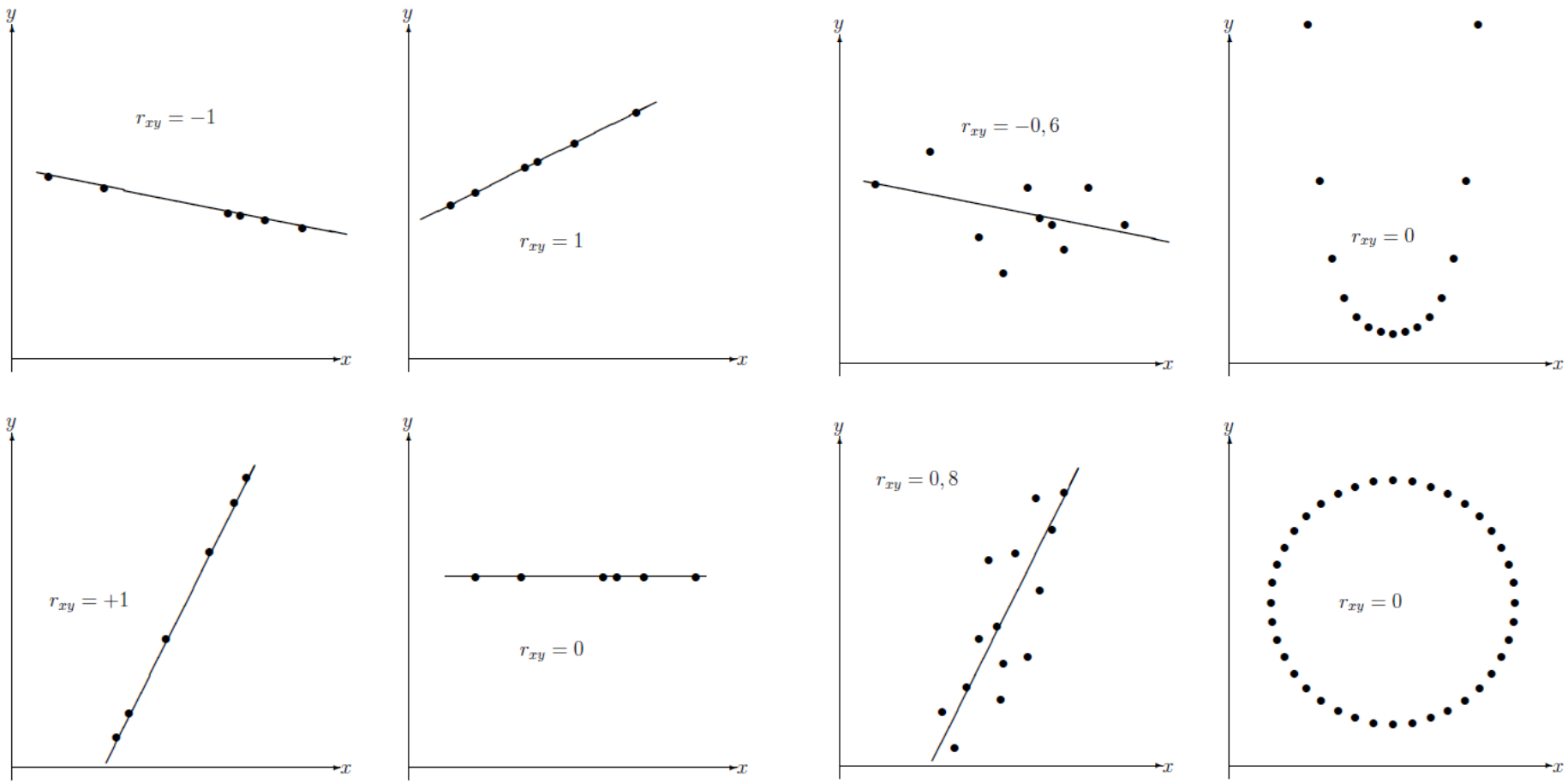
$r_{xy} = -1 \quad \Leftrightarrow$ Alle Beobachtungen liegen auf einer Geraden mit negativer Steigung.

$r_{xy} = 0 \quad \Leftrightarrow$ Es liegt kein **linearer** Zusammenhang vor.

**Es kann aber sehr wohl ein nichtlinearer Zusammenhang vorliegen, wenn eine Korrelation von Null ausgewiesen wird!**

# 8.1 Korrelation nach Bravais-Pearson

## Was misst der Korrelationskoeffizient nach Bravais-Pearson genau?



# 8.1 Korrelation nach Bravais-Pearson

**Beispiel:**

*Die Alpenklinik hat an ausgewählten Tagen die Anzahl $x$ der verkauften Tageskarten (in Tausend) im benachbarten Skigebiet sowie die Anzahl $y$ der Aufnahmen in der Unfallchirurgie erfasst:*

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 5 | 12 |
| 2 | 6 | 14 |
| 3 | 5,5 | 9 |
| 4 | 2 | 4 |
| 5 | 3,8 | 7 |
| 6 | 4,4 | 10 |
| 7 | 6,2 | 13 |
| 8 | 5,6 | 12 |
| 9 | 4,2 | 7 |
| 10 | 5,9 | 15 |

*Berechnen Sie die Korrelation.*

# 8.1 Korrelation nach Bravais-Pearson

## Lösung 1:

| $i$ | $x_i$ | $y_i$ | $(x_i - \overline{x})$ | $(y_i - \overline{y})$ | $(x_i - \overline{x})^2$ | $(y_i - \overline{y})^2$ | $(x_i - \overline{x})(y_i - \overline{y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 5 | 12 | 0,14 | 1,7 | 0,0196 | 2,89 | 0,238 |
| 2 | 6 | 14 | 1,14 | 3,7 | 1,2996 | 13,69 | 4,218 |
| 3 | 5,5 | 9 | 0,64 | -1,3 | 0,4096 | 1,69 | -0,832 |
| 4 | 2 | 4 | -2,86 | -6,3 | 8,1796 | 39,69 | 18,018 |
| 5 | 3,8 | 7 | -1,06 | -3,3 | 1,1236 | 10,89 | 3,498 |
| 6 | 4,4 | 10 | -0,46 | -0,3 | 0,2116 | 0,09 | 0,138 |
| 7 | 6,2 | 13 | 1,34 | 2,7 | 1,7956 | 7,29 | 3,618 |
| 8 | 5,6 | 12 | 0,74 | 1,7 | 0,5476 | 2,89 | 1,258 |
| 9 | 4,2 | 7 | -0,66 | -3,3 | 0,4356 | 10,89 | 2,178 |
| 10 | 5,9 | 15 | 1,04 | 4,7 | 1,0816 | 22,09 | 4,888 |
| Summe: | 48,6 | 103 | x | x | 15,104 | 112,1 | 37,22 |

$$r_{xy} = \frac{37,22}{\sqrt{15,104}\sqrt{112,1}} = 0,905$$

# 8.1 Korrelation nach Bravais-Pearson

**Lösung 2:**

| $i$ | $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i \cdot y_i$ |
|---|---|---|---|---|---|
| 1 | 5 | 12 | 25 | 144 | 60 |
| 2 | 6 | 14 | 36 | 196 | 84 |
| 3 | 5,5 | 9 | 30,25 | 81 | 49,5 |
| 4 | 2 | 4 | 4 | 16 | 8 |
| 5 | 3,8 | 7 | 14,44 | 49 | 26,6 |
| 6 | 4,4 | 10 | 19,36 | 100 | 44 |
| 7 | 6,2 | 13 | 38,44 | 169 | 80,6 |
| 8 | 5,6 | 12 | 31,36 | 144 | 67,2 |
| 9 | 4,2 | 7 | 17,64 | 49 | 29,4 |
| 10 | 5,9 | 15 | 34,81 | 225 | 88,5 |
| Summe: | 48,6 | 103 | 251,3 | 1173 | 537,8 |

$$r_{xy} = \frac{537,8 - 10 \cdot 4,86 \cdot 10,3}{\sqrt{251,3 - 10 \cdot 4,86^2}\sqrt{1173 - 10 \cdot 10,3^2}} = 0,905$$

# 8.1 Korrelation nach Bravais-Pearson

**Abwägung zur Berechnung des Korrelationskoeffizienten**

Die (manuelle) Berechnung über die Verschiebungsformel benötigt lediglich drei Hilfsspalten.

Zudem sind die Werte einfacher zu berechnen, insbesondere dann, wenn bei der Mittelwertbildung von $\overline{x}$ und/oder $\overline{y}$ Nachkommastellen auftreten.

**Fazit:** Mit der Verschiebungsformel lässt sich die Korrelation manuell einfacher und schneller berechnen!

**Zeitgewinn bedeutet in der Klausur häufig Punktezuwachs…**

# 8.1 Korrelation nach Bravais-Pearson

**Korrelation nach Bravais-Pearson linear transformierter Daten**

Erfolgen lineare Transformationen der Art

$$\begin{aligned} x_i^\star &= a_x + b_x x_i \\ \text{und} \quad y_i^\star &= a_y + b_y y_i, \end{aligned}$$

so gilt für den Korrelationskoeffizienten nach Bravais-Pearson der transformierten Daten:

$$r_{x^\star y^\star} = \text{sign}(b_x \cdot b_y) r_{xy}.$$

# 8.2 Rangkorrelation nach Spearman

**Wie gelangt man zu einem Zusammenhangsmaß für ordinal skalierte Merkmale?**

**Idee:** Verwende den Korrelationskoeffizienten nach Bravais-Pearson für die **Ränge statt** für die **Beobachtungen:**

$$R_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2}}$$

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

# 8.2 Rangkorrelation nach Spearman

**Korrelationskoeffizient nach Spearman für Originaldaten (mit Bindungen)**

$$R_{xy} = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - \frac{n(n+1)^2}{4}}{\sqrt{\sum_{i=1}^{n} R(x_i)^2 - \frac{n(n+1)^2}{4}}\sqrt{\sum_{i=1}^{n} R(y_i)^2 - \frac{n(n+1)^2}{4}}}$$

**Korrelationskoeffizient nach Spearman für Originaldaten (ohne Bindungen)**

$$R_{xy} = 1 - \frac{6\sum_{i=1}^{n} \left(R(x_i) - R(y_i)\right)^2}{(n-1)n(n+1)}$$

# 8.2 Rangkorrelation nach Spearman

**Was misst der Korrelationskoeffizient nach Spearman genau?**

**Interpretation des Korrelationskoeffizienten nach Spearman**

Der Korrelationskoeffizient nach Spearman misst den **monotonen** Zusammenhang zweier Merkmale:

$R_{xy} = 1 \quad \Leftrightarrow$ Mit steigendem x-Wert steigt auch der y-Wert.

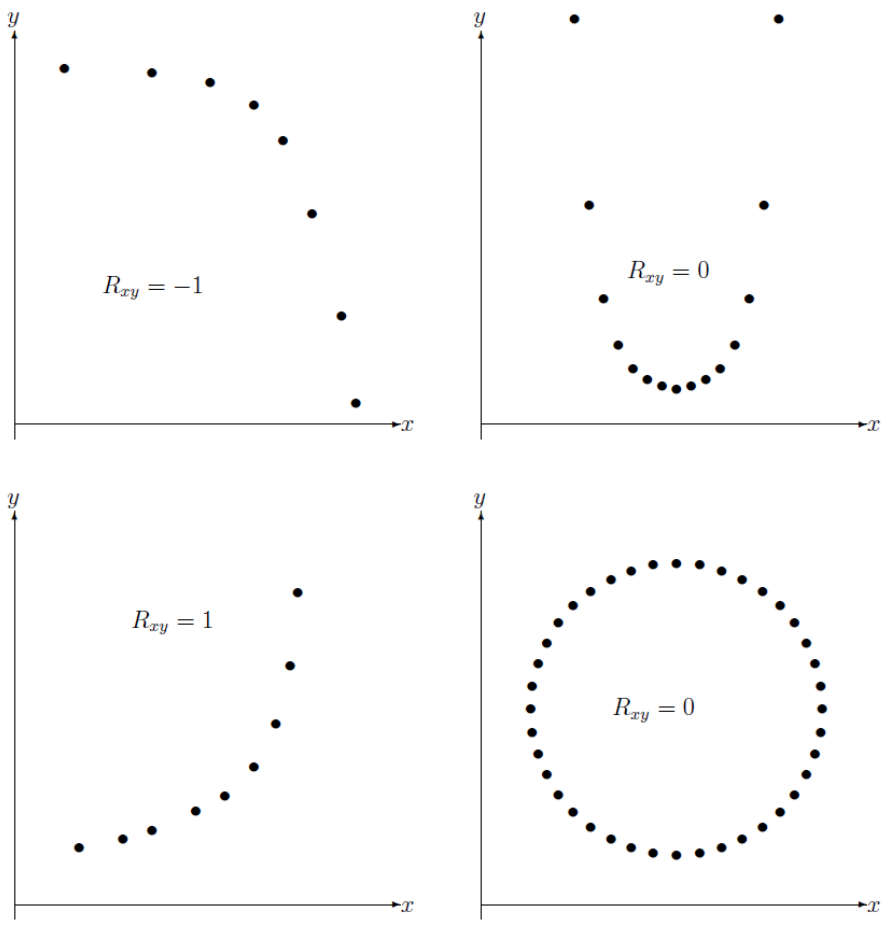
$R_{xy} = -1 \quad \Leftrightarrow$ Mit steigendem x-Wert sinkt der y-Wert.

$R_{xy} = 0 \quad \Leftrightarrow$ Es liegt kein **monotoner** Zusammenhang vor.

**Aber es kann sehr wohl ein nicht-monotoner Zusammenhang vorliegen, wenn eine Korrelation von Null ausgewiesen wird!**

# 8.2 Rangkorrelation nach Spearman

## Was misst der Korrelationskoeffizient nach Spearman genau?



# 8.2 Rangkorrelation nach Spearman

**Korrelation nach Spearman linear transformierter Daten**

Erfolgen lineare Transformationen der Art

$$\begin{aligned} x_i^\star &= a_x + b_x x_i \\ \text{und} \quad y_i^\star &= a_y + b_y y_i, \end{aligned}$$

so gilt für den Korrelationskoeffizienten nach Spearman der transformierten Daten:

$$R_{x^\star y^\star} = \text{sign}(b_x \cdot b_y) R_{xy}.$$

# 8.2 Rangkorrelation nach Spearman

## Beispiel 1:

*Die nachfolgende Tabelle gibt die Mathematik-Noten ($x_i$) und die Deutsch-Noten ($y_i$) zehn zufällig ausgewählter Schüler einer 10. Klasse an:*

| $x_i$ | $y_i$ |
|---|---|
| 1 | 2 |
| 5 | 3 |
| 2 | 3 |
| 2 | 4 |
| 4 | 3 |
| 3 | 5 |
| 4 | 4 |
| 4 | 3 |
| 3 | 5 |
| 3 | 2 |

*Berechnen Sie den Rangkorrelationskoeffizienten nach Spearman.*

# 8.2 Rangkorrelation nach Spearman

**Lösung:**

| $x$ | $y$ | $R(x_i)$ | $R(y_i)$ | $R(x_i)^2$ | $R(y_i)^2$ | $R(x_i)R(y_i)$ |
|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1,5 | 1,00 | 2,25 | 1,50 |
| 5 | 3 | 10 | 4,5 | 100,00 | 20,25 | 45,00 |
| 2 | 3 | 2,5 | 4,5 | 6,25 | 20,25 | 11,25 |
| 2 | 4 | 2,5 | 7,5 | 6,25 | 56,25 | 18,75 |
| 4 | 3 | 8 | 4,5 | 64,00 | 20,25 | 36,00 |
| 3 | 5 | 5 | 9,5 | 25,00 | 90,25 | 47,50 |
| 4 | 4 | 8 | 7,5 | 64,00 | 56,25 | 60,00 |
| 4 | 3 | 8 | 4,5 | 64,00 | 20,25 | 36,00 |
| 3 | 5 | 5 | 9,5 | 25,00 | 90,25 | 47,50 |
| 3 | 2 | 5 | 1,5 | 25,00 | 2,25 | 7,50 |
| | | Summe: | | 380,50 | 378,50 | 311,00 |

$$R_{xy} = \frac{311 - 302,5}{\sqrt{380,5 - 302,5}\sqrt{378,5 - 302,5}} = 0,11.$$

# 8.2 Rangkorrelation nach Spearman

**Beispiel 2:**

**Die nachfolgende Tabelle gibt die Noten von sechs zufällig ausgewählten Schülern in Mathematik ($x_i$) und Physik ($y_i$) wieder:**

| $x_i$ | $y_i$ |
|---|---|
| 2+ | 1- |
| 2- | 2 |
| 1 | 2- |
| 3- | 3+ |
| 5 | 4+ |
| 4- | 4 |

**Berechnen Sie den Korrelationskoeffizienten nach Spearman.**

# 8.2 Rangkorrelation nach Spearman

**Lösung:**

**Es liegen keine Bindungen vor, daher kann mit der vereinfachten Formel gerechnet werden:**

| $x_i$ | $y_i$ | $R(x_i)$ | $R(y_i)$ | $[R(x_i) - R(y_i)]^2$ |
|---|---|---|---|---|
| 2+ | 1- | 2 | 1 | 1 |
| 2- | 2 | 3 | 2 | 1 |
| 1 | 2- | 1 | 3 | 4 |
| 3- | 3+ | 4 | 4 | 0 |
| 5 | 4+ | 6 | 5 | 1 |
| 4- | 4 | 5 | 6 | 1 |
| | | Summe: | | **8** |

**Damit gilt:** $R_{xy} = 1 - \frac{6 \cdot 8}{5 \cdot 6 \cdot 7} = 0{,}77.$

# 8.3 Assoziation und Kontingenz

Ist mindestens eines der Merkmale **nominal skaliert,** so spricht man bei der Berechnung eines Zusammenhangsmaßes von **Kontingenz.**

Im **Spezialfall,** dass **beide Merkmale jeweils nur zwei Ausprägungen** besitzen, spricht man von **Assoziation.** Die entsprechende Kontingenztabelle wird dann auch **Vierfeldertafel** genannt:

| | $b_1$ | $b_2$ | |
|---|---|---|---|
| $a_1$ | $n_{11}$ | $n_{12}$ | $n_{1\bullet}$ |
| $a_2$ | $n_{21}$ | $n_{22}$ | $n_{2\bullet}$ |
| | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n$ |

# 8.3 Assoziation

## Populärstes Maß für die Vierfeldertafel:

**Assoziationskoeffizient nach Yule**

$$Y = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \in [-1; 1]$$

*Eine Erhebung zum Rauchverhalten unter 200 Studierenden ergab folgende Verteilung:*

| | *Raucher* | *Nichtraucher* | |
|---|---|---|---|
| *weiblich* | *30* | *70* | *100* |
| *männlich* | *50* | *50* | *100* |
| | *80* | *120* | *200* |

*Berechnen Sie den Yule'schen Assoziationskoeffizienten.*

# 8.3 Kontingenzkoeffizient

**Liegt eine beliebig große Kontingenztabelle vor, so ist der Yule'sche Assoziationskoeffizient nicht mehr anwendbar.**

**Idee:**

- **Betrachte eine Kontingenztabelle mit gleichen Randverteilungen aber Unabhängigkeit:**

$$\Leftrightarrow \quad r_{ij} = r_{i\bullet} r_{\bullet j}$$

$$\Leftrightarrow \quad n_{ij} = \underbrace{\frac{n_{i\bullet} n_{\bullet j}}{n}}_{=\widehat{n}_{ij}}$$

- **Messe die Abweichung zwischen den Zelleinträgen beider Tabellen:**

$$(n_{ij} - \widehat{n}_{ij})^2$$

# 8.3 Kontingenzkoeffizient

**$\chi^2$ (Chiquadrat) in beliebigen Kontingenztabellen**

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

**$\chi^2$ (Chiquadrat) in Vierfeldertafeln**

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} \qquad (9.4)$$

# 8.3 Kontingenzkoeffizient

**Kontingenzkoeffizient nach Pearson**

$$K_P = \sqrt{\frac{\chi^2}{\chi^2 + n}} \in \left[0; \sqrt{\frac{M-1}{M}}\right]$$

wobei $M = \min\{k, \ell\}$.

**Korrigierter Kontingenzkoeffizient nach Pearson**

$$K_P^\star = \sqrt{\frac{M}{M-1}} K_P \in [0; 1]$$

wobei $M = \min\{k, \ell\}$.

# 8.3 Kontingenzkoeffizient

**Beispiel 1: Berechnen Sie den korrigierten Kontingenzkoeffizienten für folgenden Datensatz:**

| | *Raucher* | *Nichtraucher* | |
|---|---|---|---|
| *weiblich* | *30* | *70* | *100* |
| *männlich* | *50* | *50* | *100* |
| | *80* | *120* | *200* |

**Lösung:**

**1.**

$$\chi^2 = \frac{200(30 \cdot 50 - 70 \cdot 50)^2}{100 \cdot 100 \cdot 80 \cdot 120} = \frac{25}{3} = 8,\overline{3}.$$

**2.**

$$K_P = \sqrt{\frac{\frac{25}{3}}{\frac{25}{3} + 200}} = \sqrt{\frac{1}{25}} = 0,2.$$

**3.**

$$M = \min\{2;2\} = 2 \qquad K_P^* = \sqrt{\frac{2}{2-1}} K_P = \sqrt{2} \cdot 0,2 = 0,283.$$

# 8.3 Kontingenzkoeffizient

**Beispiel 2:**

Betrachten Sie folgenden Datensatz:

| Altersgruppen | Geschlecht | | |
|---|---|---|---|
| | männlich | weiblich | Insgesamt |
| unter 18 Jahre | 6.740.581 | 6.393.771 | 13.134.352 |
| 18 bis unter 65 Jahre | 25.383.029 | 25.184.193 | 50.567.222 |
| 65 Jahre und mehr | 7.022.331 | 9.495.790 | 16.518.121 |
| Insgesamt | 39.145.941 | 41.073.754 | 80.219.695 |

Bevölkerung Deutschlands am 09.05.2011 (Zensusstichtag)
nach Geschlecht und Altersgruppen (Quelle: Destatis)

Berechnen Sie den Kontingenzkoeffizienten nach Pearson für die Merkmale Alter und Geschlecht.

**Lösung: siehe Vorlesung**

# 8.3 Kontingenzkoeffizient

**Zur Interpretation der Kontingenz:**

**Steht etwa ein Wert von 0,2 für eine starke oder schwache Abhängigkeit?**

- **Interpretation in der Praxis erfolgt häufig falsch! Anwender neigen mitunter dazu, Größenwerte ähnlich wie den absoluten Wert eines Korrelationskoeffizienten zu interpretieren, das ist falsch!**
- **In der Theorie kann der korrigierte Kontingenzkoeffizient Werte in [0;1] annehmen, in der Praxis liegen Werte deutlich unter 1!**
- **Aussagen über den Zusammenhang sollten daher statistisch in Form eines Kontingenztests (siehe Statistik 2) abgesichert werden.**
- **Zudem kann man die Art der Abhängigkeit nur über die bedingten Verteilungen klären:**

# 8.3 Kontingenzkoeffizient

**Zur Interpretation der Kontingenz: (vgl. Beispiel 2):**

**Geschlechtsverteilung in der Klasse der Kinder und Jugendlichen:**

| männlich | weiblich |
|---|---|
| 0,513 | 0,487 |

**Geschlechtsverteilung in der Klasse der Personen im Erwerbsalter:**

| männlich | weiblich |
|---|---|
| 0,502 | 0,498 |

**Geschlechtsverteilung in der Klasse der Personen im Rentenalter:**

| männlich | weiblich |
|---|---|
| 0,425 | 0,575 |

# 9. Einfache lineare Regression

**Korrelation:** x <-> y d.h. wechselseitige Beziehung

**Regression:** x -> y d.h. einseitige Einflussnahme von x auf y.

Ziel ist es, eine Gerade der Form

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \qquad i = 1, \ldots, n$$

an die Punktewolke

$$(x_1, y_1), \ldots, (x_n, y_n)$$

anzupassen und dabei die Summe der Abweichungsquadrate zu minimieren:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$$

# 9. Einfache lineare Regression

**Kleinstquadratemethode für die Lineare Regression**

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

mit

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

# 9. Einfache lineare Regression

**Beispiel:** *Dina Vier stellt in ihrer Druckerei hochwertige Bildbände her. Je nach Auflage $x_i$ (in 1.000 Exemplaren) konnte Sie die Kosten $y_i$ (in Tausend €) in einer Stichprobe wie folgt erheben:*

| *Kosten (T€)* | *Produktionsmenge [1.000)* |
|---:|---:|
| *22,1* | *3,1* |
| *16,3* | *2,2* |
| *17,8* | *2,1* |
| *25,9* | *2,9* |
| *20,5* | *2,4* |
| *28,4* | *3,3* |
| *12,1* | *1,5* |
| *22,5* | *3,3* |

*a) Ermitteln Sie die lineare Regressionsgerade, die die Kosten durch die Auflagenhöhe erklärt.*

*b) Dina Vier erhält einen neuen Druckauftrag über 2.800 Bildbände. Welche Gesamtkosten kann sie erwarten?*

# 9. Einfache lineare Regression

**Lösung:**

| $i$ | $y_i$ | $x_i$ | $x_i y_i$ | $x_i^2$ |
|---|---|---|---|---|
| 1 | 22,1 | 3,1 | 68,51 | 9,61 |
| 2 | 16,3 | 2,2 | 35,86 | 4,84 |
| 3 | 17,8 | 2,1 | 37,38 | 4,41 |
| 4 | 25,9 | 2,9 | 75,11 | 8,41 |
| 5 | 20,5 | 2,4 | 49,2 | 5,76 |
| 6 | 28,4 | 3,3 | 93,72 | 10,89 |
| 7 | 12,1 | 1,5 | 18,15 | 2,25 |
| 8 | 22,5 | 3,3 | 74,25 | 10,89 |
| Summe | 165,6 | 20,8 | 452,18 | 57,06 |

$$\overline{y} = \frac{1}{8} \cdot 165,6 = 20,7.$$

$$\overline{x} = \frac{1}{8} \cdot 20,8 = 2,6.$$

$$\widehat{b} = \frac{452,18 - 8 \cdot 20,7 \cdot 2,6}{57,06 - 8 \cdot 2,6^2} = 7,255 \qquad \widehat{a} = 20,7 - 7,255 \cdot 2,6 = 1,837$$

$$\widehat{y}(2,8) = 1,837 + 7,255 \cdot 2,8 = 22,151$$

# 9. Einfache lineare Regression

**Streuungszerlegung**

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{Gesamtstreuung}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{\text{erklärte Streuung}} + \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{nichterklärte Streuung}} \qquad (10.1)$$

**Idee für ein Gütemaß:**

**Messe den Anteil der Streuung der abhängigen Variablen, der durch die Regression erklärt wird:**

# 9. Einfache lineare Regression

**Bestimmtheitsmaß**

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} \hat{y}_i^2 - n\overline{y}^2}{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2} \in [0;1]$$

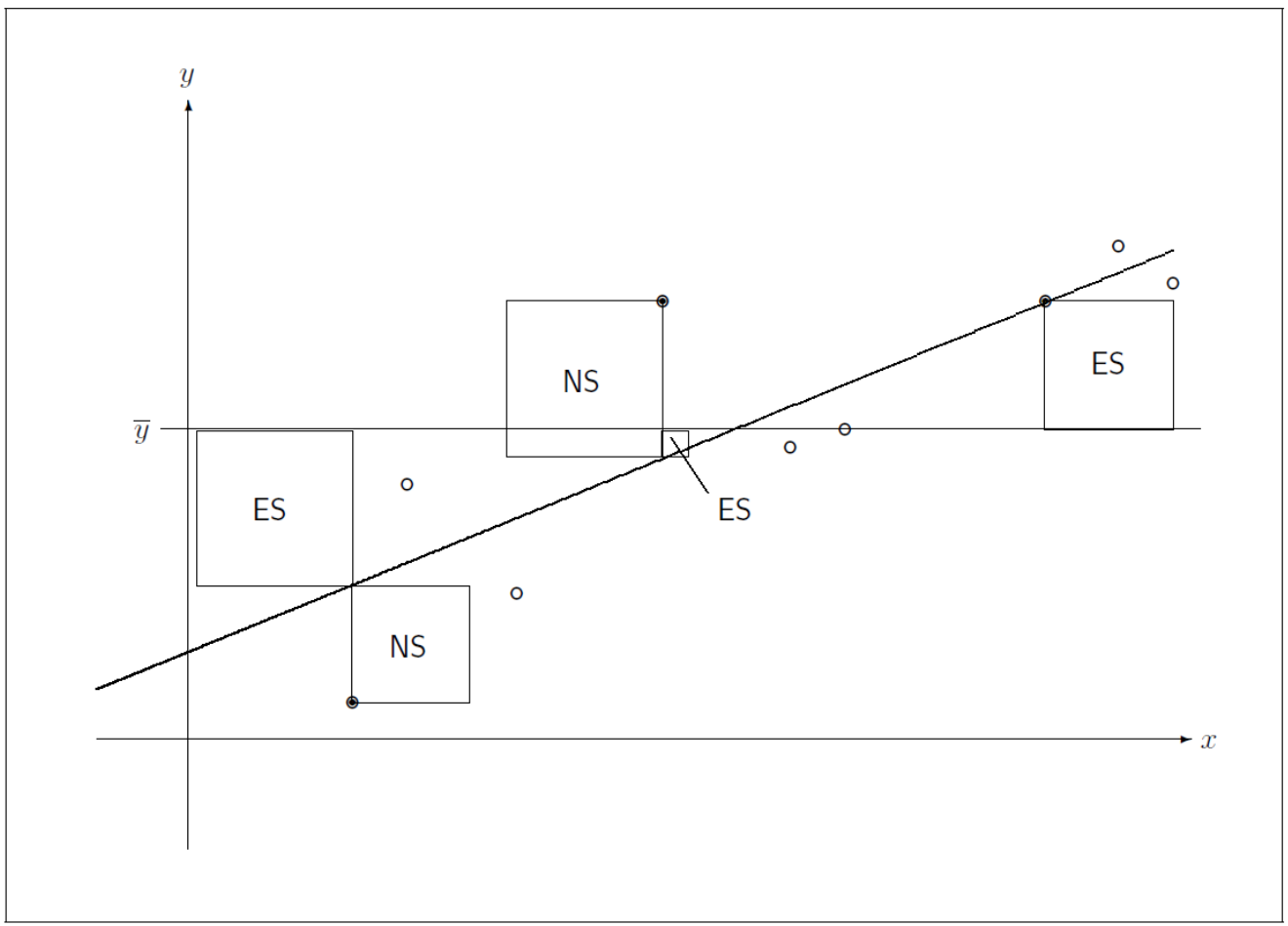
**Ziel: möglichst hohes Bestimmtheitsmaß!**

**Zusammenhang Bestimmtheitsmaß $\leftrightarrow$ Korrelationskoeffizient nach Bravais-Pearson**

$$R^2 = r_{xy}^2$$

# 9. Einfache lineare Regression

## Darstellung erklärte Streuung (ES) und nicht erklärte Streuung (NS):



# 9. Einfache lineare Regression

**Beispiel: Berechnen Sie das Bestimmtheitsmaß für das vorangegangene Beispiel (Dina Vier)**

1. **Lösungsmöglichkeit: ergänze vorherige Tabelle um eine Spalte:**

| $i$ | $y_i$ | $x_i$ | $x_i y_i$ | $x_i^2$ | $y_i^2$ |
|---|---|---|---|---|---|
| 1 | 22,1 | 3,1 | 68,51 | 9,61 | **488,41** |
| 2 | 16,3 | 2,2 | 35,86 | 4,84 | **265,69** |
| 3 | 17,8 | 2,1 | 37,38 | 4,41 | **316,84** |
| 4 | 25,9 | 2,9 | 75,11 | 8,41 | **670,81** |
| 5 | 20,5 | 2,4 | 49,2 | 5,76 | **420,25** |
| 6 | 28,4 | 3,3 | 93,72 | 10,89 | **806,56** |
| 7 | 12,1 | 1,5 | 18,15 | 2,25 | **146,41** |
| 8 | 22,5 | 3,3 | 74,25 | 10,89 | **506,25** |
| Summe | 165,6 | 20,8 | 452,18 | 57,06 | **3.621,22** |

$$r_{xy} = \frac{452,18 - 8 \cdot 2,6 \cdot 20,7}{\sqrt{57,06 - 8 \cdot 2,6^2}\sqrt{3.621,22 - 8 \cdot 20,7^2}} = 0,9 \qquad R^2 = 0,9^2 = 0,81$$

# 9. Einfache lineare Regression

**2. Möglichkeit:**

| $i$ | $y_i$ | $x_i$ | $\widehat{y}_i$ | $(\widehat{y}_i - \overline{y})^2$ | $(y_i - \overline{y})^2$ |
|---|---|---|---|---|---|
| 1 | 22,1 | 3,1 | **24,3** | **12,96** | **1,96** |
| 2 | 16,3 | 2,2 | **17,8** | **8,41** | **19,36** |
| 3 | 17,8 | 2,1 | **17,1** | **12,96** | **8,41** |
| 4 | 25,9 | 2,9 | **22,9** | **4,84** | **27,04** |
| 5 | 20,5 | 2,4 | **19,2** | **2,25** | **0,04** |
| 6 | 28,4 | 3,3 | **25,8** | **26,01** | **59,29** |
| 7 | 12,1 | 1,5 | **12,7** | **64,00** | **73,96** |
| 8 | 22,5 | 3,3 | **25,8** | **26,01** | **3,24** |
| Summe | 165,6 | 20,8 | 165,6 | **157,44** | **193,30** |

$$R^2 = \frac{157,44}{193,3} = 0,81$$

**Einmal mehr: Zeit sind Punkte in der Klausur!**

# 9. Einfache lineare Regression

**Standardfehler**

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

**Relativer Standardfehler**

$$s_0 = \frac{s}{\overline{y}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\overline{y}}$$

# 9. Einfache lineare Regression

**Beispiel:** Berechnen Sie Standardfehler und relativen Standardfehler für den vorangegangenen Datensatz (Dina Vier)

**Lösung:**

Die nicht erklärte Streuung war $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 193,30 - 157,44 = 35,86$

und somit $s^2 = \frac{1}{8-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{35,86}{6} = 5,98$

Damit erhält man den **Standardfehler** $s = \sqrt{s^2} = \sqrt{5,98} = 2,45$

und den **relativen Standardfehler:** $s_0 = \frac{s}{\overline{y}} = \frac{2,45}{20,7} = 0,118 = 11,8\%$