

Deskriptive Statistik

Prof. Dr. Christoph Hanck

Wintersemester 2022/2023

Überblick

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

0 Motivation

- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

Motivation

Anwendungsbereiche für statistische Methoden

- Wirtschaftswissenschaften (Empirische Wirtschaftsforschung, Ökonometrie)
- Ingenieurwissenschaften (Technometrie)
- Biologie/Medizin (Biometrie)
- Verhaltenswissenschaften (Psychometrie)
- u.v.a.m. – überall dort, wo Daten anfallen!

- Aufdeckung von Zusammenhängen
(z.B. zwischen Arbeitslosigkeit und Inflation)
- Überwachung ökonomischer Aktivität
(z.B. Aktienkurse, Wechselkurse, Zinssätze, Rohstoff- und Immobilienpreise)
- Überprüfung von Theorien anhand von Daten
(z.B. Zusammenhang zwischen verfügbarem Einkommen und Konsumausgaben)

Flüchtlinge, Our World in Data

Motivation

- Wie hoch ist der Anteil aller anwesenden weiblich Studierenden im Hörsaal?
- Welche Körpergröße wird von 30 Prozent aller im Hörsaal anwesenden Personen nicht überschritten?
- Wie stark ist der Zusammenhang zwischen der Entwicklung der VW- und der BMW-Aktie?
- Welches Bruttoinlandsprodukt kann für 2023 erwartet werden?
- Welcher Anteil des gesamten deutschen Stromabsatzes entfällt auf die beiden größten Anbieter?
- Um wie viel Prozent ist das Preisniveau in Deutschland im Monat September 2022 gegenüber dem Vorjahresmonat gestiegen?

Coronavirus, Westen und Osten, Anteil Studierender



Statistisches Bundesamt

Niedriglöhne steigen stärker als Gehälter von Besserverdienden

In den vergangenen Jahren sind die Niedriglöhne erstmals stärker gewachsen als die Gehälter der Besserverdiener - besonders in Ostdeutschland. Als Grund führen die Statistiker den Mindestlohn an.

14.09.2020, 10:59 Uhr

Erstmals sind die Gehälter im Niedriglohnsektor prozentual stärker gestiegen als die der Besserverdiener. Damit habe es 2018 erstmals eine Tendenz zur Lohngleichung zwischen Gering- und Besserverdienden gegeben, teilte das Statistische Bundesamt am Montag mit. Besserverdieneende erzielten demnach das 3,27-Fache des Bruttostundenverdiensts von Geringverdienden, während es 2014 noch das 3,48-Fache gewesen sei.

"Besonders deutlich schließt sich die Lohnschere in Ostdeutschland", hieß es. Hier erzielten Besserverdiende im Jahr 2018 einen um das 2,80-Fache höheren Bruttostundenverdienst als Geringverdiende. 2014 war es noch das 3,31-Fache. In Westdeutschland war dieser Trend deutlich schwächer (3,47 im Jahr 2014 und 3,29 im Jahr 2018). Das Statistische Bundesamt hatte bereits 2014 einen Stopp der sogenannten Lohnspreizung konstatiert.

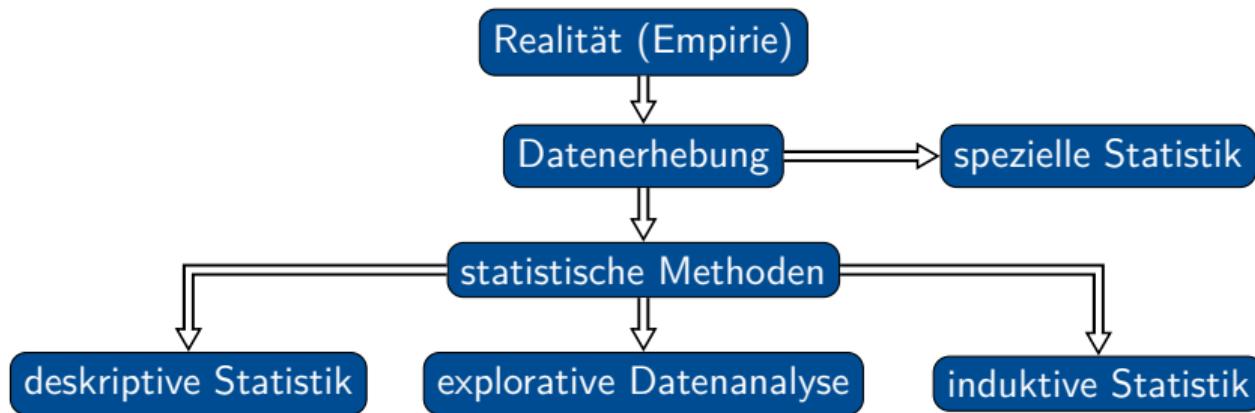
ANZEIGE

Das absolute **Lohngefälle in Deutschland** verringert sich damit allerdings nicht. Denn aufgrund des niedrigeren Ausgangswerts stieg der durchschnittliche Stundenlohn im Niedriglohnbereich um 1,37 Euro, während er bei den Besserverdienden um 2,74 Euro zulegte.

- Die Statistik liefert Werkzeuge zur Beantwortung solcher Fragen.
- Die Statistik hat drei Aufgabengebiete:
 - ① Statistische Erhebung
 - ② Statistische Aufbereitung: Gegenstand der **Deskriptiven Statistik**
 - ③ Statistische Analyse: Gegenstand der **Induktiven Statistik**

Motivation

Die verschiedenen Aufgabenbereiche und ihre Verbindungen:



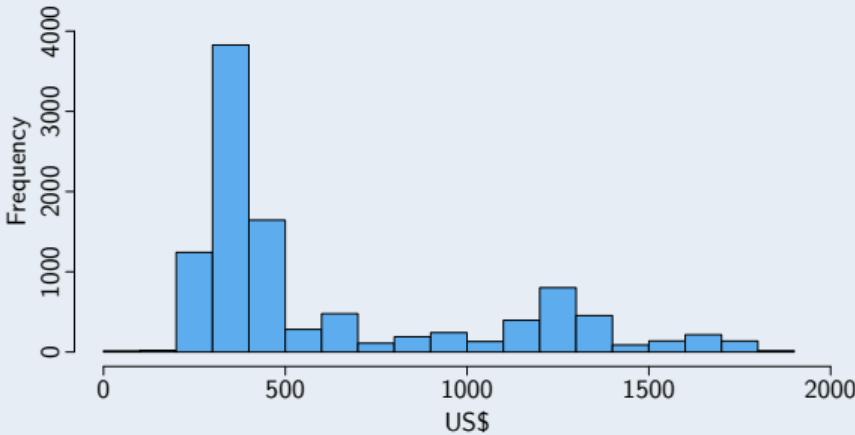
- **Deskriptive** (beschreibende) **Statistik**: Aufbereitung der statistischen Daten.
- Ziel: Übersicht mittels tabellarischer und grafischer Repräsentationen sowie geeigneter Kenngrößen.
- Vergleich von Datensätzen und Ableitung von Handlungssimplikationen.

Beispiel 0.1: Kommunalwahl NRW 2020.

Stimmenanteile Kommunalwahl NRW 2020

Motivation

Beispiel 0.2: Goldpreis (US-Dollar/Unze).



Daten: World Gold Council

- Wir befassen uns in der **induktiven** (schließenden) **Statistik** mit der Datenanalyse auf Basis von Wahrscheinlichkeitsmodellen.
- Ziel hierbei: Verifikation theoretischer Modelle anhand von Daten, Testen von Hypothesen über unbekannte Parameter.

Motivation

Datenquellen

- Verschiedene Institutionen stellen statistische Informationen bereit:
⇒ „amtliche“ und „nicht-amtliche“ Statistik.
- Amtliche Statistik: Deutsches Statistisches Bundesamt (DESTATIS), Statistische Landes- sowie kommunalstatistische Ämter.
- Nicht-amtliche Statistik: Verbände, Wirtschaftsforschungs- sowie Markt- und Meinungsforschungsinstitute: z.B. das Deutsche Institut für Wirtschaftsforschung (DIW) in Berlin, das IFO-Institut in München sowie das Rheinisch-Westfälische Institut für Wirtschaftsforschung (RWI) in Essen.
- Daten fallen aber auch und gerade außerhalb von Instituten und Ämtern an.

Obama, Statistische Analphabeten, Data Scientists1, Data Scientists2

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

Grundzüge der Datenerhebung

Merkmal, statistische Einheit, statistische Masse

- Vor der statistischen Analyse ist das Untersuchungsziel zu klären: Festlegung des zu quantifizierenden Phänomens und Operationalisierung des **theoretischen Konstrukts**.
- Keine klare Definition von z.B. Intelligenz, Bildung, Wohlfahrt und Inflation.
- Operationale Definitionen ordnen theoretischen Konstrukten Zählbegriffe der Statistik zu („**Adäquation**“).
- Eventuelle Diskrepanz zwischen Zählbegriff und theoretischem Konstrukt: **Adäquationsproblem**.
Homer

Beispiel 1.1:

Bildung \Rightarrow Schul- oder Studienabschluss, Anzahl der Schul- bzw. Studienjahre
sozialer Status \Rightarrow monatliches Einkommen, berufliche Stellung etc.

Grundzüge der Datenerhebung

Merkmal, statistische Einheit, statistische Masse

- Statistischer Zählbegriff: definiert eine beobachtbare Eigenschaft ⇒ statistisches **Merkmals**.
- Mögliche Erscheinungsformen des statistischen Merkmals: Merkmalswerte, **Merkmalsausprägungen** oder einfach Ausprägungen.
- Anzahl der Ausprägungen: endlich oder unendlich.
- Objekte, an denen das Merkmal in Erscheinung tritt: statistische Einheit, Untersuchungseinheit, **Merkmalsträger** oder kurz Element.
- Unterscheidung zwischen qualitativen (klassifikatorischen), ordinalen (komparativen) und quantitativen (metrischen bzw. kardinalen) Merkmalen.

Grundzüge der Datenerhebung

Merkmal, statistische Einheit, statistische Masse

Beispiel 1.2:

Merksam	Merkmalsausprägung	Merksamsträger
Studiendauer	1, 2, ..., n Semester	Absolventen der UDE im SS 2022
Verkaufszahlen	50, 100, 120, 200, ...	Handelstage eines Supermarktes im April 2022
Investitionen	500.000, 600.000, ... €	Unternehmen in NRW im März 2021
Energieverbrauch	3.500, 4.000, 6.000, ... kWh	Haushalte in Essen im Jahr 2022
Studienrichtung	VWL, BWL, ...	Studierende der UDE im WS 22/23
Religionszugehörigkeit	ev., rk., musl., ...	im Jahr 2021 in Deutschland lebende Personen

- **Qualitative Merkmale:** Unterscheidung der Ausprägungen durch ihre Art
⇒ höchstens abzählbar viele Ausprägungen. Beispiele: Haarfarbe, Geschlecht, Familienstand.
- **Ordinale Merkmale:** Rangordnung der Ausprägungen. Beispiele: Zensuren, Güteklassen, Windstärke.
- **Quantitative Merkmale:** Zählen bzw. Messen der Ausprägungen ⇒ Abstände interpretierbar.
Beispiele: Körpergröße, Einkommen, Beschäftigte.
- Unterteilung in **diskrete und stetige Merkmale:**
 - ▶ Diskretes Merkmal: abzählbar viele Ausprägungen (Anzahl Mitarbeiter).
 - ▶ Stetiges Merkmal: überabzählbar unendlich viele Ausprägungen (Köpergewicht).
 - ▶ Diskretes Merkmal mit sehr vielen Ausprägungen: quasi-stetig.

- Die Festlegung der Ausprägungen geschieht durch Zählen oder Messen.
- Messen ist die regelbasierte Zuordnung von Zahlen zu den Ausprägungen.
- Eine **Skala** stellt sicher, dass nach dem Messen dieselbe Ordnung der Merkmalsträger gemäß ihrer Ausprägungen vorliegt („relationstreue Abbildung in ein Zahlensystem“).
- Je nach Merkmalstyp verwendet man unterschiedliche Skalentypen.

- Qualitative Merkmale: Zuordnung von Zahlen zu den einzelnen Ausprägungen (**Nominalskala**). Idee: Unterscheidbarkeit der MA, zulässige Operationen: $=, \neq$. (Beispiele: Haarfarbe: 1 = rot, 2 = schwarz, 3 = blond, 4 = braun, 5 = weiß; Postleitzahlen).
- Ordinale Merkmale: **Ordinalskala**. Idee: Zuordnung drückt die Rangfolge der MA aus, Abstände sind nicht definiert. Zulässige Operationen: $=, \neq, <, >$.
- **Intervallskala**: zusätzliche Definition des Abstands zwischen je zwei Ausprägungen, das Verhältnis hingegen ist nicht definiert (Beispiel: Temperaturmessung). Zulässige Operationen: $=, \neq, <, >, +, -$. Transformation durch die Funktion $y = ax + b, a > 0$ ist zulässig, ohne dass sich der Skalentyp ändert.

Grundzüge der Datenerhebung

Messen und Skalieren

- Definition des Verhältnis zweier Ausprägungen: **Verhältnisskala** (Ratioskala).
- Verhältnisskalierte Merkmale haben einen natürlichen Nullpunkt \Rightarrow nur linear homogene Transformationen $y = ax, a > 0$ zulässig. Zulässige Operationen: $=, \neq, <, >, +, -, \cdot, /$; Beispiele: Entferungen (km, Meilen), Währungen (€, \$), Körpergröße (cm, Fuß)
- Besitzen Merkmale zusätzlich eine natürliche Skaleneinheit, verwendet man eine **Absolutskala**, die nicht transformiert werden kann (Beispiel: Anzahl Kinder in einem HH).
- Nominal- und Ordinalskala heißen topologische Skalen; Intervall-, Verhältnis- und Absolutskala bezeichnet man als **Kardinal- bzw. metrische Skalen**.

- Ein Merkmal bildet durch Messen seiner Ausprägungen jeden Merkmalsträger $\omega_j \in \Omega$ in eine Skala S ab, die Teilmenge der reellen Zahlen \mathbb{R} ist: $S \subset \mathbb{R}$.
- Man bezeichnet das Merkmal auch als „**statistische Variable** X “. Formal: $X : \Omega \longrightarrow S \subset \mathbb{R}$.

Definition 1.3: Beobachtung.

Das Bild von $\omega_j \in \Omega$ unter X heißt **Beobachtung** von X und wird mit x_j bezeichnet: $x_j = X(\omega_j)$. Die Gesamtheit aller Beobachtungen x_j sind die statistischen Daten (Datensatz), mit $j = 1, \dots, n$.

- Die einzelnen Beobachtungen müssen nicht alle verschieden sein, da mehrere Merkmalsträger dieselbe Ausprägung haben können.

Grundzüge der Datenerhebung

Messen und Skalieren

- Hingegen sind alle Elemente der Menge $\{X(\omega_j), \omega_j \in \Omega\}$ wegen der Mengendefinition verschieden.
- D.h. wir unterscheiden zwischen allen vorliegenden Beobachtungen x_j (z.B. 5, 6, 7, 7, 7, 8, 8, 9, 10) und allen verschiedenen Ausprägungen x_i (5, 6, 7, 8, 9, 10). Allgemein:

Definition 1.4: Merkmalsausprägung.

Die unterschiedlichen Ausprägungen von X werden als x_i bezeichnet: $x_i \in \{X(\omega_j), \omega_j \in \Omega\}$ mit $i = 1, \dots, m$, sodass $m \leq n$.

Grundzüge der Datenerhebung

Messen und Skalieren

- Zerlegen einer Skala S in abzählbar viele, halboffene Intervalle heißt **Klassierung** bzw. Klasseneinteilung.
- Die Klassenbildung kann entweder durch rechtsgeschlossene $(x'_{k-1}, x'_k]$ oder linksgeschlossene $[x'_{k-1}, x'_k)$ Intervalle mit $k \in \mathbb{N}$ erfolgen.
- Die Klassengrenzen x'_{k-1} und x'_k müssen nicht zu den Ausprägungen gehören.
- Eine Klassierung ist sinnvoll, wenn fast genauso viele Ausprägungen wie Beobachtungen vorliegen ($m \approx n$).

- Die Datengewinnung erfolgt durch **(Daten-)Erhebung**.
- Bei einer **Vollerhebung** werden alle Merkmalsträger einer Masse, bei einer **Teilerhebung** nur bestimmte Merkmalsträger aus Ω untersucht.
- Teilerhebungen können durch Ausgliedern nach bestimmten Ausprägungen (z.B. Bevölkerung unter 40 Jahren) oder durch Zufallsauswahl entstehen.
- Eine Teilerhebung ist leichter, schneller und vor allem billiger als eine Totalerhebung; dafür sind die Ergebnisse bei Zufallsauswahlen aber unsicherer als bei Vollerhebungen.

Grundzüge der Datenerhebung

Datengewinnung

- Je nach zeitlichem Bezug der Datenerhebung unterscheidet man zwischen **Längsschnitt- und Querschnitterhebung**.
- Bei einer Längsschnitterhebung werden Beobachtungen für aufeinanderfolgende Zeitpunkte/Perioden erhoben. Es resultiert eine **Zeitreihe** (z.B. die Entwicklung des BIPs 2000-2022).
- Bei Querschnitterhebungen haben alle Beobachtungen denselben Zeitbezug (z.B. Konsumausgaben der Haushalte in der 36. Woche eines Jahres).
- Die Kombination beider Erhebungsarten liefert **Paneldaten**.

Grundzüge der Datenerhebung

Datengewinnung

- Bei einer statistischen Masse lassen sich oft mehrere statistische Variablen X_1, X_2, \dots, X_g beobachten.
- Jeder Merkmalsträger $\omega_j, j = 1, \dots, n$ weist für jede Variable eine Beobachtung auf. Es liegen also insgesamt ng Beobachtungen vor.
- Ein solcher Datensatz, der aus mehreren Variablen besteht, nennt sich multivariat (mehrdimensional).
- Eine einzelne Variable liefert einen univariaten (eindimensionalen) Datensatz.

- Nachbereitung: Kapitel 1 und 2 des Buches von Prof. Assenmacher



- Kapitel 2 behandelt die Verteilung eindimensionaler Datensätze. Hiermit lassen sich Datensätze kompakt beschreiben.
- Vorbereitung: Kapitel 3 des Buches von Prof. Assenmacher

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen**
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

Urliste und Klassierung

Aufbereitung der Daten in der **Urliste** \Rightarrow möglichst kompakte Zusammenfassung der Informationen.

Beispiel 2.1: Fiktive Daten.

X bilde die Merkmalsträger einer Grundgesamtheit Ω im Umfang von $n = 20$ in die Beobachtungen $x_j, j = 1, \dots, n$ ab. Die Urliste sei:

11, 13, 15, 16, 12, 18, 14, 15, 17, 14, 12, 16, 13, 15, 17, 16, 15, 14, 13, 15.

Es gibt also $m = 8$ verschiedene Ausprägungen $x_i, i = 1, \dots, m$ von X :

11, 12, 13, 14, 15, 16, 17, 18.

Urliste und Klassierung

Bereits für $n = 20$ ist das Datenmaterial recht unübersichtlich. Ordne daher zunächst die Beobachtungen aufsteigend:

11, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 17, 17, 18.

Zähle, wie oft Ausprägung x_i vorkommt (wie viele ω_j in Ausprägung x_i abbilden), also wie sich die Beobachtungen auf die Ausprägungen verteilen.

Ausprägung	Anzahl
11	1
12	2
13	3
14	3
15	5
16	3
17	2
18	1

Urliste und Klassierung

- Nützlich: Aufteilung der Beobachtungen auf **Klassen** ⇒ Reduktion der Anzahl unterschiedlicher Ausprägungen durch **Klassierung**.
- Anstatt m verschiedener Ausprägungen liegen jetzt nur noch $K \ll m$ unterschiedliche Klassen vor.
- Das ist übersichtlicher, jedoch geht die Kenntnis der Verteilung der Daten innerhalb der Klassen verloren – ein *Trade-off*.

Urliste und Klassierung

Nützliche Klassierungen hängen vom jeweiligen Untersuchungsziel ab. Orientierungspunkte hierbei sind:

- ① Die Klassen sollten gleich breit (äquidistant) sein.
- ② Die Klassen sind disjunkt, d.h. überlappen sich nicht.
- ③ Die Klassen sollten angrenzen, d.h. keine Werte sollten ausgelassen werden.
- ④ Alle Daten sollten durch die Klassen erfasst werden.
- ⑤ Die Anzahl der Beobachtungen in den Randklassen (das sind die erste und letzte Klasse) sollte nicht zu gering sein.
- ⑥ Die am häufigsten vorkommende Ausprägung sollte in der Mitte ihrer Klasse liegen.

- Wegen 6. entwickelt man die Klassierung um die häufigste Beobachtung der Urliste; im obigen Beispiel ist dies $x_i = 15$.
- Bei einer Klassenbreite von 2 ergeben sich bei rechts- bzw linksoffenen Klassen folgende Häufigkeiten:

[von ... bis unter ...)	(über ... bis ...]
[10, 12) : 1	(10, 12] : 3
[12, 14) : 5	(12, 14] : 6
[14, 16) : 8	(14, 16] : 8
[16, 18) : 5	(16, 18] : 3
[18, 20) : 1	

- Die Klassenbildung „von... bis unter...“ suggeriert eine symmetrische Datenstruktur. Die Klassierung „über... bis...“ stimmt hier besser mit der nicht symmetrischen Verteilung der Beobachtungen überein.
- Wir werden Klassen nach dem Prinzip $(x'_{k-1}, x'_k]$ bilden.
- Die **absoluten Häufigkeiten** der Ausprägungen bezeichnen wir bei nicht klassierten Daten mit n_i bzw. bei klassierten Daten mit n_k .

Absolute und relative Häufigkeitsverteilungen

Uraliste und Klassierung - in

```
# Uraliste als einfachen Vektor abspeichern
x <- c(11, 12, 12, 13, 13, 13, 14, 14, 14, 15,
      15, 15, 15, 15, 16, 16, 16, 17, 17, 18)
# Anzeigen der Häufigkeiten mittels der Funktion table
table(x)

## x
## 11 12 13 14 15 16 17 18
## 1  2  3  3  5  3  2  1

# Klassieren des Vektors mit der Funktion cut
table(cut(x, breaks = c(10, 12, 14, 16, 18, 20), right = F))

##
## [10,12) [12,14) [14,16) [16,18) [18,20)
##       1       5       8       5       1

table(cut(x, breaks = c(10, 12, 14, 16, 18, 20)))

##
## (10,12] (12,14] (14,16] (16,18] (18,20]
##       3       6       8       3       0
```

Absolute und relative Häufigkeitsverteilungen

Absolute und relative Häufigkeitsfunktionen

Definition 2.2: Abs. bzw. rel. Häufigkeiten (auch: Häufigkeitsfunktionen).

Absolute Häufigkeiten bei nicht klassierten und klassierten Daten:

$$n(X = x_i) = n_i \quad i = 1, \dots, m \quad \text{bzw.} \quad n(x'_{k-1} < X \leq x'_k) = n_k \quad k = 1, \dots, K$$

n_i bzw. n_k dividiert durch n liefern **relative Häufigkeiten** h_i bzw. h_k :

$$h(X = x_i) = \frac{n_i}{n} = h_i \quad i = 1, \dots, m \quad \text{bzw.}$$

$$h(x'_{k-1} < X \leq x'_k) = \frac{n_k}{n} = h_k \quad k = 1, \dots, K$$

Definitionsgemäß gilt:

$$\sum_{i=1}^m n_i = n, \quad \sum_{k=1}^K n_k = n, \quad \sum_{i=1}^m h_i = 1, \quad \sum_{k=1}^K h_k = 1.$$

Absolute und relative Häufigkeitsfunktionen

- Relative Häufigkeiten h_i bzw. h_k multipliziert mit 100 liefern Prozentsätze.
- Eine **Häufigkeitsverteilung** ordnet die Häufigkeiten den entsprechenden Ausprägungen x_i , $i = 1, \dots, m$ zu.
- Bei nicht klassierten Daten ist $\{(x_1, n_1), (x_2, n_2), \dots, (x_m, n_m)\}$ die absolute, $\{(x_1, h_1), (x_2, h_2), \dots, (x_m, h_m)\}$ die relative Häufigkeitsverteilung.

Absolute und relative Häufigkeitsverteilungen

Absolute und relative Häufigkeitsfunktionen - in R

```
# Absolute Häufigkeiten können, wie gesehen, mittels table() ermittelt werden.  
# Für relative Häufigkeiten teilen wir die abs. Hfgk. einfach durch n.  
table(x) / length(x)  
  
## x  
## 11   12   13   14   15   16   17   18  
## 0.05 0.10 0.15 0.15 0.25 0.15 0.10 0.05  
  
table(cut(x, breaks = c(10, 12,14,16,18,20))) / length(x)  
  
##  
## (10,12] (12,14] (14,16] (16,18] (18,20]  
##     0.15    0.30    0.40    0.15    0.00
```

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Die einfachste Form der Darstellung ist die Häufigkeitstabelle:

nicht klassiert			klassiert		
x_i	n_i	h_i	$x'_{k-1} < X \leq x'_k$	n_k	h_k
11	1	0,05	$k = 1, \dots, 4$		
12	2	0,10	(10,12]	3	0,15
13	3	0,15	(12,14]	6	0,30
14	3	0,15	(14,16]	8	0,40
15	5	0,25	(16,18]	3	0,15
16	3	0,15			
17	2	0,10			
18	1	0,05			
\sum		20	20	1,00	1,00

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Beispiel 2.3: Studierende nach Fächergruppen in Deutschland WS 2019/20.

x_i	n_i	h_i
Geisteswissenschaften	332 440	0,1150
Sport	29 207	0,0101
Rechts-, Wirtschafts- und Sozialwissenschaften	1 082 326	0,3744
Mathematik, Naturwissenschaften	322 086	0,1114
Humanmedizin/Gesundheitswissenschaften	186 835	0,0646
Agrar-, Forst- und Ernährungswissenschaften, Veterinärmedizin	63 381	0,0219
Ingenieurwissenschaften	774 687	0,2680
Kunst, Kunstwissenschaft	95 521	0,0330
Sonstige Fächer und ungeklärt	4 566	0,0016
\sum	2891049	1

Quelle: Statistisches Bundesamt: DESTATIS

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Beispiel 2.4: Monatliches Haushaltsnettoeinkommen 2016 — rel. Hfkt.

$[x'_{k-1}, x'_k)$	h_k
0 bis 1300 €	0,163
1300 bis 1700 €	0,091
1700 bis 2600 €	0,206
2600 bis 3600 €	0,178
3600 bis 5000 €	0,175
5000 bis 18000 €	0,186
\sum	1

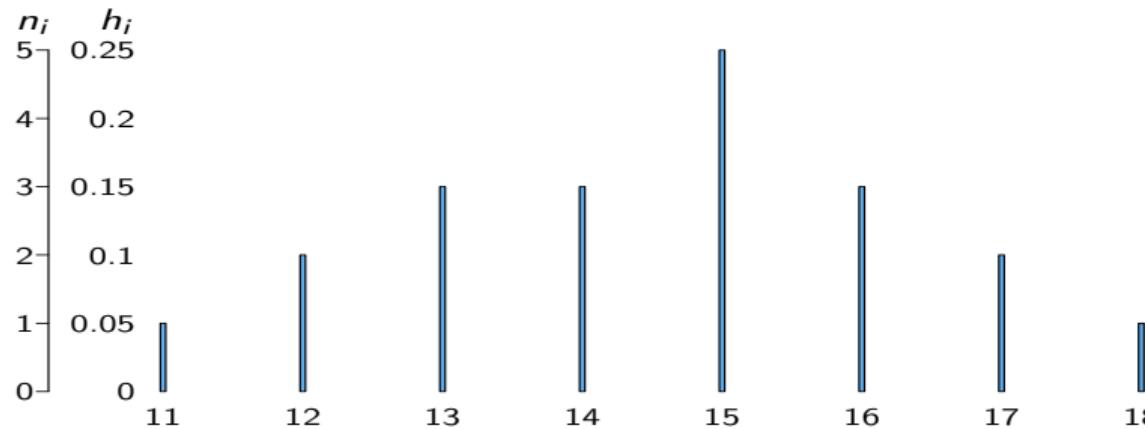
Quelle: Statistisches Bundesamt: DESTATIS

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

- Grafische Darstellung veranschaulicht, kann aber anfällig für optische Manipulation sein.
- Die einfachste Grafik ist das **Stabdiagramm**:

```
barplot(table(x), col = "steelblue2", axes = F, space = 25)
```



- Da der Übergang von absoluten zu relativen Häufigkeiten nur eine Maßstabsänderung ist, sind n_i und h_i an derselben Ordinate abgetragen.

- Zeichnet man die Stäbe dicker, entsteht ein Säulen- bzw. Balkendiagramm.
- Bei einem Rechteckdiagramm schließen die Balken ohne Freiräume an.
- Die grafische Darstellung von Häufigkeitsverteilungen klassierter Daten geschieht über **Histogramme**.
- Ein Histogramm besteht aus Rechtecken, die über den an der Abszisse abgetragenen Klassen so errichtet werden, dass die Flächen proportional zu den Klassenhäufigkeiten sind („Flächentreue“).
Tödliche Tiere
- Wir entwickeln Histogramme über die **Häufigkeitsdichtefunktion**.

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

- Häufigkeitsdichtefunktion: Unterscheidung zwischen Klassierung mit äquidistanter und variabler Klassenbreite ist überflüssig.
- Häufigkeitsdichte: n_k^* bzw. h_k^* ist der Quotient aus Klassenhäufigkeit n_k bzw. h_k und Klassenbreite $\Delta_k = x'_k - x'_{k-1}$.

Definition 2.5: Absolute bzw. relative Häufigkeitsdichtefunktion n_k^* bzw. h_k^* .

$$n_k^* = \begin{cases} n(x'_{k-1} < X \leq x'_k) / \Delta_k & \text{für } x'_{k-1} < x \leq x'_k \\ 0 & \text{sonst} \end{cases}$$

$$h_k^* = \begin{cases} h(x'_{k-1} < X \leq x'_k) / \Delta_k & \text{für } x'_{k-1} < x \leq x'_k \\ 0 & \text{sonst} \end{cases}$$

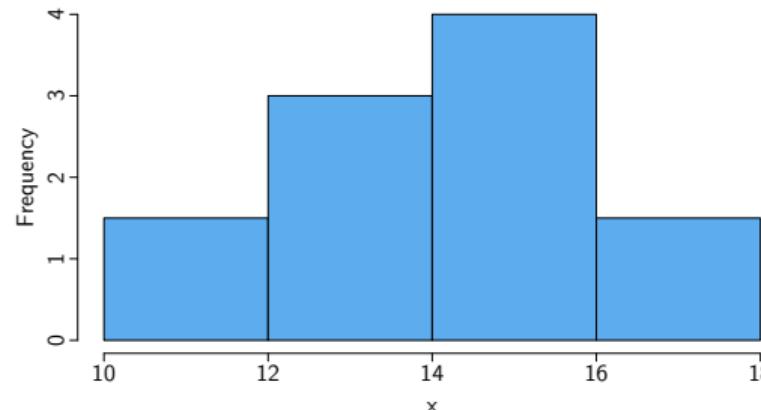
Histogramms.R

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Für unser Beispiel ergibt sich:

$$n_k^* = \begin{cases} 1,5 & \text{für } 10 < x \leq 12 \\ 3 & \text{für } 12 < x \leq 14 \\ 4 & \text{für } 14 < x \leq 16 \\ 1,5 & \text{für } 16 < x \leq 18 \\ 0 & \text{sonst} \end{cases} \quad h_k^* = \begin{cases} 0,075 & \text{für } 10 < x \leq 12 \\ 0,15 & \text{für } 12 < x \leq 14 \\ 0,20 & \text{für } 14 < x \leq 16 \\ 0,075 & \text{für } 16 < x \leq 18 \\ 0 & \text{sonst} \end{cases} .$$

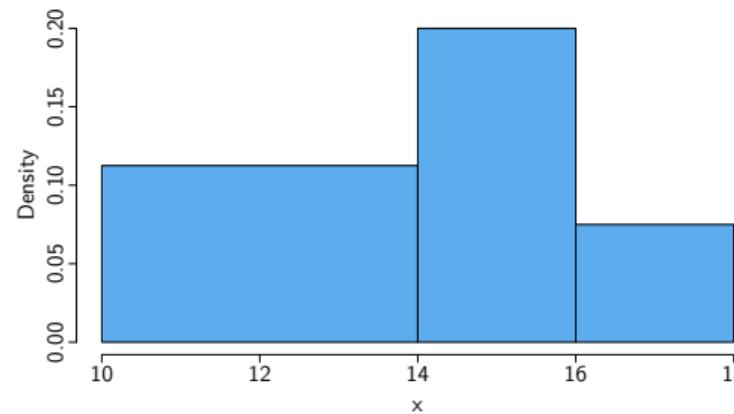


Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Fasst man die erste und zweite Klasse zusammen, resultieren unterschiedliche Klassenbreiten:

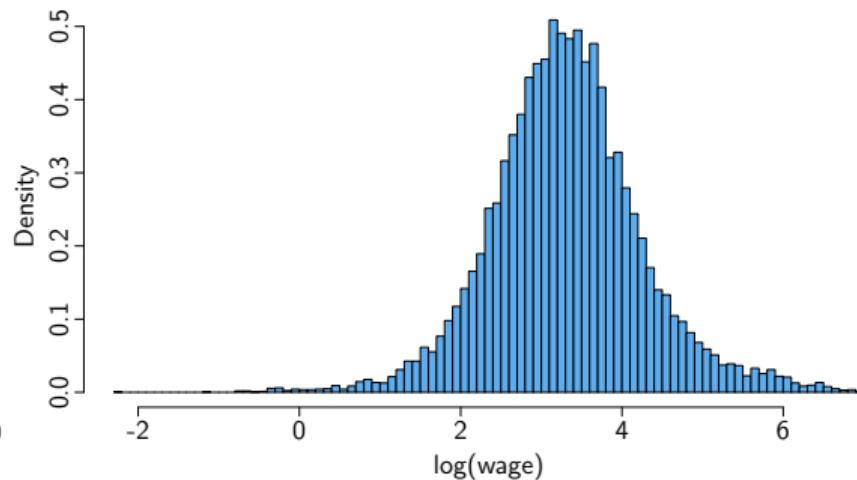
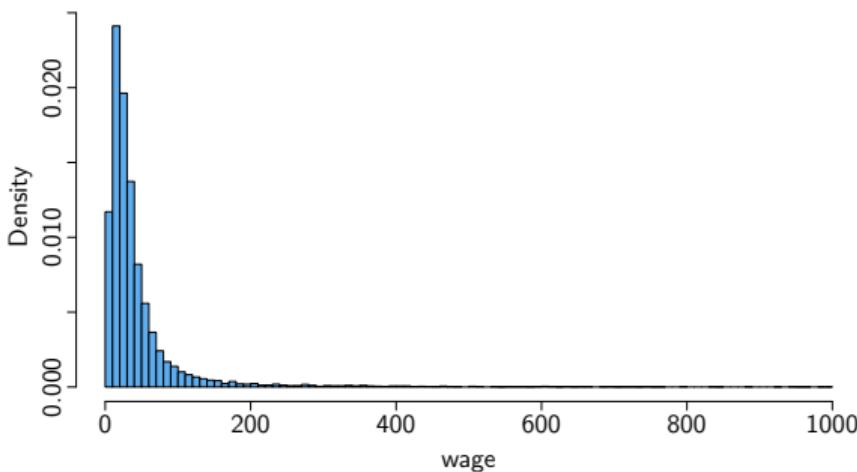
$x'_{k-1} < X \leq x'_k, k = 1, 2, 3$	n_k	n_k^*	h_k	h_k^*
(10,14]	9	2,25	0,45	0,1125
(14,16]	8	4,00	0,40	0,2000
(16,18]	3	1,50	0,15	0,0750



Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken - Histogramme in R

```
load("Daten/soep2013.rda")
sl <- bdpequiv$i1110313 / bdpequiv$e1110113 # Haushaltseinkommen / Arbeitsstunden
sl <- sl[sl > 0 & sl < 1000 & !is.na(sl)]    # Entferne Ausreißer, fehlende Werte
hist(sl, breaks = 100, col = "steelblue2", main = "", freq = F, xlab = "wage")
hist(log(sl), breaks = 100, col = "steelblue2", main = "", freq = F, xlab = "log(wage)")
```



Deutsche Stundenlöhne (Euro/Stunde), Quelle: SOEP 2013

- Beim **Kreissektorendiagramm** verhalten sich, analog zum Histogramm, die Flächeninhalte der Kreissektoren proportional zu den Häufigkeiten; es kann auch für nicht klassierte Daten erstellt werden.
- Das Kreissektorendiagramm wird oft bei nominal skalierten Merkmalen angewendet.

www.graphitti-blog.de

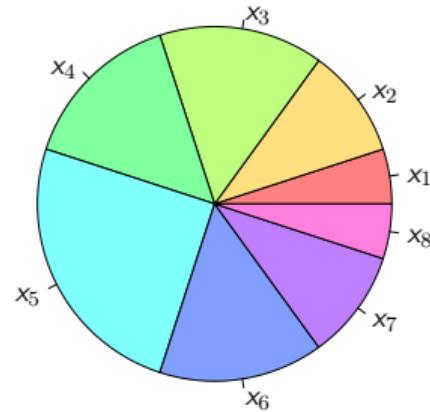
Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Um die proportionalen Kreissektorwinkel zu erhalten, dividiert man 360° durch n und multipliziert dann mit den einzelnen absoluten Häufigkeiten n_i .

x_i	11	12	13	14	15	16	17	18
n_i	1	2	3	3	5	3	2	1
α_i	18°	36°	54°	54°	90°	54°	36°	18°

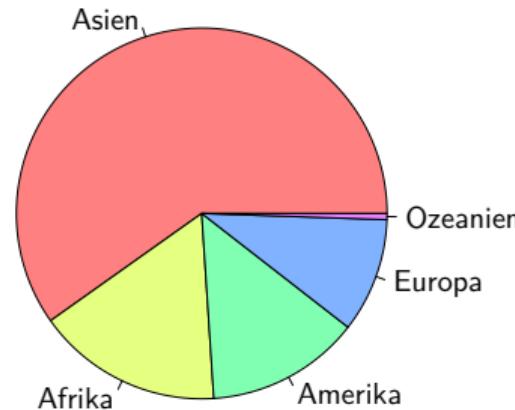
```
pie(table(x), labels = paste0("$x_",$ 1:8, "$"), col = rainbow(8, s = 0.5))
```



Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

```
Weltbev <- c(4437, 1203, 997, 740, 40)
names(Weltbev) <- c("Asien", "Afrika", "Amerika", "Europa", "Ozeanien")
pie(Weltbev, col = rainbow(5, s = 0.5))
```

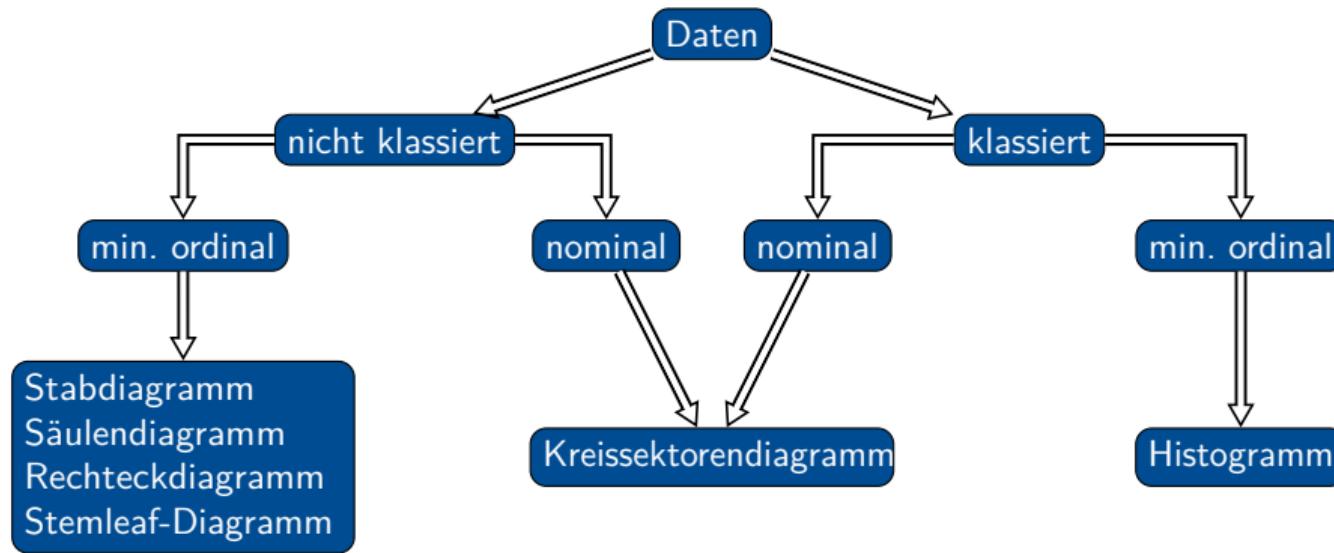


Wie auch hier ersichtlich, sind Kreisdiagramme oft suboptimal. Einige Empfehlungen und Beispiele, wie man es nicht machen sollte.

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Zusammenfassung einiger Möglichkeiten der grafischen Darstellung der Daten:



Beispiel 2.6: Geodaten.

Es sind aber in letzter Zeit noch viele andere Arten entwickelt worden, Daten grafisch darzustellen. Siehe zum Beispiel [hier](#).

Kahoot!

In den Vorlesungunterlagen finden Sie Verweise auf Aufgaben, die mit Hilfe der Lern- und Quiz-App Kahoot während der Vorlesung bearbeitet werden. Die App können Sie im App-Store bzw. Playstore herunterladen. Die Teilnahme ist auch über den Browser per <https://kahoot.it> möglich.

Kahoot für Android:



Kahoot für iOS:



Tabellen und Grafiken

Kahoot!

Absolute und relative Häufigkeitsverteilungen

Absolute und rel. Häufigkeitssummenfunktion

- Bei mindestens ordinal skalierten Merkmalen, für die „größer, kleiner, gleich“ definiert sind, ist von Interesse, welche(r) Anzahl/Anteil der Beobachtungen nicht größer als ein Wert $x \in \mathbb{R}$ ist.
- Hierzu kumulieren wir die Häufigkeiten (Aufsummierung) \Rightarrow Unterscheidung zwischen klassierten und nicht klassierten Daten.
- Die Anzahl der Beobachtungen, die höchstens gleich x sind, ist die Summe der absoluten Häufigkeiten n_i der x_i , für die gilt: $x_i \leq x$:

Definition 2.7: Kumulierte absolute Häufigkeit.

Diese Summe $N(X \leq x) = \sum_i n_i$ mit $x_i \leq x$ heißt **kumulierte absolute Häufigkeit**.

Absolute und relative Häufigkeitsverteilungen

Absolute und relative Häufigkeitssummenfunktion

Definition 2.8: Absolute Häufigkeitssummenfunktion.

$$N(x) = \begin{cases} 0 & \text{für } x < x_{i=1} \text{ (kleinste Ausprägung)} \\ N(X \leq x_i) & \text{für } x_i \leq x < x_{i+1}, i = 1, \dots, m-1 \\ n & \text{für } x \geq x_{i=m} \text{ (größte Ausprägung).} \end{cases}$$

Absolute und relative Häufigkeitsverteilungen

Absolute und relative Häufigkeitssummenfunktion

Der Anteil der Beobachtungen, die einen Wert x nicht überschreiten, ergibt sich analog als kumulierte relative Häufigkeit:

Definition 2.9: Empirische Verteilungsfunktion.

$$H(X \leq x) = \sum_i h_i = \frac{1}{n} N(X \leq x)$$

Variiert x , erhält man die **empirische Verteilungsfunktion** (auch: relative Häufigkeitssummenfunktion):

$$H(x) = \begin{cases} 0 & \text{für } x < x_{i=1} \text{ (kleinste Ausprägung)} \\ H(X \leq x_i) & \text{für } x_i \leq x < x_{i+1}, \quad i = 1, \dots, m-1 \\ 1 & \text{für } x \geq x_{i=m} \text{ (größte Ausprägung).} \end{cases}$$

SOEP.R

Absolute und relative Häufigkeitssummenfunktion

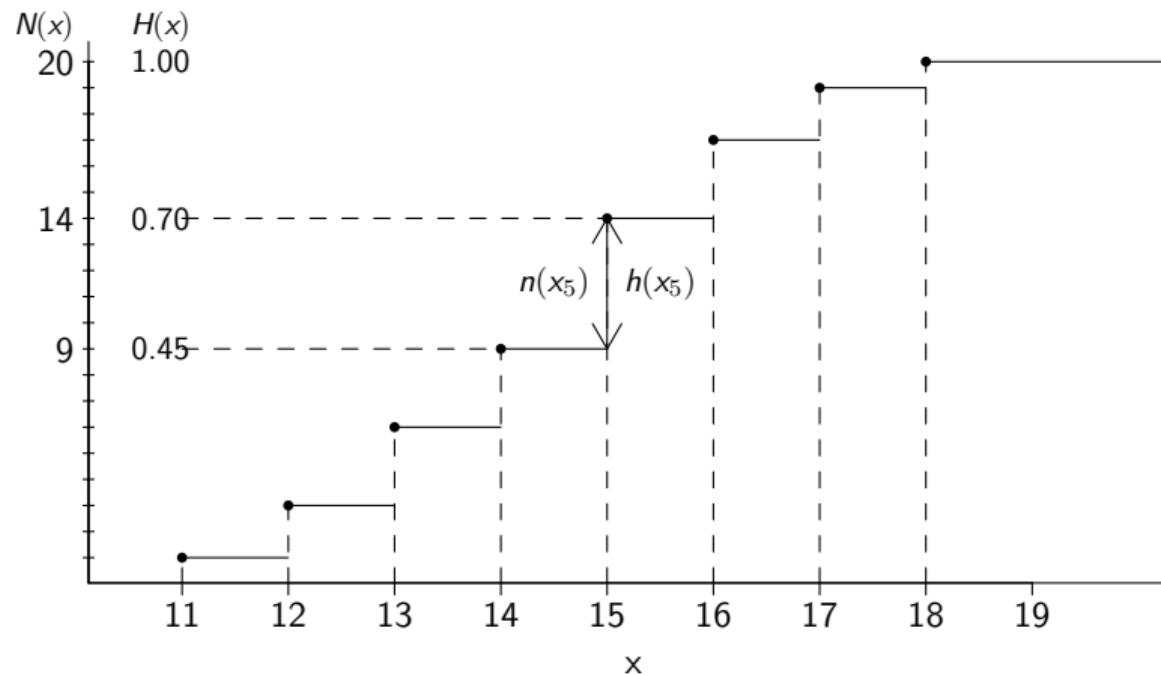
Für unser Beispiel ergeben sich folgende Häufigkeitssummen:

i	x_i	n_i	$N(X \leq x_i)$	h_i	$H(X \leq x_i)$
1	11	1	1	0,05	0,05
2	12	2	3	0,10	0,15
3	13	3	6	0,15	0,30
4	14	3	9	0,15	0,45
5	15	5	14	0,25	0,70
6	16	3	17	0,15	0,85
7	17	2	19	0,10	0,95
8	18	1	20	0,05	1,00

Absolute und relative Häufigkeitsverteilungen

Absolute und relative Häufigkeitssummenfunktion

Es folgt:



Absolute und relative Häufigkeitsverteilungen

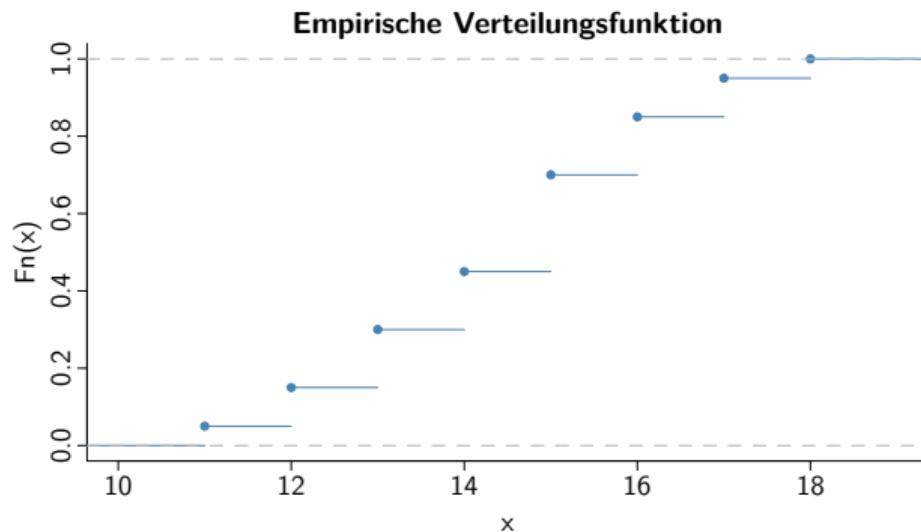
Absolute und relative Häufigkeitssummenfunktion - in **R**

```
H_x <- ecdf(x) # „empirical cumulative distribution function“
```

```
H_x(14)
```

```
## [1] 0.45
```

```
plot(H_x, col = "steelblue", main = "Empirische Verteilungsfunktion")
```



Absolute und relative Häufigkeitssummenfunktion

- Für jedes x liefern Grafik/kumulierte Häufigkeitstabelle die Anzahl bzw. den Anteil der Beobachtungen, die x nicht übersteigen.
- 6 bzw. 30% der Beobachtungen sind kleiner als $x = 13,5$; ebenfalls 6 bzw. 30% sind kleiner als oder gleich 13.
- Die strenge Ungleichung resultiert für Werte von x , die nicht auch als Beobachtungen vorliegen.
- Die Anzahl bzw. den Anteil der Beobachtungen, die größer als $x = a$, aber nicht größer als $x = b > a$ sind, berechnet man als

$$N(a < X \leq b) = N(b) - N(a) \quad \text{bzw.}$$

$$H(a < X \leq b) = H(b) - H(a).$$

- Anteil der Beobachtungen, die größer als 14, aber nicht größer als 17 sind:

$$H(14 < X \leq 17) = H(17) - H(14) = 0,95 - 0,45 = 0,50$$

Quantile

- Quantile ergeben sich aus der Umkehrung der zur Häufigkeitssummenfunktion führenden Fragestellung.
- Gesucht wird jetzt eine Ausprägung der Variablen X , die von vorgegebenen $p \cdot 100\%$ der Beobachtungen ($0 < p < 1$) nicht überschritten wird.
- Diesen Wert nennt man **p -Quantil** x_p .
- x_p , das nicht notwendigerweise im Datensatz vorkommen muss, teilt die Beobachtungen in zwei Teile so auf, dass $p \cdot 100\%$ der Beobachtungen kleiner oder gleich und $(1 - p) \cdot 100\%$ größer als x_p sind.

Quantile

- Ist X eine stetige Variable, kann x_p leicht bestimmt werden: Die Vorgabe von p legt $H(x)$ fest als: $H(x) = p$; Auflösen nach x liefert das p -Quantil.
- In der Praxis liegen in Datensätzen immer nur endlich viele Ausprägungen vor.
- Bei einer diskreten Variablen hat $H(x)$ Sprungstellen \Rightarrow für bestimmte p existiert das oben definierte p -Quantil nicht.
- Wir definieren daher x_p als die Ausprägung, bei der mindestens $p \cdot 100\%$ aller Beobachtungen denselben oder einen kleineren, und mindestens $(1 - p)100\%$ denselben oder einen größeren Wert aufweisen.
- Aus der Definition folgt, dass von n Beobachtungen mindestens np (gerundet) Beobachtungen kleiner oder gleich und mindestens $(1 - p)n$ (gerundet) Beobachtungen größer oder gleich x_p sind.

Absolute und relative Häufigkeitsverteilungen

Quantile

- Berechnung der p -Quantile: Sortiere zunächst die Beobachtungen aufsteigend:
 $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$.
- Das Produkt np bestimmt die Beobachtung, die den Datensatz auf die gewünschte Weise unterteilt.
- Da der Platzierungsindex immer ganzzahlig ist, bestimme den ganzzahligen Teil g von np :
 $g = \text{int}(np)$.
- Die Abkürzung „int“ steht für integer (ganze Zahl); z.B. $\text{int}(7,89) = 7$.

Dann gilt:

Definition 2.10: Quantil (Einzelbeobachtungen).

$$x_p = \begin{cases} x_{(g+1)} & \text{für } np > \text{int}(np) = g \\ x_{(g)} & \text{für } np = \text{int}(np). \end{cases}$$

Absolute und relative Häufigkeitsverteilungen

Quantile - in

```
n <- 11
# zufällig generierte und der Übersichtlichkeit halber sortierte Zahlen:
(x <- sort(round(rchisq(n, df = 2), 3)))
## [1] 0.179 0.339 0.391 0.594 1.204 1.576 2.470 2.622 3.119 3.864
## [11] 10.862

p <- 0.5                      # Median
quantile(x, type = 1, p = p)   # beachte "type = 1"
## 50%
## 1.576

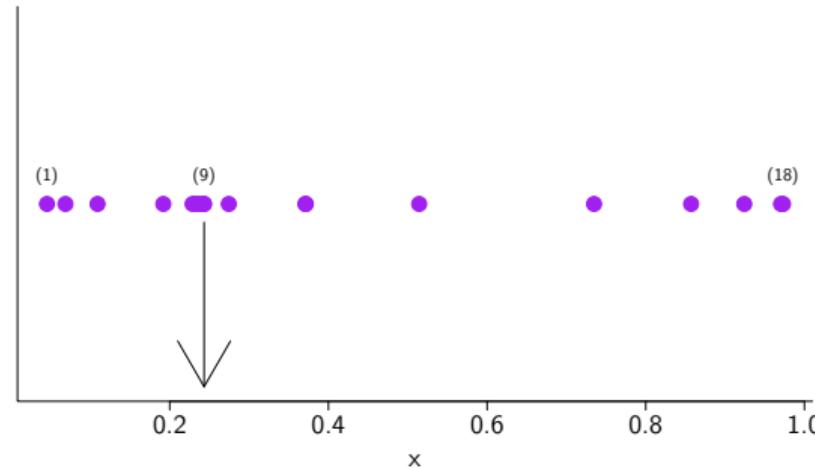
# per "Hand":
(g <- floor(n * p))
## [1] 5

x[g + 1]
## [1] 1.576
```

Absolute und relative Häufigkeitsverteilungen

Quantile - in

```
n <- 18
x <- rbeta(n, shape1 = .7, shape2 = .9) # Ein paar Zufallszahlen erzeugen
p <- 0.5
(xMedian <- quantile(x, type = 1, p = 0.5)) # Datenerzeugung unterdrückt
##           50%
## 0.2433145
```



Quantile - in

Es gibt noch viele weitere Definitionen für Quantile (vgl. bspw. `?quantile`), die etwa auf verschiedene Art und Weise interpolieren, wenn $n \cdot p > g$.

```
n <- 100
x <- rnorm(n)           # 100 zufällig generierte Zahlen
p <- 0.1                 # erstes Dezil

quantile(x, type = 1, p = p)
##          10%
## -1.052453

quantile(x, p = p)       # der default (type=7)
##          10%
## -0.9733217
```

Manchmal wird auch das arithmetische Mittel aus $x_{(g)}$ und $x_{(g+1)}$ gebildet (s. Buch).

Quantile

Beispiel 2.11: Quantile.

Wenn wir der Einfachheit halber davon ausgehen, dass es konstant eine Milliarde Chinesen gibt, findet sich hier ein Artikel über die 50 reichsten Chinesen, also der Entwicklung des $100 \cdot (1 - 50/1\text{e}9) = 99,99995\%$ -Quantils der chinesischen Einkommensverteilung: SZ

Absolute und relative Häufigkeitsverteilungen

Quantile

- Erstelle alternativ zunächst die empirische Verteilungsfunktion.
- Suche dann das x_{i^*} , für das gilt: $H(x_{i^*}) = p$.
- Andernfalls: Existiert für ein p ein x_{i^*} mit $H(x_{i^*-1}) < p$ und $H(x_{i^*}) > p$, so ist das Quantil $x_p = x_{i^*}$.

Definition 2.12: Quantil (Häufigkeitsverteilung).

$$x_p = \begin{cases} x_{i^*} & \text{für } H(x_{i^*-1}) < p \text{ und } H(x_{i^*}) > p \\ x_{i^*} & \text{für } H(x_{i^*}) = p. \end{cases}$$

Quantile

- Es kann hilfreich sein, Intervallgrenzen für die Beobachtungen so festzulegen, dass pro Intervall (nahezu) gleiche Besetzungszahlen bzw. -anteile resultieren.
- So unterscheidet man z.B. in Terzile mit $x_{0,\bar{3}}$ und $x_{0,\bar{6}}$ oder Quartile mit $x_{0,25}$, $x_{0,5}$ und $x_{0,75}$.
- Analog: Aufteilung auf fünf Intervalle mit Quintilen, auf 10 Intervalle mit Dezilen und auf 100 Intervalle mit Perzentilen.

Beispiel 2.13: Quintile.

Hier sehen Sie ein Beispiel für Reichtumsquintile.

Quantile

Kahoot!

- Dieses Kapitel behandelte erste Schritte zur deskriptiven Analyse von Daten durch bspw. Häufigkeitssummenfunktionen \Rightarrow Entwicklung von Quantilen und Betrachtung verschiedener Arten von Datensätzen: ungeordnete Datensätze, Häufigkeitsverteilungen und klassierte Daten.
- Nachbearbeitung: Kapitel 3 des Buches von Prof. Assenmacher.
- Nächste Vorlesung: Parameter eindimensionaler Häufigkeitsverteilungen \Rightarrow Möglichkeit des Vergleichs verschiedener Datensätze.
- Vorbereitung: Kapitel 4.1 und 4.2 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter**
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

- Obwohl Häufigkeitsverteilungen und Häufigkeitssummenfunktionen die Informationen im Datensatz bündeln, reicht dies oft noch nicht aus.
- Problematisch wird dies insbesondere beim Vergleich mehrerer großer Datensätze.
- Daher sind Maßzahlen nützlich, die Eigenschaften eines Datensatzes zusammenfassen.
- Solche Maßzahlen heißen Parameter eines Datensatzes bzw. einer Verteilung.

- Die Nützlichkeit von Parametern hängt von der Fragestellung und der Skalierung der statistischen Variablen ab.
- Die meisten Parameter sind lediglich für metrisch skalierte Merkmale sinnvoll.
- Arten von Parametern:
 - ▶ Lageparameter
 - ▶ Streuungsparameter
 - ▶ Kurtosisparameter
 - ▶ Schiefeparameter

Lageparameter

- **Lageparameter** (auch: Lagemaße) charakterisieren komprimiert die Lage des Datensatzes bzw. seiner Häufigkeitsverteilung.
- Sie haben daher dieselbe Dimension wie das erfasste Merkmal.
- Lageparameter müssen bestimmte Mindestanforderungen erfüllen, so genannte **axiomatische Grundlagen**.
- Hierzu gehören:
 - ▶ Identitätsaxiom
 - ▶ Inklusionsaxiom
 - ▶ Translationsaxiom
 - ▶ Homogenitätsaxiom.

- Haben alle n Beobachtungen denselben Wert c , soll auch der Lageparameter Θ_L diesen Wert annehmen (**Identitätsaxiom**):

$$x_1 = x_2 = \dots = x_n = c \Rightarrow \Theta_L = c.$$

- Θ_L soll zwischen der kleinsten und größten Beobachtung liegen (**Inklusionsaxiom**):

$$x_{(1)} = \min_j x_j \leq \Theta_L \leq x_{(n)} = \max_j x_j, \quad j = 1, \dots, n.$$

- Eine Verschiebung des gesamten Datensatzes auf der Merkmalsachse um $d \neq 0$ soll Θ_L ebenfalls um d verschieben (**Translationsaxiom**):

$$\Theta_L(x_1 + d, \dots, x_n + d) = \Theta_L(x_1, \dots, x_n) + d$$

Lageparameter

Axiomatische Grundlagen

- Eine Veränderung aller absoluten Häufigkeiten n_i , $i = 1, \dots, m$, mit dem Faktor $\lambda > 0$ beeinflusst Θ_L nicht:

$$\Theta_L(x_1, \dots, x_m, n_1, \dots, n_m) = \Theta_L(x_1, \dots, x_m, \lambda n_1, \dots, \lambda n_m).$$

Dieses **Homogenitätsaxiom** verlangt also, dass Lageparameter homogen vom Grade null in den n_i sind. Datensätze mit gleichen relativen Häufigkeitsverteilungen haben dann auch gleiche Lageparameter.

- Es existieren verschiedene Lageparameter, deren Anwendbarkeit von der Skalierung der Variablen abhängt.

- Der **Modus** ist der einfachste Lageparameter und kann für jede Skalierung erstellt werden.

Definition 3.1: Modus x_M .

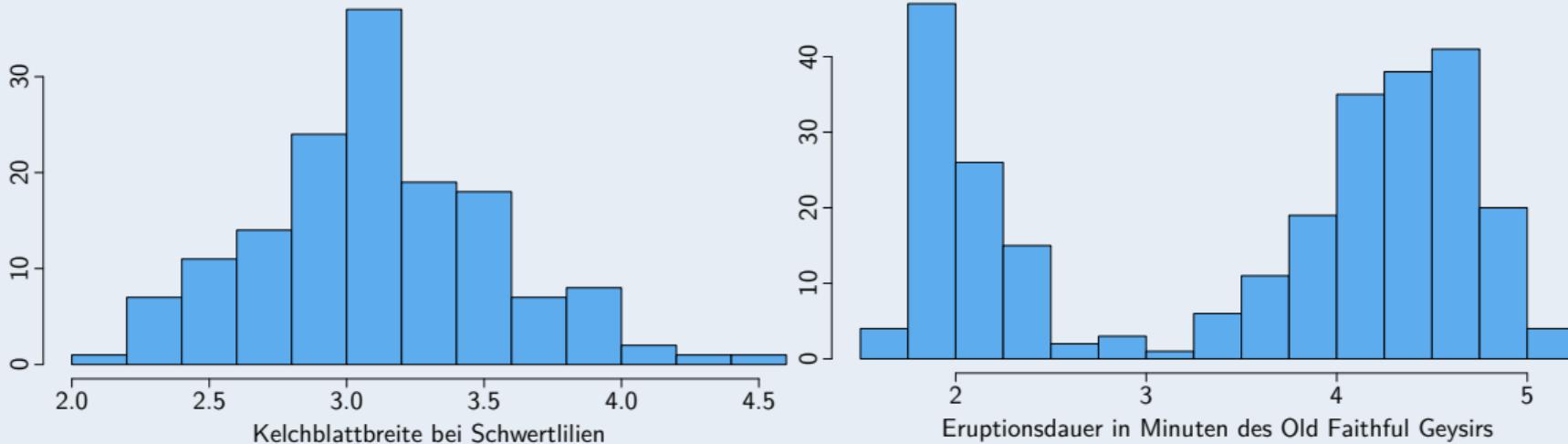
Es gilt: $x_M = x_{i^*}$, wobei i^* der Index der Ausprägungen mit der größten absoluten Häufigkeit ist.

- x_M ist mglw. für multimodale (mehrgipflige) Verteilungen nicht aussagekräftig.

Lageparameter

Modus x_M

Beispiel 3.2: Unimodale vs. bimodale Verteilung.

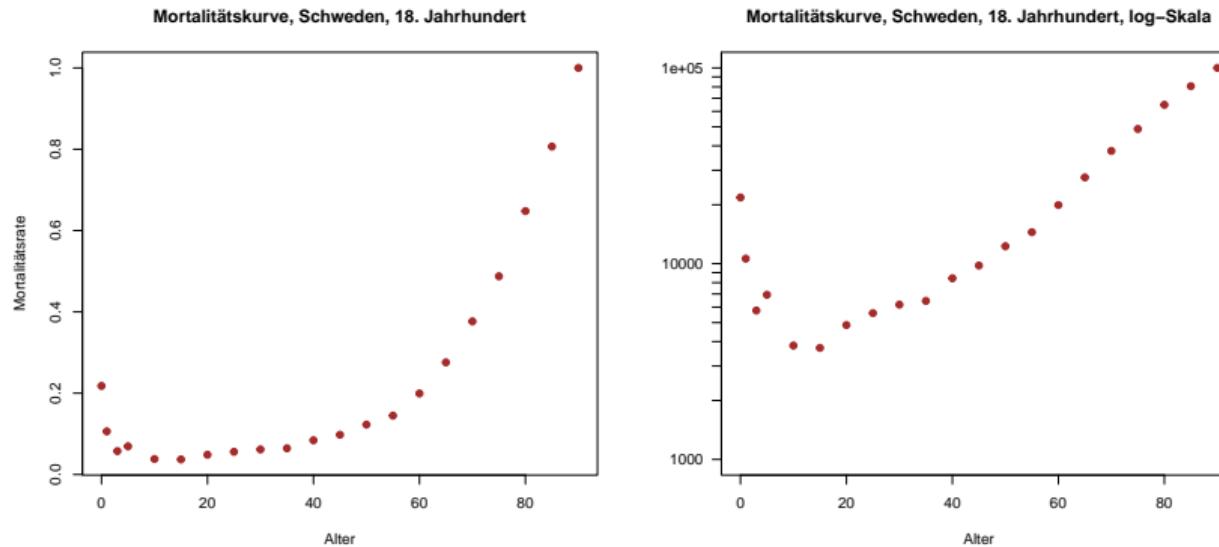


Unimodale (links) im Vergleich mit einer bimodalen Verteilung.

Lageparameter

Modus x_M

- Hier ein Beispiel für eine multimodale Verteilung anhand der Mortalitätskurve für Schweden im 18. Jahrhundert.



Quelle: [Human Life Table Database](#).

Lageparameter

Median/0,5-Quantil x_{Med}

- Als Lageparameter eignet sich die Beobachtung, die den Datensatz in zwei (fast) gleich große Hälften teilt, das **0,5-Quantil** bzw. **Median** oder Zentralwert.
- x_{Med} ist also der Wert, bei dem mindestens 50% aller Beobachtungen kleiner oder gleich und mindestens 50% aller Beobachtungen größer oder gleich x_{Med} sind.

Definition 3.3: Median x_{Med} .

$$x_{\text{Med}} = \begin{cases} x_{(\frac{n+1}{2})} & \text{für } n \text{ ungerade} \\ x_{(\frac{n}{2})} & \text{für } n \text{ gerade} \end{cases}$$

Lageparameter

Median/0,5-Quantil x_{Med}

- Zum Thema: scienceblogs.de
- Bei einer stetigen statistischen Variablen mit metrischer Skala berechnet man manchmal $x_{\text{Med}} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$ (s. Buch).
- Wie alle Quantile kann er für mindestens ordinal skalierte Merkmale berechnet werden.
- x_{Med} reagiert unempfindlich auf extreme Werte (**statistische Ausreißer**).
- Der Median besitzt die **Minimierungseigenschaft**:

$$\sum_{j=1}^n |x_j - x_{\text{Med}}| \leq \sum_{j=1}^n |x_j - a| \quad \text{für } a \in \mathbb{R} \quad \text{und } a \neq x_{\text{Med}}$$

Lageparameter

Arithmetisches Mittel \bar{x}

- Der am häufigsten verwendete Lageparameter ist das **arithmetische Mittel \bar{x}** (umgangssprachlich: Durchschnitt).
- Es ist definiert als Summe aller Beobachtungen, dividiert durch die Anzahl der Beobachtungen \Rightarrow nur aussagekräftig bei metrisch skalierten Daten.

Definition 3.4: Arithmetisches Mittel \bar{x} (unklassierte Daten).

- Daten als Urliste vorhanden:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

- Daten als Häufigkeitsverteilung:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^m x_i n_i = \sum_{i=1}^m x_i h_i$$

Lagemaße 1

Kahoot!

Lageparameter

Arithmetisches Mittel \bar{x}

Definition 3.5: Arithmetisches Mittel \bar{x} (klassierte Daten).

Das exakte arithmetische Mittel kann bei klassierten Daten nur dann berechnet werden, wenn die arithmetischen Klassenmittel \bar{x}_k bekannt sind. Berechne andernfalls mit den Klassenmitten m_k ein approximatives arithmetisches Mittel $\hat{\bar{x}}$:

$$\hat{\bar{x}} = \frac{1}{n} \sum_{k=1}^K m_k n_k = \sum_{k=1}^K m_k h_k$$

bzw.

$$\bar{x} = \frac{1}{n} \sum_{k=1}^K \bar{x}_k n_k = \sum_{k=1}^K \bar{x}_k h_k$$

- \bar{x} hat wesentliche Eigenschaften:

- Schwerpunkteigenschaft:

$$\sum_{j=1}^n (x_j - \bar{x}) = \sum_{j=1}^n x_j - n\bar{x} = 0,$$

wegen $\sum_{j=1}^n x_j = n\bar{x}$.

- Transformationseigenschaft: $y_j = \alpha + \beta x_j \Rightarrow \bar{y} = \alpha + \beta \bar{x}$
- Minimierungseigenschaft: (siehe Übungsaufgaben für einen Beweis)

$$\sum_{j=1}^n (x_j - \bar{x})^2 \leq \sum_{j=1}^n (x_j - a)^2 \quad \text{für } a \in \mathbb{R}$$

Lageparameter

Arithmetisches Mittel \bar{x}

- Die **Schwerpunkteigenschaft** besagt, dass die Summe aller Abweichungen von \bar{x} gleich Null ist.
- Beispiel: Für $x_j = \{3, 4, 4, 9\}$ gilt $n = 4$ und

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{1}{4}(3 + 4 + 4 + 9) = 5.$$

Die Summe aller Ausprägungen $\sum_{j=1}^n x_j = 20$ lässt sich auch schreiben als $n\bar{x} = 4 \cdot 5 = 20$.

- Ebenso einfach lässt sich die **Transformationseigenschaft** beweisen. Transfomiert man die Originaldaten x_j linear zu $y_j = \alpha + \beta x_j$, ergibt sich $\sum_{j=1}^n y_j = \sum_{j=1}^n (\alpha + \beta x_j) = n\alpha + \beta \sum_{j=1}^n x_j$. Nach Division durch n folgt

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j = \alpha + \beta \frac{1}{n} \sum_{j=1}^n x_j = \alpha + \beta \bar{x}.$$

Lageparameter

Arithmetisches Mittel \bar{x}

Beispiel 3.6: Transformationseigenschaft.

Temperaturmessung in Grad Celsius (${}^{\circ}\text{C} \Rightarrow x$) und Grad Fahrenheit (${}^{\circ}\text{F} \Rightarrow y$) möglich

Umrechnungsformel: $y = 32 + 1,8x$

Tagesdurchschnittstemperaturen in Grad Celsius: $x = \{23, 25, 24, 19, 22, 23, 24\}$.

Wochendurchschnittstemperaturen in Grad Celsius: $\bar{x} = 22,8571$

Tagesdurchschnittstemperaturen in Grad Fahrenheit: $y = \{73,4; 77; 75,2; 66,2; 71,6; 73,4; 75,2\}$.

Wochendurchschnittstemperaturen in Grad Fahrenheit: $\bar{y} = 73,1428$

Transformation: $\bar{y} = 32 + 1,8\bar{x} = 32 + 1,8 \cdot 22,8571 = 73,1428$

- Das arithmetische Mittel reagiert empfindlich auf **statistische Ausreißer**:
- Da \bar{x} die Summe der quadrierten(!) Abweichungen minimiert, haben sehr große und sehr kleine Beobachtungen großen Einfluss auf \bar{x} .
- Bei Vorliegen von Ausreißern kann \bar{x} daher irreführend sein.
- Verfügbares Einkommen.R
- Wenn diese Beobachtungen „untypisch“ sind, kann man sie eliminieren oder ihren Einfluss durch eine geringe Gewichtung reduzieren.
- Man erhält so **robuste** arithmetische Mittel.

Lageparameter

Arithmetisches Mittel \bar{x}

- Sollen die $\alpha \cdot 100\%$ kleinsten und $\alpha \cdot 100\%$ größten Beobachtungen für das arithmetische Mittel ausgeschlossen werden, bestimmt sich die Anzahl der zu eliminierenden Beobachtungen aus $g = \text{int}(\alpha n)$.
- Entferne nun die g kleinsten und die g größten Beobachtungen und berechne das arithmetische Mittel für die verbleibenden $n - 2g$ Daten:

$$\bar{x}_\alpha = \frac{1}{n - 2g} \sum_{j=g+1}^{n-g} x_{(j)}$$

- Man bezeichnet \bar{x}_α als das **α -getrimmte arithmetische Mittel**.

Lageparameter

Modus, Median und Arithmetisches Mittel - in 

```
x <- c(11, 12, 12, 13, 13, 13, 14, 14, 14, 15,  
      15, 15, 15, 15, 16, 16, 16, 17, 17, 18)
```

```
which.max(table(x)) # Welcher Wert kommt am häufigsten vor?
```

```
## 15  
## 5
```

```
median(x) # Median (anhand des sortierten Vektors bereits ablesbar)
```

```
## [1] 15
```

```
# Arithm. Mittel  
sum(x) / length(x)
```

```
## [1] 14.55
```

```
mean(x)
```

```
## [1] 14.55
```

```
mean(x, trim = 0.1) # entferne die kleinsten/größten 10% der Beobachtungen
```

```
## [1] 14.5625
```

```
mean(x[3:18])
```

```
## [1] 14.5625
```

Lageparameter

Geometrisches Mittel \bar{x}_G

- Das arithmetische Mittel gibt z.B. bei zeitabhängigen Messzahlen nicht den „richtigen“ Durchschnitt an.
- **Zeitabhängige Messzahlen** resultieren, wenn zwei Beobachtungen einer statistischen Variable zu unterschiedlichen Zeitpunkten ins Verhältnis gesetzt werden: **Wachstums- bzw. Aufzinsungsfaktoren**.

Beispiel 3.7: Verkehrstote.

Siehe hierzu [diese Berechnungen](#) des Verkehrsclubs Deutschland.

- Sie werden meist für äquidistante Zeitpunkte oder Perioden erstellt.
- Wachstumsfaktoren verbinden Beobachtungen so über die Zeit, dass Nachfolger das Produkt aus Vorgänger und Wachstumsfaktor ist.
- Dies ist nur bei metrisch skalierten Variablen sinnvoll.

Lageparameter

Geometrisches Mittel \bar{x}_G

- Für eine Zeitreihe y_0, y_1, \dots, y_n (z.B. der Kapitalstock einer Volkswirtschaft) sind die entsprechenden Wachstumsfaktoren x_j pro Periode j definiert als

$$x_j = \frac{y_j}{y_{j-1}}, \quad j = 1, \dots, n.$$

- Analog hierzu erhält man den Gesamtwachstumsfaktor als y_n/y_0 .
- Wegen

$$\frac{y_n}{y_0} = \frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdot \dots \cdot \frac{y_{n-1}}{y_{n-2}} \cdot \frac{y_n}{y_{n-1}} = x_1 \cdot x_2 \cdot \dots \cdot x_n$$

lässt sich y_n mit dem Produktoperator \prod darstellen als

$$y_n = y_0 x_1 \cdot \dots \cdot x_n = y_0 \prod_{j=1}^n x_j.$$

Lageparameter

Geometrisches Mittel \bar{x}_G

Das **geometrische Mittel** \bar{x}_G ist der **durchschnittliche Wachstumsfaktor**, der über die n Perioden konstant bleibt und y_0 auf seinen Endwert y_n anwachsen lässt.

Definition 3.8: Geometrisches Mittel \bar{x}_G .

Somit gilt: $y_0(\bar{x}_G)^n = y_n$, oder, nach \bar{x}_G aufgelöst:

$$\bar{x}_G = \sqrt[n]{x_1 \cdot \dots \cdot x_n} = \left(\prod_{j=1}^n x_j \right)^{\frac{1}{n}}.$$

- Alternativ gilt wegen $\exp(\log(x)) = x$ und $\log(a \cdot b) = \log(a) + \log(b)$, dass

$$\bar{x}_G = \exp\left(\frac{1}{n} \sum_{j=1}^n \log(x_j)\right)$$

- Aus der Definition der Wachstumsrate w_y folgt

$$w_{y_j} = \frac{y_j - y_{j-1}}{y_{j-1}} = \frac{y_j}{y_{j-1}} - 1 = x_j - 1$$

- Die **durchschnittliche Wachstumsrate** folgt aus \bar{x}_G als

$$\bar{w}_y = \bar{x}_G - 1.$$

Lageparameter

Geometrisches Mittel \bar{x}_G

Beispiel 3.9: Bruttoinlandsprodukt der EU.



BIP der EU 27 (links) sowie die entsprechenden Wachstumsfaktoren.

Quelle: STATISTA

Lageparameter

Geometrisches Mittel \bar{x}_G

Beispiel 3.9: Fortsetzung.

Von 1995 bis 2019 wuchs das Bruttoinlandsprodukt der Europäischen Union mit den Raten (in %)

4.73; 2.71; 4.25; 4.50; 5.92; 4.70; 3.64; 2.69; 4.56; 4.25; 5.86; 6.13; 3.26 -4.51; 3.68; 3.19; 0.53; 1.14; 2.26; 3.74; 2.70; 3.98; 3.45; 3.26.

Wandele die Wachstumsraten zunächst in Wachstumsfaktoren um. Zur ersten Wachstumsrate von 4,73% gehört der Wachstumsfaktor $x_1 = 1.0473$, zur zweiten $x_2 = 1.0271$ usw.

Beispiel 3.9: Fortsetzung.

Als durchschnittlichen Wachstumsfaktor erhält man dann

$$\bar{x}_G = (1,0473 \cdot 1,0271 \cdot \dots \cdot 1,0326)^{\frac{1}{24}} \approx 1.0334.$$

Die durchschnittliche Wachstumsrate beträgt somit 3,34%. Das arithmetische Mittel der Wachstumsraten würde ein „falsches“ Ergebnis liefern. Bei großem Anfangswert y_0 und/oder langer Laufzeit kann ein geringfügiger Fehler bereits zu einer beträchtlichen Reaktion von y_n führen.

Lageparameter

Geometrisches Mittel \bar{x}_G - in 

```
# Dieselben Daten wie im vorhergehenden Beispiel
## load("Daten/BIPEU.rda")
library(psych)
head(BIP, n = 3)
## [1] 6.34 6.64 6.82

w_fkt <- BIP[2:25] / BIP[1:24]
geometric.mean(w_fkt)
## [1] 1.033373
```

Lagemaße 2

Kahoot!

Lageparameter

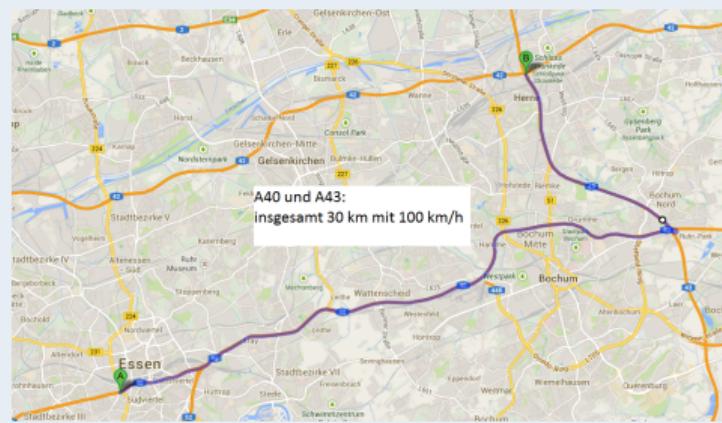
Harmonisches Mittel \bar{x}_H

- Zahlreiche Merkmale haben eine Dimension, die aus verschiedenen Grunddimensionen hervorgeht; sie sind daher mehrdimensional.
- So ist die Geschwindigkeit als Kilometer pro Stunde (km/h) definiert, d.h. ihre Dimension ist ein Quotient aus der Dimension „Länge“ im Zähler und der Dimension „Zeit“ im Nenner.
- Bei Merkmalen, deren Dimension als Quotient vorliegt, können die Häufigkeiten in der Dimension des Zählers oder des Nenners vorliegen.
- Haben sie die Dimension des Nenners, nutze das übliche \bar{x} ; haben sie die Dimension des Zählers, nutze das **harmonische Mittel**.
- Es setzt metrisch skalierte Merkmale mit nur positiven Ausprägungen voraus.

Lageparameter

Harmonisches Mittel \bar{x}_H

Beispiel 3.10: Falls Sie nach Herne wollen.



Lageparameter

Harmonisches Mittel \bar{x}_H

Definition 3.11: Harmonisches Mittel \bar{x}_H .

Das harmonische Mittel \bar{x}_H ist bei Einzelbeobachtungen bzw. häufigkeitsverteilten Daten als Kehrwert des arithmetischen Mittels der reziproken Beobachtungen definiert: Aus

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n \frac{1}{x_j}$$

folgt

$$\bar{x}_H = \bar{x}^{-1} = \frac{n}{\sum_{j=1}^n \frac{1}{x_j}} \quad \text{bzw.} \quad \bar{x}_H = \frac{n}{\sum_{i=1}^m \frac{n_i}{x_i}} = \frac{1}{\sum_{i=1}^m \frac{h_i}{x_i}}$$

Lageparameter

Harmonisches Mittel \bar{x}_H

Beispiel 3.12: Geschwindigkeiten.

Ein Auto fährt eine Strecke von 1000km mit den angegebenen Geschwindigkeiten und der dazugehörigen Dauer bzw. Streckenlänge.

x_i (km/h)	60	100	110	120	Σ
n_i (Stunden)	1,5	3	5	0,5	10

Die Dauer, mit der eine bestimmte Geschwindigkeit gefahren wird, stellt die Häufigkeiten in der Dimension Zeit dar; d.h. in der Nenner-Dimension.

Daher ist die Durchschnittsgeschwindigkeit als gewogenes arithmetisches Mittel zu berechnen. Da die gesamte Fahrzeit $n = 10$ Stunden beträgt, ergibt sich:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^4 x_i n_i = \frac{1}{10} (60 \cdot 1,5 + 3 \cdot 100 + 5 \cdot 110 + 0,5 \cdot 120) = \frac{1}{10} 1000 = 100 \text{ (km/h)}.$$

Lageparameter

Harmonisches Mittel \bar{x}_H

Beispiel 3.12: Fortsetzung.

Liegen die Häufigkeiten so vor, dass 90 km mit 60 km/h, 300 km mit 100 km/h usw. gefahren wurden, haben sie die Zähler-Dimension „Länge“.

x_i	(km/h)	60	100	110	120	Σ
n_i	(km)	90	300	550	60	1000

Berechne daher die Durchschnittsgeschwindigkeit mit dem harmonischen Mittel:

$$\bar{x}_H = \frac{1000}{\frac{90}{60} + \frac{300}{100} + \frac{550}{110} + \frac{60}{120}} = 100 \text{ (km/h)}.$$

Das arithmetische Mittel wäre:

$$\bar{x} = \frac{1}{1000}(60 \cdot 90 + 100 \cdot 300 + 110 \cdot 550 + 120 \cdot 60) = 103,1$$

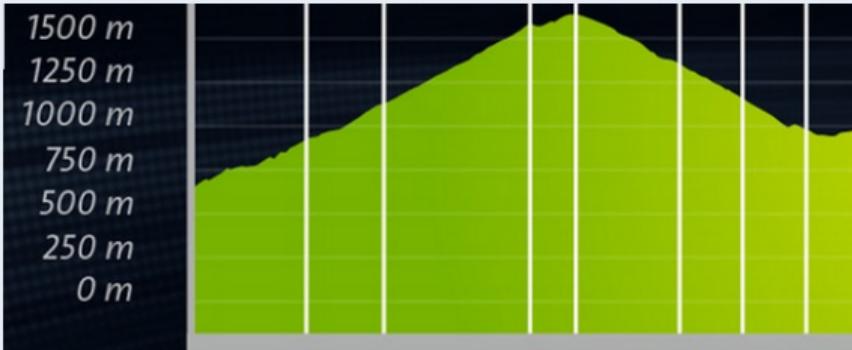
HarmonicMean.R

```
# library(psych)
kmh <- c(60, 100, 110, 120)
n.i <- c(90, 300, 550, 60)
harmonic.mean(rep(kmh, n.i))
## [1] 100
```

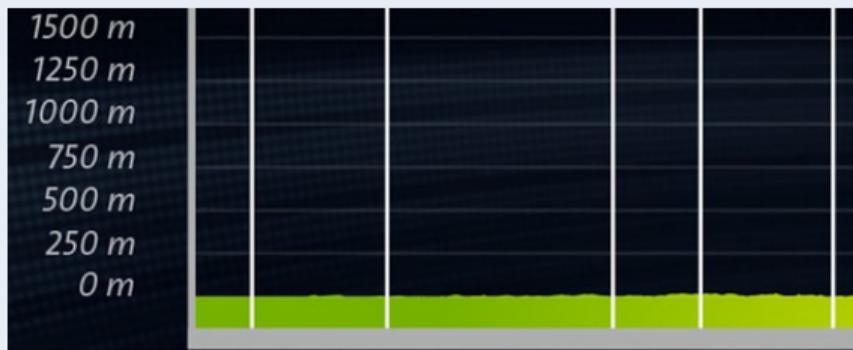
Lageparameter

Harmonisches Mittel \bar{x}_H - Kahoot

Beispiel 3.13: Tour de France.



10km bergauf mit 20km/h, 10km bergab mit 60km/h



20km mit 40 km/h

Lagemaße 3

Kahoot!

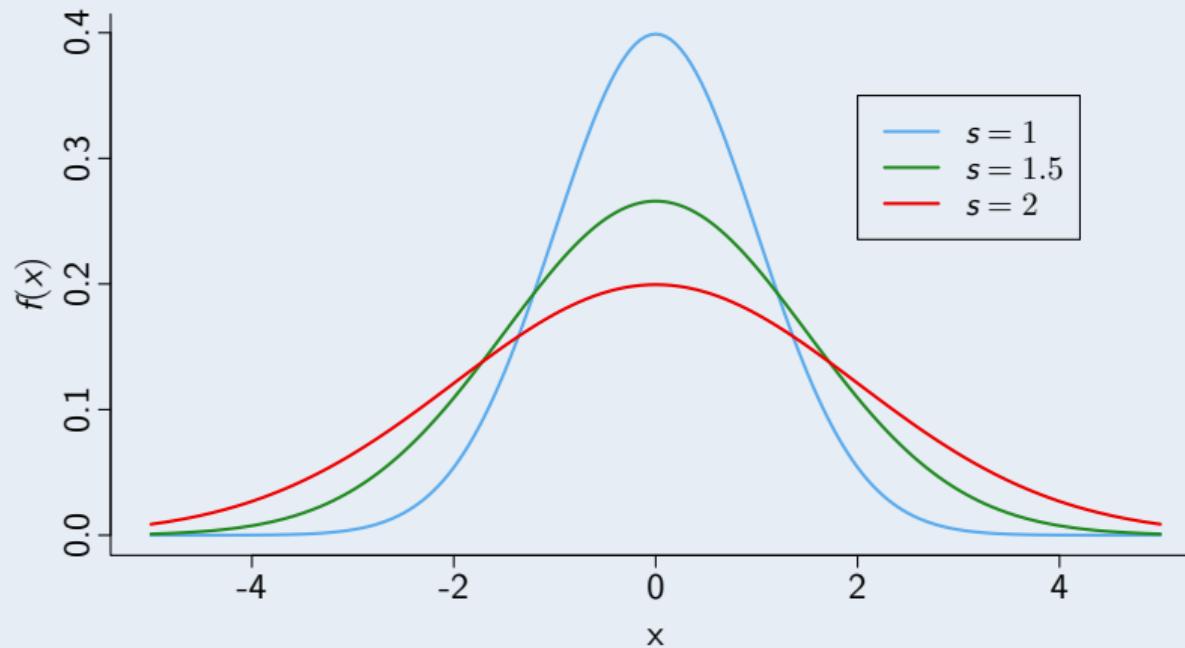
- Wir haben Lageparameter eindimensionaler Datensätze besprochen. Diese Parameter beschreiben einen Datensatz kompakt und machen verschiedene Datensätze vergleichbar.
- Das Verständnis des hier behandelten Stoffes ist für den weiteren Verlauf der Vorlesung elementar! Arbeiten Sie daher nicht nur die Zusammenhänge auf, sondern üben Sie auch die neuen Methoden anhand der Aufgaben.
- Nachbearbeitung: Kapitel 4.1 und 4.2 des Buches von Prof. Assenmacher.
- Die nächste Vorlesung behandelt Streuungsparameter, die uns z.B. darüber Auskunft geben, ob eine Aktie risikobehafteter ist als eine andere.
- Vorbereitung: Kapitel 4.3 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 **Streuungsparameter**
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

„Then there is the man who drowned while crossing a stream that was, on average, 6 inches deep.“
—W.I.E. Gates

- Lageparameter beschreiben Datensätze nur teilweise.
- Ebenso bedeutsam ist die Streuung der Daten. **Streuungsparameter** Θ_S bzw. Streuungsmaße liefern hierüber Information.
- Da Streuungsparameter immer eine Abstandsmessung voraussetzen, sind sie nur bei metrischen Merkmalen sinnvoll.
- Auch Streuungsmaße müssen bestimmte Axiome erfüllen.

Beispiel 4.1: Streuung verschiedener Normalverteilungen (Induktive Statistik).



Lageparameter vs Streuung.R

Das Konzept der Streuung

- Haben alle Beobachtungen dieselbe Ausprägung c , so streuen die Daten nicht (Einpunktverteilung). Θ_S soll gleich null sein: $x_1 = x_2 = \dots = x_n = c \Rightarrow \Theta_S = 0$.
- Sind mindestens zwei Beobachtungen verschieden, liegt Streuung vor: $\Theta_S \neq 0$. Da nur der Abstand der Beobachtungen zu einem Bezugspunkt, nicht aber ihre Richtung relevant ist, soll Θ_S positiv sein: $\Theta_S > 0$ für $x_i \neq x_j, i, j \in 1, \dots, n$.
- Eine Verschiebung des gesamten Datensatzes um $d \neq 0$ lässt die Abstände der Beobachtungen und damit auch ihre Streuung unverändert; Θ_S muss von der Lage der Daten unabhängig sein (Translationsinvarianz): $\Theta_S(x_1 + d, \dots, x_n + d) = \Theta_S(x_1, \dots, x_n)$.
- Datensätze mit gleicher empirischer Verteilungsfunktionen haben auch die gleiche Streuung. Θ_S soll daher auch homogen vom Grade null in den absoluten Häufigkeiten sein.

Das Konzept der Streuung

- Streuungsparameter unterscheiden sich in der zugrunde liegenden Abstandsmessung.
- Θ_S können die Abstände aller Beobachtungen untereinander zugrunde liegen. Alternativ lassen sich die Abweichungen aller Beobachtungen von einer Bezugsgröße bilden. Hierzu sind Lageparameter geeignet.
- Hierauf basierende Maßzahlen heißen **absolute Streuungsparameter**.
- Häufig nimmt jedoch mit dem Niveau der Daten auch ihre Streuung zu. Um diesen Größeneffekt zu kompensieren, benutzt man **relative Streuungsparameter**. Diese sind Quotienten eines absoluten Streuungsparameters und eines geeigneten Lageparameters.

Absolute Streuungsparameter

Spannweite R

Die einfachste Maßzahl ist die **Spannweite R** (auch *range* oder Variationsbreite).

Definition 4.2: Spannweite R .

Die Spannweite ist die Differenz zwischen größter und kleinster Beobachtung:

$$R = \max_j(x_j) - \min_j(x_j), \quad j = 1, \dots, n \text{ bzw. } R = x_{(n)} - x_{(1)}.$$

Beispiel 4.3: Nobelpreisträger.

Alter nach Fachdisziplinen

Absolute Streuungsparameter

Quartilsabstand Q

Die Spannweite ist ein recht grobes Streuungsmaß, das von **Ausreißern** abhängt. Der **Quartilsabstand** schaltet deren Einfluss aus.

Definition 4.4: Quartilsabstand Q .

Der Quartilsabstand ist die Differenz des dritten und ersten Quartils:

$$Q = x_{0,75} - x_{0,25}$$

Division des Quartilsabstands, auch Interquartilsbreite genannt, durch 2 ergibt den mittleren Quartilsabstand (**Semiquartilsabstand**).

Absolute Streuungsparameter

Quartilsabstand Q

Beispiel 4.5: Urliste aus Kapitel 2.

11, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 17, 17, 18.

Die Spannweite lässt sich berechnen als

$$R = x_{(20)} - x_{(1)} = 18 - 11 = 7.$$

Für den Quartilsabstand benötigen wir die Quantile

$$x_{0,25} = x_{(5)} = 13 \quad \text{da} \quad np = \text{int}(np) = 5$$

$$x_{0,75} = x_{(15)} = 16 \quad \text{da} \quad np = \text{int}(np) = 15.$$

Es folgt $Q = x_{0,75} - x_{0,25} = 3$

Absolute Streuungsparameter

Spannweite und Quartilsabstand - in 

```
x <- c(11, 12, 12, 13, 13, 13, 14, 14, 14, 15,
      15, 15, 15, 15, 16, 16, 16, 17, 17, 18)
max(x) - min(x) # Spannweite
## [1] 7

IQR(x) # "Interquartile range"
## [1] 3

diff(quantile(x, c(0.25, 0.75), type = 1))
## 75%
## 3
```

Absolute Streuungsparameter

Quartilsabstand Q

Beispiel 4.6: Klausurpunkte.

Die unterschiedliche Aussagekraft von R und Q zeigt folgendes Beispiel. Bei einer Klausur haben 20 Studierende folgende Anzahl an Punkten erreicht:

$$0, 0, 4, 6, 20, 20, 21, 21, 22, 23, 23, 25, 26, 27, 31, 31, 34, 42, 51, 60.$$

Die Spannweite für diese Daten beträgt 60.

Das für Q benötigte erste und dritte Quartil erhält man als

$$x_{0,25} = x_{(5)}, \quad \text{da} \quad np = 20 \cdot 0,25 = 5$$

und

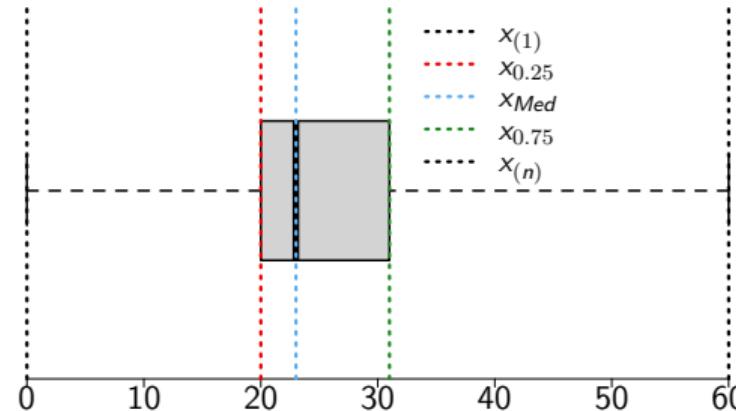
$$x_{0,75} = x_{(15)}, \quad \text{da} \quad np = 20 \cdot 0,75 = 15.$$

Daher liegen die erreichten Punkte der mittleren 50% der Ergebnisse um höchstens $Q = 31 - 20 = 11$ Punkte auseinander.

Absolute Streuungsparameter

Box-Plot

- Ein **Box-Plot** veranschaulicht den Quartilsabstand.
- Hierzu benötigt man neben den drei Quartilen $x_{0,25}$, $x_{0,5}$ und $x_{0,75}$ noch den kleinsten und größten Beobachtungswert $x_{(1)}$ und $x_{(n)}$.
- Zwischen $x_{(1)}$ und $x_{0,25}$ sowie zwischen $x_{0,75}$ und $x_{(n)}$ liegen jeweils mindestens 25% der Beobachtungen, zwischen $x_{0,25}$ und $x_{0,75}$ mindestens 50% aller Beobachtungen.



Box-Plot

- Wie das 5-Zahlen-Schema für die Klausurdaten verdeutlicht, sind die fünf Punkte nicht äquidistant; auch muss der Median nicht in der Mitte der Box liegen. Die Medianpunktzahl von 23 Punkten ist kleiner als die Mitte der Box (25,5).
- Liegen Ausreißer vor, verwendet man anstelle von $x_{(1)}$ bzw. $x_{(n)}$ z.B. das 0,1- und 0,9-Quantil als äußere Punkte des Schachteldiagramms.
- Mit Box-Plots können verschiedene Datensätze gut verglichen werden. [Earnings](#), [Icehockey](#)

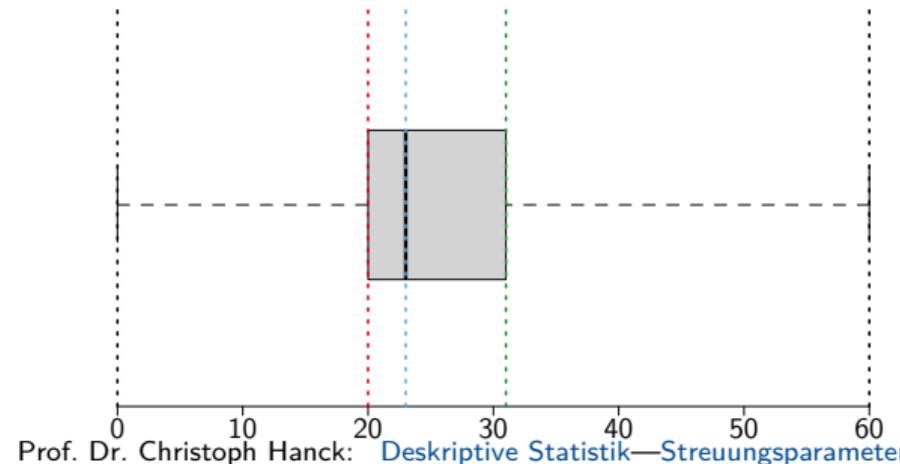
Absolute Streuungsparameter

Box-Plot - in R

- Beachte, dass boxplot die Quantile anders berechnet als wir. Im aktuellen Beispiel sind die Daten so gewählt, dass es keinen Unterschied macht, ob man `quantile(x, 0.5, type = 1)` oder `quantile(x, 0.5)` nutzt.

```
x <- c(0, 0, 4, 6, 20, 20, 21, 21, 22, 23,  
      23, 25, 26, 27, 31, 31, 34, 42, 51, 60)
```

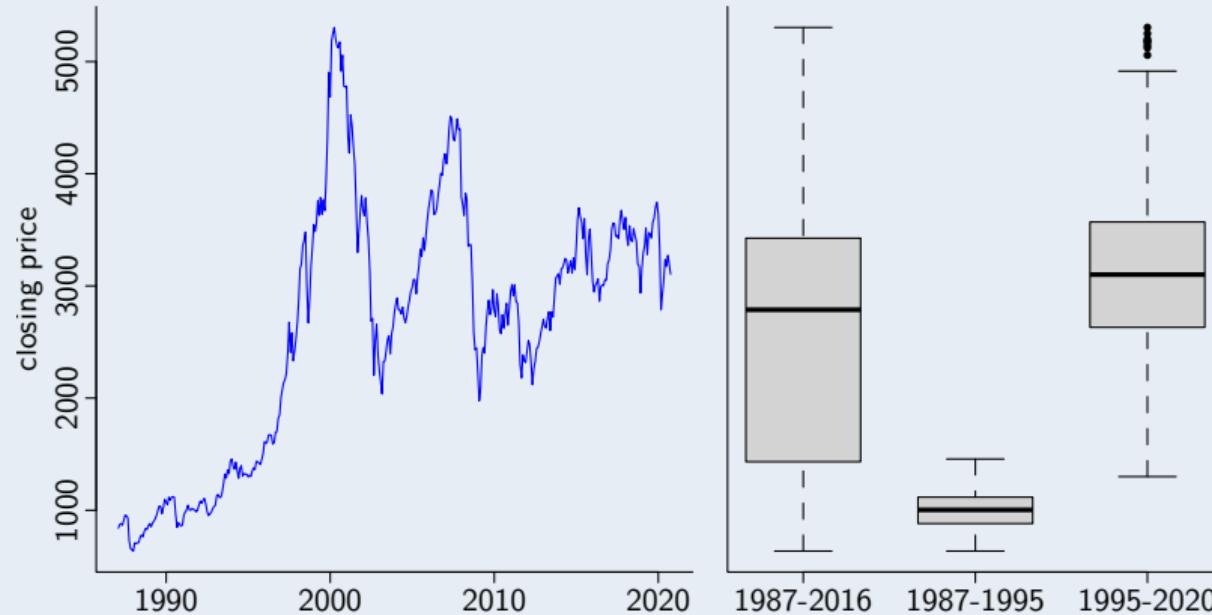
```
bxpl_data <- boxplot(x, horizontal = T, range = 0, cex.axis = 1.2, frame = F)  
abline(v = bxpl_data$stats, lty = 3, lwd = 2,  
       col = c("black", "red2", "steelblue2", "forestgreen"))
```



Absolute Streuungsparameter

Box-Plot

Beispiel 4.7: Euro Stoxx 50 (Aktienindex) 1987-2020.



Daten: Europäische Zentralbank

Absolute Streuungsparameter

Box-Plot

Beispiel 4.8: Selbststudium von BWL-Studierenden (in Stunden pro Tag).

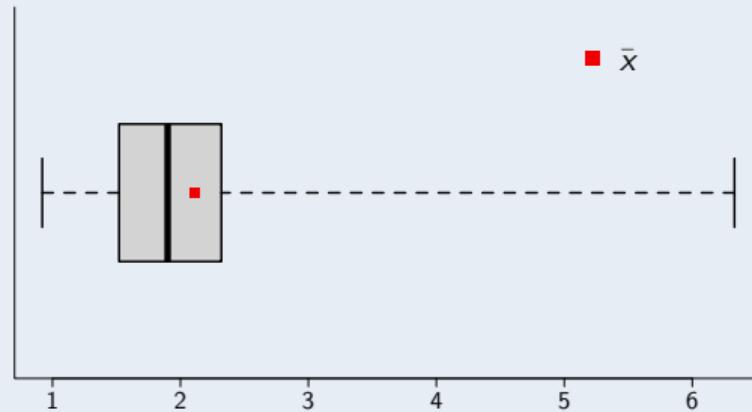
$$\bar{x} = 2.11$$

$$x_{\text{Med}} = 1.90$$

$$R = 6.33 - 0.92 = 5.41$$

$$Q = 2.32 - 1.52 = 0.8$$

$$s = 0.97$$



Quelle: anonym

Absolute Streuungsparameter

Varianz s^2

- Die Streuung lässt sich auch durch die Abweichungen der Beobachtungen von einem Bezugspunkt $a \in \mathbb{R}$ erfassen.
- Für ein aussagefähiges Θ_S sollte a ein Lageparameter sein.
- Da sowohl positive als auch negative Abweichungen von a zur Streuung beitragen, muss Θ_S so spezifiziert werden, dass sie sich nicht kompensieren und dadurch die Streuung zu gering erscheint.
- Dieser Kompensationseffekt wird vermieden, wenn man Abweichungen quadriert: $(x_j - a)^2$, $a \in [x_{(1)}, x_{(n)}]$.

Absolute Streuungsparameter

Varianz s^2

- Als Lageparameter wählt man für a wegen seiner Minimierungseigenschaft bei Summen quadrierter Abweichungen \bar{x} .
- Die durchschnittliche quadratische Abweichung mit $a = \bar{x}$ heißt **Varianz s^2** .

Definition 4.9: Varianz.

Für Einzelbeobachtungen bzw. häufigkeitsverteilte Daten ist s^2 definiert als

$$s^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 \quad \text{bzw.} \quad s^2 = \frac{1}{n} \sum_{i=1}^m (x_i - \bar{x})^2 n_i = \sum_{i=1}^m (x_i - \bar{x})^2 h_i.$$

Absolute Streuungsparameter

Varianz s^2

Definition 4.10: Allgemeiner und spezieller Verschiebungssatz.

Der **allgemeine Verschiebungssatz** für ein beliebiges $a \in \mathbb{R}$ lautet

$$s^2 = \frac{1}{n} \sum_{j=1}^n (x_j - a)^2 - (\bar{x} - a)^2.$$

Für $a = 0$ folgt der **spezielle Verschiebungssatz**, mit dem s^2 oft einfach berechnet werden kann:

$$s^2 = \frac{1}{n} \sum_{j=1}^n x_j^2 - \bar{x}^2.$$

Die Verschiebungssätze gelten für Häufigkeitsverteilte Daten analog. Für $a = 0$ ergibt sich etwa

$$s^2 = \frac{1}{n} \sum_{i=1}^m x_i^2 n_i - \bar{x}^2 = \sum_{i=1}^m x_i^2 h_i - \bar{x}^2.$$

Streuungsmaße 1

Kahoot!

Absolute Streuungsparameter

Varianz bei klassierten Daten

- Hat man nur klassierte Beobachtungen, kann die Varianz nur über die arithmetischen Klassenmittel oder Klassenmitten berechnet werden.
- Da die so ermittelte Varianz meist von der Varianz der Urliste abweicht, wird sie mit s_K^2 bei bekannten und mit \hat{s}_K^2 bei unbekannten arithmetischen Klassenmitteln bezeichnet.

Definition 4.11: Varianz von klassierten Daten.

Die Varianzformeln für klassierte Daten lauten:

$$s_K^2 = \frac{1}{n} \sum_{k=1}^K (\bar{x}_k - \bar{x})^2 n_k = \sum_{k=1}^K (\bar{x}_k - \bar{x})^2 h_k$$

bzw.

$$\hat{s}_K^2 = \frac{1}{n} \sum_{k=1}^K (m_k - \bar{x})^2 n_k = \sum_{k=1}^K (m_k - \bar{x})^2 h_k.$$

Absolute Streuungsparameter

Varianz bei klassierten Daten

Definition 4.12: Varianz von klassierten Daten.

Mit dem speziellen Verschiebungssatz gehen die Gleichungen über in:

$$s_K^2 = \frac{1}{n} \sum_{k=1}^K \bar{x}_k^2 n_k - \bar{x}^2 = \sum_{k=1}^K \bar{x}_k^2 h_k - \bar{x}^2$$

bzw.

$$\hat{s}_K^2 = \frac{1}{n} \sum_{k=1}^K m_k^2 n_k - \hat{\bar{x}}^2 = \sum_{k=1}^K m_k^2 h_k - \hat{\bar{x}}^2.$$

- Stimmen in jeder Klasse die Beobachtungen überein, misst s_K^2 die Varianz der Originalreihe; streuen die Daten in mindestens einer Klasse, gilt immer $s_K^2 < s^2$, da die Streuung innerhalb der Klassen unberücksichtigt bleibt.
- Das Buch (S. 99) leitet bei Vorliegen von Einzelbeobachtungen eine Formel für die Differenz von s^2 und s_K^2 her.

Absolute Streuungsparameter

Varianz bei klassierten Daten

- Bei Verwendung der Klassenmitten m_k bleibt zwar die Streuung innerhalb der Klassen ebenfalls unberücksichtigt, jedoch wird mit \hat{s}_K^2 die Varianz der Urliste dann meist zu groß ausgewiesen, wenn die Daten in den Klassen sehr asymmetrisch zur Klassenmitte verteilt sind.
- Bei gleichen Klassenbreiten $\Delta_k = \Delta$ für $k = 1, \dots, K$ lässt sich diese „Überschätzung“ der Varianz der Urliste mit der **Sheppard-Korrektur** kompensieren: Verwende anstelle von \hat{s}_K^2 die korrigierte Varianz

$$(\hat{s}_K^*)^2 = \hat{s}_K^2 - \Delta^2/12.$$

Absolute Streuungsparameter

Varianz s^2

- Die Varianzen zweier Datensätze stehen in einer festen Beziehung, wenn die y_j eine Lineartransformation der x_j sind: $y_j = \alpha + \beta x_j$, $j = 1, \dots, n$ (vgl. Transformationseigenschaft des arithm. Mittels).
- Für die Varianz der Beobachtungen y_j , s_y^2 gilt (Beweis siehe Übung)

$$s_y^2 = \beta^2 s_x^2.$$

- s_y^2 nimmt mit dem Quadrat des Skalenfaktors β zu; der Verschiebungsparameter α hingegen hat keinen Einfluss: s^2 ist also translationsinvariant.

Absolute Streuungsparameter

Varianz s^2

```
# Umrechnung von Temperaturdaten (siehe Bsp. 3.6)
```

```
x <- c(23, 25, 24, 19, 22, 23, 24)
```

```
y <- 32 + 1.8 * x
```

```
n <- length(x)
```

```
# Varianz von y
```

```
(n - 1) / n * var(y)
```

```
## [1] 10.57959
```

```
# über die Transformationseigenschaft
```

```
1.8^2 * ((n-1)/n * var(x))
```

```
## [1] 10.57959
```

Den Grund dafür, dass `var(x)` nicht direkt den passenden Wert ausgibt, diskutieren wir in Induktive Statistik.

- Die Varianz hat wegen des Quadrierens eine andere Dimension als das betrachtete Merkmal.
- Diesen Nachteil beseitigt die positive Wurzel der Varianz. Sie heißt **Standardabweichung** und wird mit s bezeichnet: $s = \sqrt{s^2}$.
- Sie besitzt dieselbe Dimension wie das betrachtete Merkmal.

Streuungsmaße 2

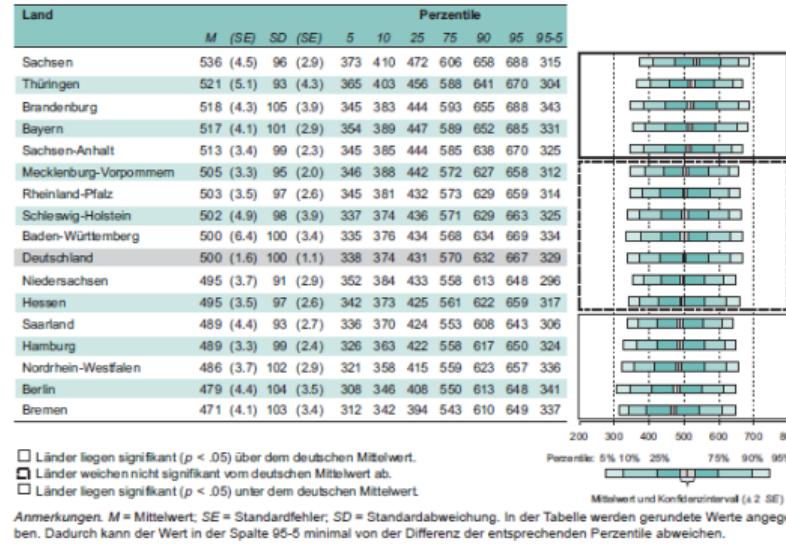
Kahoot!

Absolute Streuungsparameter

Percentilabstände, Box-Plot und Standardabweichung

Beispiel 4.13: Schulvergleich der Bundesländer Mathematik.

Abbildung 5.2: Mittelwerte, Streuungen, Percentile und Percentilbänder der von Schülerinnen und Schülern der 9. Jahrgangsstufe erreichten Kompetenzstände im Fach Mathematik (Globalskala)



Quelle: Roppelt et al. (2013) „IQB-Ländervergleich 2012“

Absolute Streuungsparameter

Varianz und Standardabweichung - in

```
# Beispieldaten aus Kapitel 2
x <- c(11, 12, 12, 13, 13, 13, 14, 14, 14, 15,
      15, 15, 15, 15, 16, 16, 16, 17, 17, 18)
(n <- length(x))
## [1] 20

(var.x <- (n - 1) / n * var(x))
## [1] 3.2475

(sd.x <- sqrt(var.x))
## [1] 1.802082

# alternativ (vgl. empirische Momente in Kapitel 5)
library(moments)
moment(x, order = 2, central = T)
## [1] 3.2475
```

Relative Streuungsparameter

- Relative Streuungsparameter sind Quotienten eines absoluten Streuungsparameters Θ_S zu einem Lageparameter $\Theta_L > 0$, wobei beide Parameter dieselbe Dimension besitzen müssen. Ein relatives Streuungsmaß ist daher dimensionslos.
- Relative Streuungsparameter eignen sich zum Vergleich der Streuung von:
 - ▶ Merkmalen mit verschiedenen Dimensionen, wie Körpergröße und Gewicht,
 - ▶ Merkmalen mit verschiedenen Messeinheiten, z.B. der in EUR oder in Mio. EUR gemessene Umsatz eines Unternehmens oder die Einkommensverteilung von Volkswirtschaften mit unterschiedlichen Währungen,
 - ▶ Daten, deren Messniveau und damit auch ihre Lageparameter stark differieren, z.B. Inlandsprodukt- und Zinssatzdaten.

Relative Streuungsparameter

Variationskoeffizient v

Der **Variationskoeffizient** beruht auf der Standardabweichung und dem arithmetischen Mittel.

Definition 4.14: Variationskoeffizient.

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

- Diese Vorlesung themisierte die Streuungsparameter eindimensionaler Datensätze. Der Anwendungsbereich dieser Parameter ist vielfältig und für das weitere Verständnis der Vorlesung und darüber hinaus auch fachübergreifend elementar. Es empfiehlt sich daher zur Nachbereitung die neuen Methoden mit Hilfe der Aufgaben einzuüben.
- Nachbearbeitung: Kapitel 4.3 des Buches von Prof. Assenmacher.
- In der nächsten Vorlesung werden die Parameter der Schiefe und Kurtosis behandelt, die uns weitere Informationen über die Charakteristik eines Datensatzes liefern.
- Vorbereitung: Kapitel 4.4 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

Kurtosis

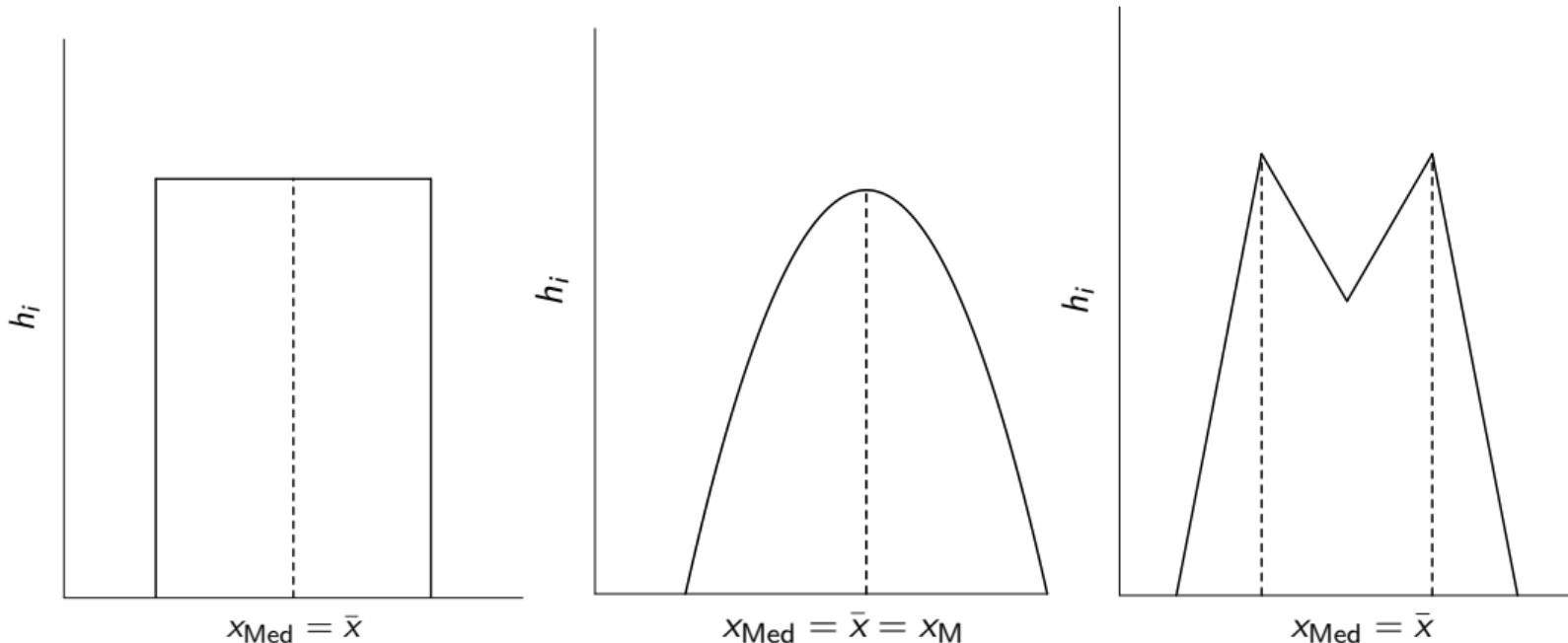
Das Konzept der Kurtosis

- Wenn alle Ausprägungen spiegelbildlich zum Median sind und die Häufigkeiten gleich weit vom Median entfernt liegender Ausprägungen übereinstimmen, heißen die Ausprägungen (axial-)symmetrisch zum Median.
- Bei **symmetrischen Häufigkeitsverteilungen** sind x_{Med} und \bar{x} gleich. Hat eine symmetrische Verteilung einen eindeutigen Modus, so ist auch er gleich x_{Med} und \bar{x} .

Kurtosis

Das Konzept der Kurtosis

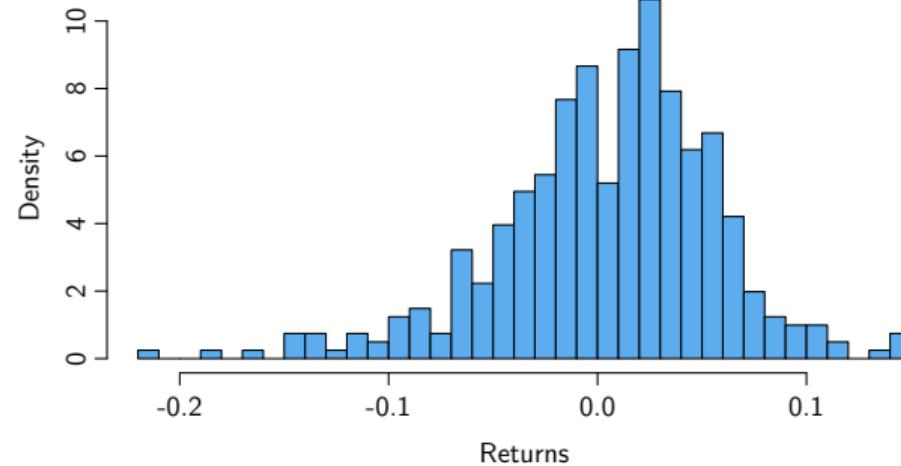
Hier sind drei symmetrische Häufigkeitsverteilungen wiedergegeben. Nur die zweite besitzt einen eindeutigen Modus.



- Symmetrische Verteilungen lassen sich durch einen Lageparameter und ein Streuungsmaß oft gut beschreiben.
- Symmetrische Verteilungen mit gleichen Lage- und Streuungsparametern müssen jedoch - auch wenn sie **unimodal** sind - nicht dieselbe Form besitzen.

Kurtosis

Das Konzept der Kurtosis

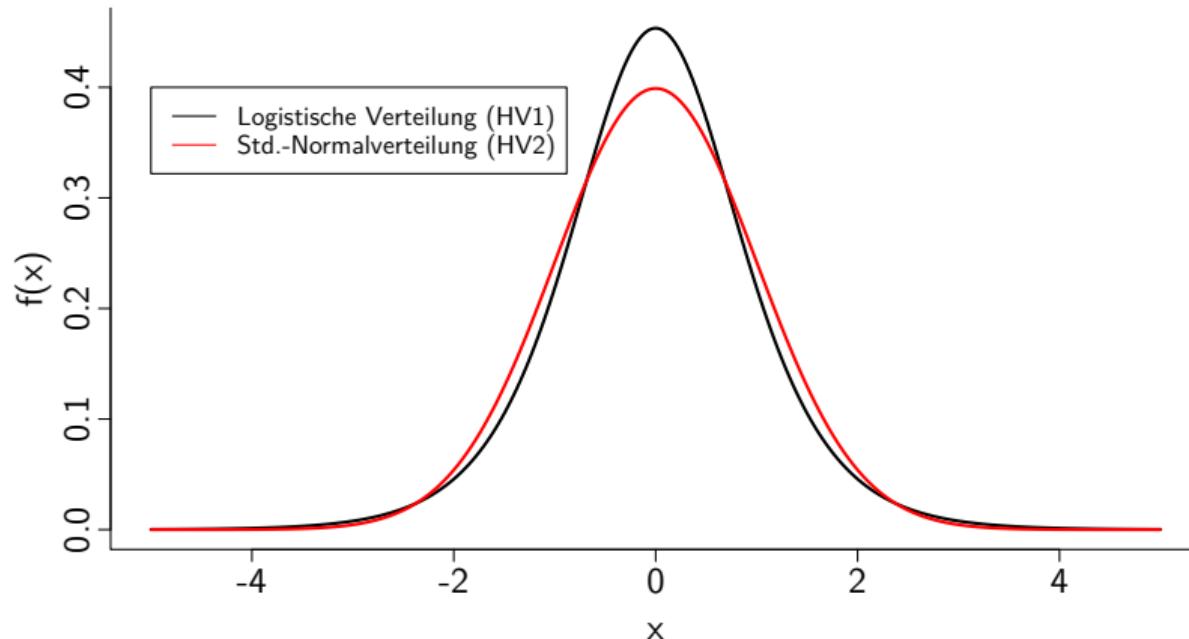


Renditen des Euro Stoxx 50 (01/1987 - 10/2020, Daten: Yahoo Finance)

Kurtosis

Das Konzept der Kurtosis

Hier sind zwei unimodale Häufigkeitsverteilungen mit gleichem Lagemaß ($= 0$) und gleicher Varianz ($= 1$) wiedergegeben.



Kurtosis

Das Konzept der Kurtosis

- Parameter für die **Kurtosis** einer Verteilung basieren zumeist auf **empirischen Momenten** und messen die Häufigkeit „extremer“ Ausprägungen ⇒ nur bei metrisch skalierten Merkmalen anwendbar.

Kurtosis

Empirische Momente

Definition 5.1: Empirische Momente.

Empirische Momente sind als arithmetische Mittel bestimmter Funktionen $f(X)$ einer statistischen Variablen X definiert, wobei $f(X)$ festgelegt ist durch:

$$f(X) = \left(\frac{X - a}{b} \right)^\alpha, \quad \text{mit } a, b \in \mathbb{R}, \quad b > 0 \quad \text{und} \quad \alpha \in \mathbb{N} \cup \{0\}.$$

- Momente hängen von den Parametern a, b und α der Funktion $f(X)$ ab; bezeichne sie daher mit $m(a, b)_\alpha$.
- Der Parameter α gibt die Ordnung des Moments an.

Kurtosis

Empirische Momente

Definition 5.2: Empirische Momente.

Je nachdem, ob die Daten als Einzelbeobachtungen oder häufigkeitsverteilt vorliegen, sind Momente definiert als:

$$m(a, b)_\alpha = \frac{1}{n} \sum_{j=1}^n \left(\frac{x_j - a}{b} \right)^\alpha \quad \text{oder}$$

$$m(a, b)_\alpha = \sum_{i=1}^m \left(\frac{x_i - a}{b} \right)^\alpha h_i.$$

Kurtosis

Empirische Momente

- Für $\alpha = 0$ gilt für alle a und b : $m(a, b)_0 = 1$.
- Ist $\alpha \neq 0$, lassen sich drei wichtige Klassen von Momenten gewinnen. Ist $a = 0$ und $b = 1$, erhält man die Klasse der Anfangs- bzw. Nullmomente der Ordnung α , geschrieben als $m(0)_\alpha$.
- Für Einzelbeobachtungen folgt

$$m(0)_\alpha = \frac{1}{n} \sum_{j=1}^n x_j^\alpha.$$

- Für $\alpha = 1$ ergibt sich das Anfangsmoment erster Ordnung (kurz: erstes Anfangsmoment) \bar{x} :

$$m(0)_1 = \frac{1}{n} \sum_{j=1}^n x_j = \bar{x}.$$

Kurtosis

Empirische Momente

- Die wichtigste Klasse der Zentralmomente der Ordnung α resultiert aus $a = \bar{x}$ und $b = 1$, geschrieben m_α . Für Einzelbeobachtungen folgt

$$m_\alpha = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^\alpha.$$

- Für $\alpha = 2$ entspricht das Zentralmoment der Varianz.
- Die dritte Klasse resultiert aus $a = \bar{x}$ und $b = s_x$. Die Momente dieser Klasse, z_α , heißen Standardmomente der Ordnung α :

$$z_\alpha = \frac{1}{n} \sum_{j=1}^n \left(\frac{x_j - \bar{x}}{s_x} \right)^\alpha.$$

- Das erste Standardmoment ($\alpha = 1$) ist wegen der Schwerpunkteigenschaft von \bar{x} null: $z_1 = 0$. Ferner ist $z_2 = 1$; dies wurde bereits mit der Varianz linear transformierter Merkmale gezeigt.

Kurtosis

Empirische Momente

- Alle Zentralmomente lassen sich durch Anfangsmomente darstellen. Es gilt:

$$m_\alpha = \sum_{r=0}^{\alpha} \binom{\alpha}{r} m(0)_{\alpha-r} (-\bar{x})^r.$$

- Das zweite Zentralmoment ($\alpha = 2$) ist

$$\begin{aligned} m_2 &= \underbrace{\binom{2}{0} m(0)_2}_{=1} \underbrace{(-\bar{x})^0}_{=1} + \underbrace{\binom{2}{1} m(0)_1}_{=2} \underbrace{(-\bar{x})^1}_{=\bar{x}} \underbrace{(-\bar{x})^1}_{=-\bar{x}} + \underbrace{\binom{2}{2} m(0)_0}_{=1} \underbrace{(-\bar{x})^2}_{=1} \underbrace{(-\bar{x})^2}_{=\bar{x}^2} \\ &= m(0)_2 - 2\bar{x}^2 + \bar{x}^2 = m(0)_2 - \bar{x}^2 \\ &= \frac{1}{n} \sum_{j=1}^n x_j^2 - \bar{x}^2. \end{aligned}$$

- Die letzte Umformung ist der spezielle Verschiebungssatz der Varianz.

Kurtosis

Kurtosisparameter

- Maßzahlen für die Kurtosis einer Verteilung basieren auf den Abweichungen der Beobachtungen von einem Lageparameter. Dabei dürfen sich negative und positive Abweichungen nicht kompensieren.
- Zudem muss der Parameter mit dem Ausmaß der Kurtosis steigen, etwa indem große Abweichungen vom Lageparameter mit großem Gewicht in den Parameter eingehen.
- **Zentralmomente** gerader Ordnung erfüllen diese Erfordernisse: Der gerade Exponent verhindert die Kompensation positiver und negativer Abweichungen und bewirkt eine Selbstgewichtung der Abweichungen.

Kurtosis

Kurtosisparameter

- Das vierte Zentralmoment ist ein einfacher **absoluter Kurtosisparameter**:

$$\theta_K = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^4.$$

- Für den Vergleich der Kurtosis mehrerer Verteilungen mit unterschiedlichen Varianzen ist θ_K nicht sinnvoll, da θ_K die Kurtosis von Verteilungen mit großer Varianz überzeichnet.
- Vermeide dies durch Division von θ_K mit der quadrierten Varianz. So erhält man den (dimensionslosen) **relativen Kurtosisparameter**:

$$\theta_K^r = \frac{\theta_K}{s^4}.$$

- Umstellungen ergeben, dass θ_K^r gleich dem vierten Standardmoment ist: $\theta_K^r = z_4$.

Kurtosis

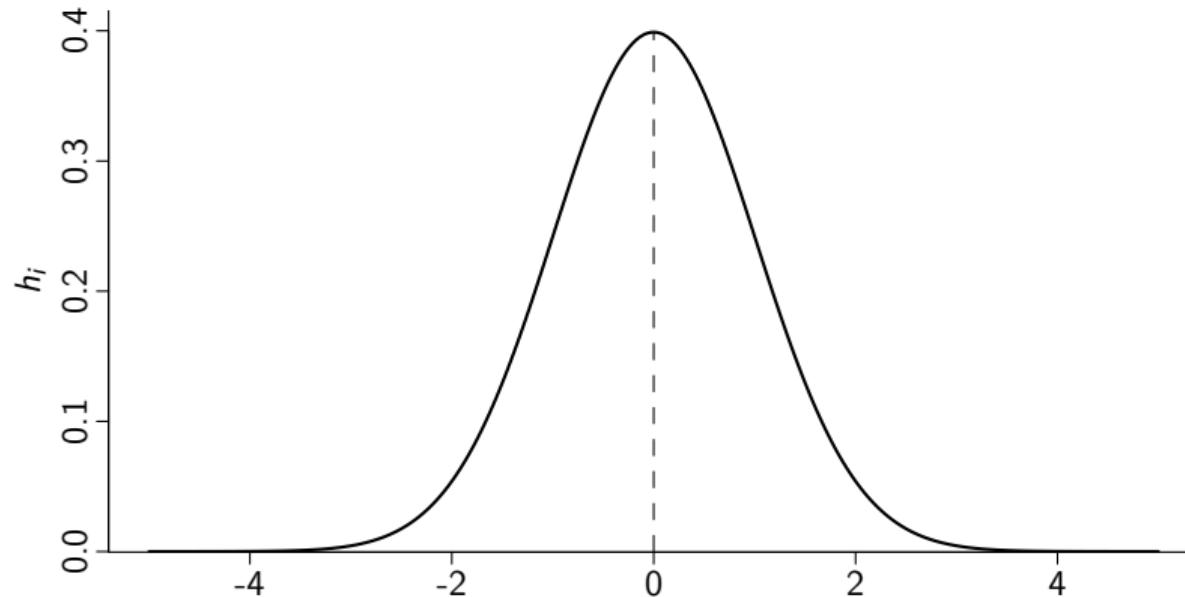
Kurtosisparameter

- Obwohl die Kurtosisparameter nur bei unimodalen und symmetrischen Häufigkeitsverteilungen verwendet werden sollten, werden sie auch bei asymmetrischen, aber unimodalen Verteilungen eingesetzt. Hier verlieren sie jedoch umso mehr an Aussagekraft, je stärker der Modus vom Lageparameter abweicht.
- Die Einschätzung der Kurtosis anhand von Parametern ist schwierig. Die Kurtosis einer konkreten Verteilung wird daher oft mit der Kurtosis der **Normalverteilung**, auch Gauß'sche Glockenkurve genannt, verglichen.

Kurtosis

Kurtosisparameter

Diese Verteilung (Normalverteilung) ist symmetrisch zu ihrem Lagemaß:



Kurtosis

Kurtosisparameter

- Da das vierte Standardmoment z_4 für jede Normalverteilung gleich drei ist, zeigt $\theta_K^N = \theta_K^r - 3$, wie die Kurtosis einer Verteilung von der Kurtosis der Normalverteilung abweicht. Diese Differenz θ_K^N heißt **zentrierter Kurtosisparameter**.
- Für $\theta_K^N = 0$ ist die Verteilung genauso wie Normalverteilung gewölbt (**mesokurtisch**). Für $\theta_K^N > 0$ liegt stärkere, bei $\theta_K^N < 0$ geringere Kurtosis als bei der Normalverteilung vor (**leptokurtisch** bzw. **platykurtisch**).
- Wegen des Bezugs auf die Normalverteilung ist auch der zentrierte Kurtosisparameter eigentlich nur bei unimodalen, symmetrischen Häufigkeitsverteilungen aussagekräftig.

Kurtosis

Kurtosisparameter

Beispiel 5.3:

Für die Beispieldaten aus Kapitel 2

11, 13, 15, 16, 12, 18, 14, 15, 17, 14, 12, 16, 13, 15, 17, 16, 15, 14, 13, 15

gilt $x_M = 15$ und $\bar{x} = 14,55$. Da die Verteilung dahingehend fast symmetrisch ist, können Kurtosisparameter berechnet werden.

Der absolute Kurtosisparameter ist $\theta_K = m_4 \approx 24.4087$. Das zweite Moment (Varianz) ist $m_2 = 3,2475$; daher ist $\theta'_K = 24,4087/(3,2475)^2 = 2,3144$. Die Ergebnisse sind nicht sehr intuitiv.

Der zentrierte Kurtosisparameter von $2,3144 - 3 = -0.6856$ zeigt an, dass die Kurtosis geringer (platykurtisch) als bei einer Normalverteilung ist.

Kurtosis

Kurtosisparameter - in R

```
library(moments)
x <- c(11, 12, 12, 13, 13, 13, 14, 14, 14, 15,
      15, 15, 15, 15, 16, 16, 16, 17, 17, 18)

kurtosis(x)
## [1] 2.314445

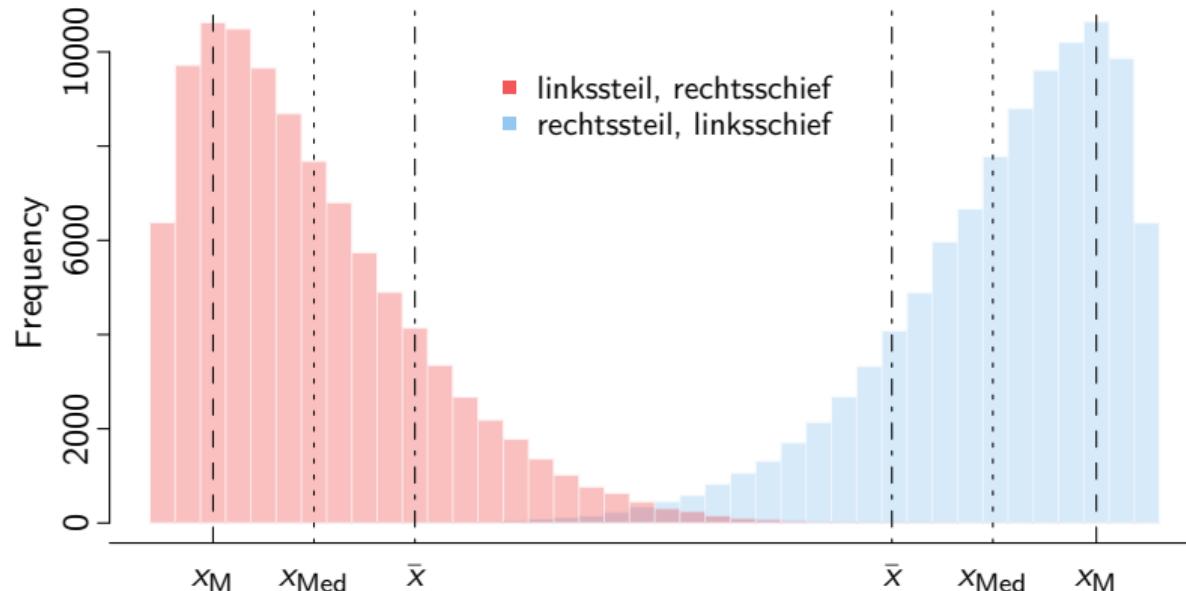
(thetaK <- mean((x - mean(x))^4))
## [1] 24.40873

(var.x <- mean((x - mean(x))^2))
## [1] 3.2475

(k <- thetaK/var.x^2)
## [1] 2.314445
```

Das Konzept der Schiefe

- Unimodale asymmetrische Häufigkeitsverteilungen heißen schief.
- Man unterscheidet **rechts- und linksschiefe Verteilungen**:



Schiefe

Das Konzept der Schiefe

- Eine rechtsschiefe Verteilung ist auf ihrer linken Seite steil („**linkssteil**“). Bei einer linksschiefen Verteilung ist die rechte Seite steil („**rechtssteil**“).
- Bei schiefen Verteilungen stimmen x_{Med} , x_M und \bar{x} nicht überein.
- Ihre Ordnung kann Informationen über die Art der Schiefe liefern. x_{Med} liegt dann zwischen x_M und \bar{x} . \bar{x} liegt wegen seiner Schwerpunkteigenschaft im „schiefen“ Teil der Verteilung (**Fechtersche Lageregel**).
- Bei $x_M < x_{\text{Med}} < \bar{x}$ bezeichnen wir eine Verteilung also als „rechtsschief“ („**linkssteil**“); als „linksschief“ („**rechtssteil**“) bei $\bar{x} < x_{\text{Med}} < x_M$.
- Schiefe hängt auch mit Abweichungen $x_j - \bar{x}$ zusammen. Bei „Rechtsschiefe“ („Linksschiefe“) sind wegen $\bar{x} > x_{\text{Med}}$ ($\bar{x} < x_{\text{Med}}$) mehr als die Hälfte der $(x_j - \bar{x})$ negativ (positiv).

Schiefe

Das Konzept der Schiefe

Beispiel 5.4: Selbststudium von BWL-Studierenden (in Stunden pro Tag).

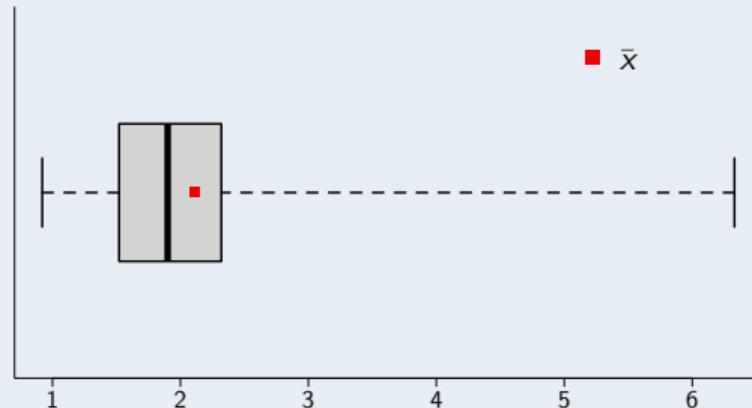
$$\bar{x} = 2.11$$

$$x_{\text{Med}} = 1.90$$

$$R = 6.33 - 0.92 = 5.41$$

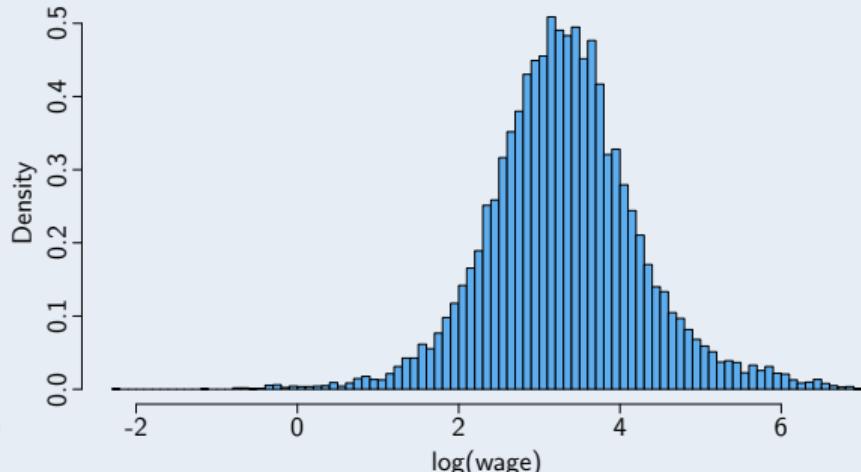
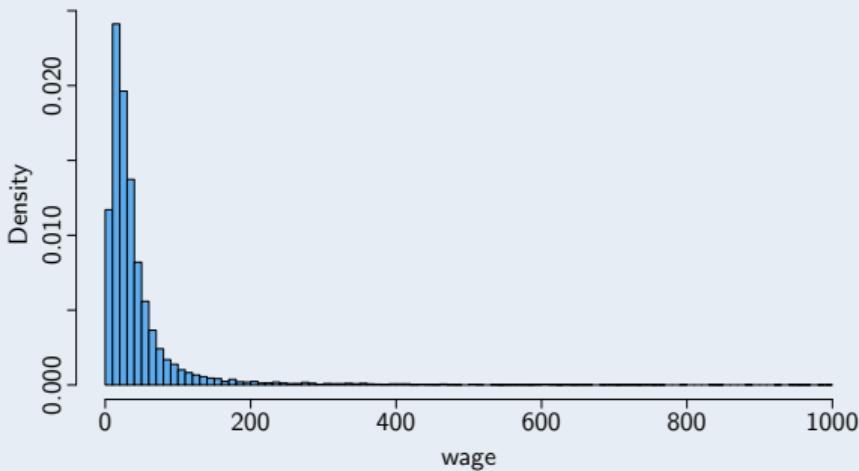
$$Q = 2.34 - 1.505 = 0.835$$

$$s = 0.97$$



Das Konzept der Schiefe

Beispiel 5.5: Deutsche Stundenlöhne (Euro/Stunde).



Deutsche Stundenlöhne (Euro/Stunde), Quelle: SOEP 2013

- Schiefeparameter nutzen den Zusammenhang zwischen Schiefe und Abweichungen. Bei linkssteilen Verteilungen sind zwar mehr als die Hälfte der Abweichungen negativ, sie sind aber vom Betrag her kleiner als die positiven. Der **Schiefeparameter** sollte dann positiv sein.
- Bei rechtssteilen Verteilungen ist es umgekehrt. Dort sollte der Parameter negativ sein.

Schiefe

Schiefeparameter

Das dritte Zentralmoment erfüllt diese Anforderungen (**absoluter Schiefeparameter**) θ_{Sch} :

Definition 5.6: Absoluter Schiefeparameter.

$$\theta_{\text{Sch}} = m_3 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^3$$

Dieser Parameter gewichtet jede Abweichung $(x_j - \bar{x})$ mit $(x_j - \bar{x})^2$. Große Abweichungen haben daher großes Gewicht und der Parameter nimmt das gewünschte Vorzeichen an. Aus Symmetrie folgt, dass $\theta_{\text{Sch}} = 0$.

Schiefe

Schiefeparameter

- Auch hier lässt sich der aus einer großen Streuung resultierende Effekt mit Division durch s^3 kompensieren. Dies liefert einen **relativen Schiefeparameter** θ_{Sch}^r , das dritte Standardmoment:

$$\theta_{\text{Sch}}^r = z_3 = \frac{\theta_{\text{Sch}}}{s^3}.$$

- Wegen seiner Dimensionslosigkeit eignet er sich zum Vergleich verschiedener Verteilungen.

```
# Fechnersche Lageregel und z_3 können sich widersprechen!
```

```
library(moments)
```

```
x1 <- 2.7 # probieren Sie auch mal x1 <- 3 und x1 <- 3.3 aus!
```

```
x <- c(x1, 15, 15, 15, 30, 30)
```

```
mean(x)
```

```
## [1] 17.95
```

```
median(x)
```

```
## [1] 15
```

```
skewness(x)
```

```
## [1] -0.02364842
```

- Das zugrunde liegende Problem ist, dass die Definition von „schief“ nicht restlos klar ist.

Schiefe und Kurtosis

Kahoot!

Schiefe

Quantil-Quantil-Diagramm

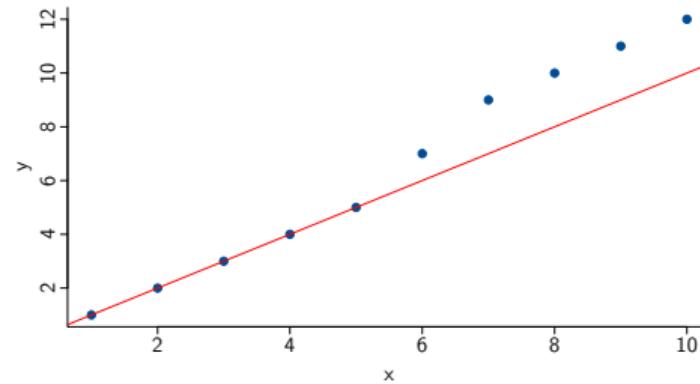
- Stimmen die relativen Häufigkeitsverteilungen zweier Datensätze überein, gilt das auch für ihre p -Quantile $x_p = y_p$.
- In einem Quantil-Quantil-Diagramm (kurz **Q-Q-Plot**) werden die Quantilspaare (x_p, y_p) für verschiedene p als Punkte in ein Koordinatensystem eingetragen.
- Liegen alle Punkte auf der 45° -Geraden, so sind die Verteilungen gleich; zunehmende Abweichung der Punkte von der Geraden zeigt Ungleichheit an.
- Liegen die Punkte annähernd auf einer Parallelen zur 45° -Geraden, unterscheiden sich die Verteilungen nur durch ihren Lageparameter.
- Verläuft die Parallele oberhalb der 45° -Geraden, ist der Lageparameter des Ordinaten-Datensatzes größer als der des Abszissen-Datensatzes.
- Entsteht das Q-Q-Diagramm auf Basis von Dezilen, sind die Zahlenpaare (x_j, y_j) die Quantilpunkte.

- Je weniger linear der Eindruck, den die Punkte vermitteln, desto unterschiedlicher sind beide Verteilungen.
- Die Ungleichheit, die aus Lage und Streuung der Daten resultiert, kann eliminiert werden, indem vor Berechnung der p -Quantile beide Datensätze standardisiert werden.
- Für die Wahl der p -Quantile gibt es keine verbindlichen Regeln. Häufig verwendet man Dezile.

Schiefe

Quantil-Quantil-Diagramm -

```
x <- 1:10 # nur 10 Beobachtungen, (sortierte) Datenpunkte also gleich Dezile
y <- c(7, 2, 11, 4, 12, 1, 10, 9, 3, 5)
# Einfacher Q-Q-Plot
qqplot(x = x, y = y, pch = 19, col = due.col$blue)
abline(a = 0, b = 1, col = "red")
```

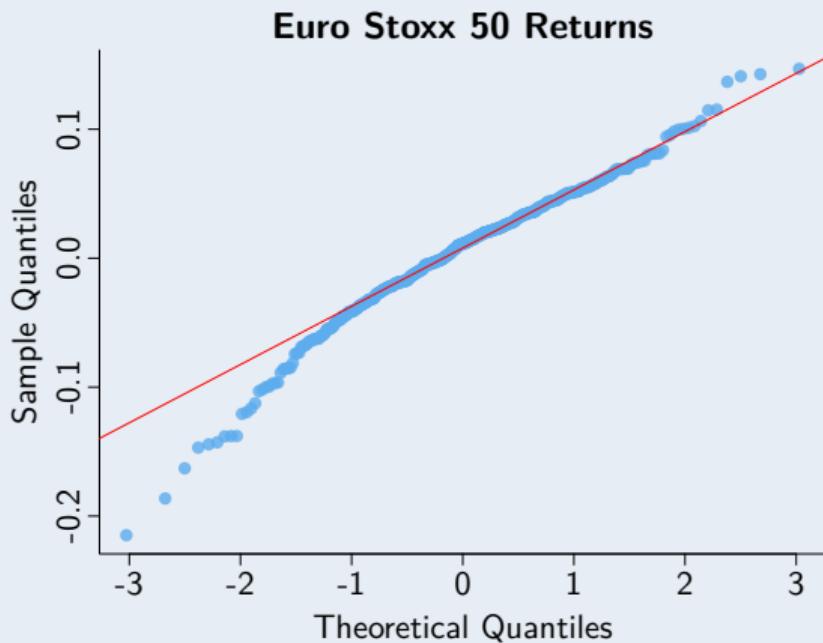
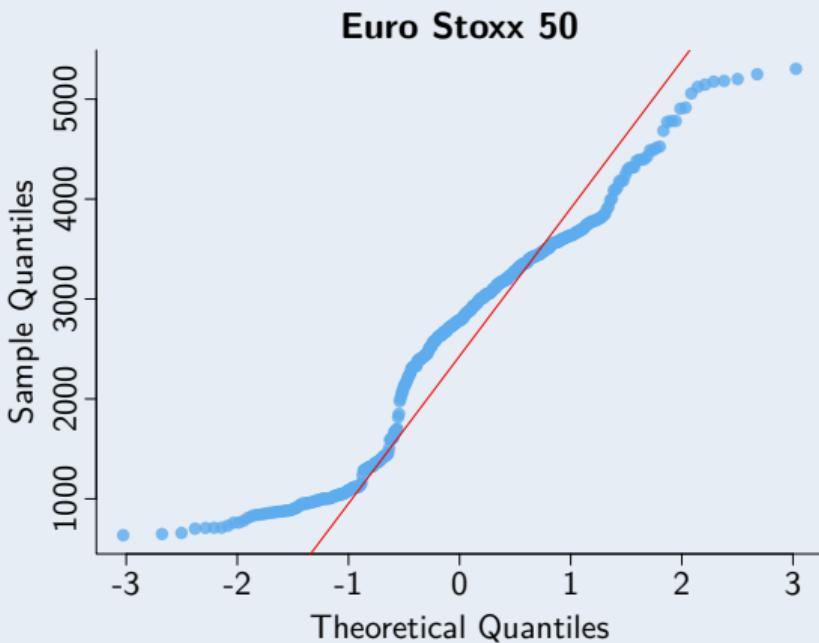


Im Q-Q-Plot ist zu erkennen, dass die unteren Dezile der Verteilungen übereinstimmen. Dies sieht man daran, dass die Daten mit der 45° -Geraden zusammenfallen. Für die Dezile oberhalb des Medians hingegen unterscheiden sich die Verteilungen. Da $\bar{y} > \bar{x}$, liegen die Punkte oberhalb der 45° -Geraden.

Schiefe

Quantil-Quantil-Diagramm

Beispiel 5.7: Euro Stoxx 50 (Aktienindex) 01/1987-05/2018.



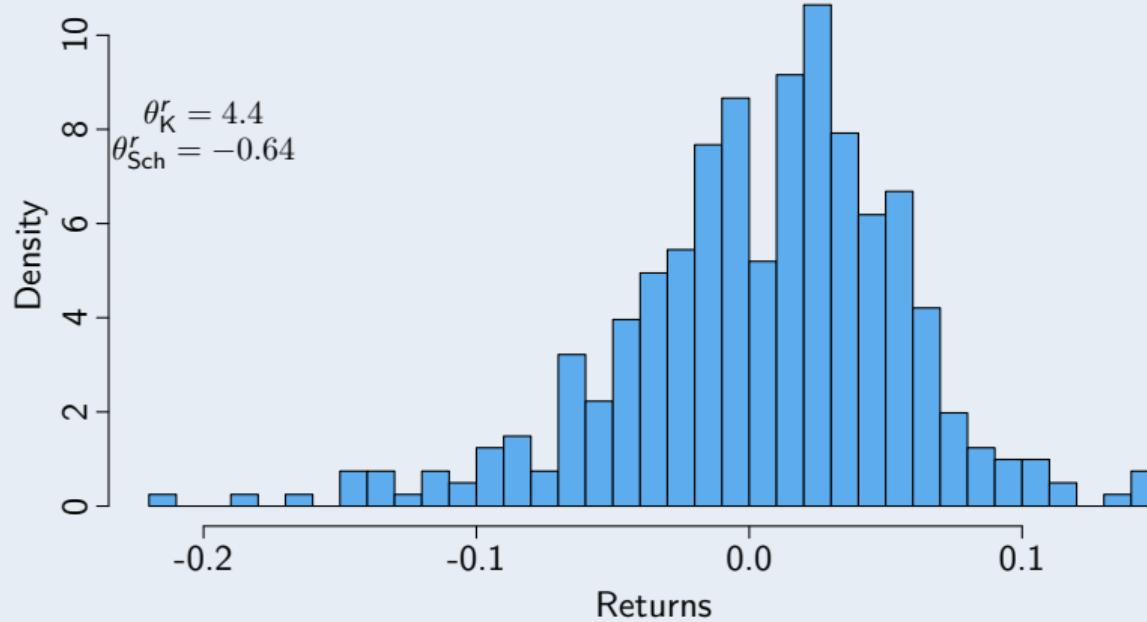
Daten: Yahoo Finance

Schiefe

Quantil-Quantil-Diagramm

Beispiel 5.7: Fortsetzung.

Histogram of EuroStoxx 50 Returns



Daten: Quandl

- Diese Vorlesung thematisierte Schiefe und Kurtosis. Hiermit können nicht-symmetrische Verteilungen beschrieben werden.
- Diese Vorlesung wiederholen Sie z.B. anhand von Kapitel 4.4 des Buches von Prof. Assenmacher.
- Die nächste Vorlesung behandelt die Konzentrationsmessung.
- Vorbereitung: Kapitel 4.5 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung**
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

Beispiel 6.1: Vermögensverteilung in Deutschland und den USA.

Zum Einstieg ins Thema:

- Ungleiche Vermögensverteilung in Deutschland? Schätzen Sie mal!
- Hier eine Studie zur Wohlstandsverteilung in Deutschland aus 2020. Überrascht?
- Vermögensverteilung 2

- Bei Merkmalen mit nicht negativen Ausprägungen können die Daten über die **Verteilung der Merkmalssumme auf die Merkmalsträger** charakterisiert werden.
- Hiermit spürt man **Konzentration** in der Verteilung auf. In der Ökonomie tritt sie z.B. als Vermögens-, Einkommens-, Umsatz-, Beschäftigungs- oder Marktmachtkonzentration auf.
- Konzentration umfasst zwei Aspekte: die Anzahl der Merkmalsträger und die Unterschiede der auf sie entfallenden Anteile der Merkmalssumme.
- So ist z.B. ein Markt mit nur zwei Anbietern und gleich großen Marktanteilen ebenso wie ein Markt mit 100 Anbietern, von denen die beiden Größten einen Marktanteil von 90% besitzen, konzentriert.

Beispiel 6.2: Armut und Ungleichheit.

Es ist wichtig sich bewusst zu sein, dass Ungleichheit etwas anderes ist als Armut!

Siehe dazu beispielsweise [hier](#).

Die im Allgemeinen sehr empfehlenswerte Quelle dieses Beispiels ist die „Unstatistik des Monats“, siehe [hier](#).

- Man unterscheidet zwei Arten statistischer Konzentration.
- **Absolute Konzentration** (Konzentration i.e.S., kurz Konzentration) berücksichtigt beide Aspekte, indem sie die Anteile an der Merkmalssumme auf die Anzahl der Merkmalsträger bezieht. Eine starke Konzentration ergibt sich, wenn auf eine kleine *Anzahl* von Merkmalsträgern ein großer Anteil der Merkmalssumme entfällt.
- Die **relative Konzentration** (auch Disparität) vernachlässigt den Anzahlaspekt, indem der Anteil der Merkmalssumme nicht zu der Anzahl, sondern zu dem Anteil der Merkmalsträger in Beziehung gesetzt wird. Hohe relative Konzentration bedeutet, dass ein kleiner *Anteil* der Merkmalsträger einen großen Anteil der Merkmalssumme auf sich vereint.

- Konzentration und Disparität werden von zwei Extremzuständen begrenzt:
 - ▶ Hat jeder Merkmalsträger den gleichen Merkmalsbetrag und ist die Anzahl der Merkmalsträger sehr groß, liegt minimale Konzentration vor („egalitäre Verteilung“).
 - ▶ Vollkommene Ungleichheit: Ein Merkmalsträger vereint die gesamte Merkmalssumme auf sich; es liegt maximale Konzentration vor (z.B. Angebotsmonopolist). Alle anderen vorhandenen Merkmalsträger müssen einen Betrag von null haben.
- Da minimale Konzentration die Anzahl der Merkmalsträger berücksichtigt, bedeutet Gleichverteilung mit wenigen Merkmalsträgern nicht zwangsläufig auch geringe absolute Konzentration, siehe obiges Beispiel mit nur zwei gleich großen Anbietern.

- Bei relativer Konzentration würde man bei Gleichverteilung auf minimale Disparität schließen.
- Wir lernen nun Verfahren zur Messung der Konzentration und Disparität kennen. Die Begrenzung der Konzentration durch die o.g. Extremzustände macht es sinnvoll, die Konzentrationsmaße (**Konzentrationsparameter**) zu normieren.
- Liegt keine Konzentration vor, soll der Konzentrationsparameter null sein; bei maximaler Konzentration eins.
- Diese Normierung erleichtert auch den Vergleich unterschiedlicher Datensätze.

Konzentrationsrate und Konzentrationskurve

- Zur Messung der absoluten Konzentration werden die n nicht negativen Beobachtungen eines Merkmals X abnehmend geordnet:

$$x_{(1)} \geq x_{(2)} \geq x_{(3)} \geq \dots \geq x_{(n)} \geq 0,$$

wobei j der Platzierungsindex ist, der im Folgenden zwecks Vereinfachung ohne Klammer geschrieben wird.

- Für häufigkeitsverteilte Daten ist die Ordnung ebenfalls möglich.
- Bei klassierten Daten ist hingegen die Verteilung innerhalb der Klassen meist unbekannt. Da Klassierung zudem zwecks Informationsverdichtung, also Konzentration von vielen Daten auf nur wenige Klassen, erfolgt, ist es nicht sinnvoll, hier die absolute Konzentration zu ermitteln.

Absolute Konzentration

Konzentrationsrate und Konzentrationskurve

Definition 6.3: Konzentrationsrate.

Die Merkmalssumme des Datensatzes ist $\sum_{j=1}^n x_j = n\bar{x}$; der auf den j -ten Merkmalsträger entfallende Anteil c_j der Merkmalssumme ist

$$c_j = \frac{x_j}{n\bar{x}}.$$

Addieren der größten c_j liefert ihren Merkmalssummenanteil

$$C_j = \sum_{r=1}^j c_r, \quad j = 1, \dots, n.$$

C_j bezeichnet man als **Konzentrationsrate** (-koeffizient). Für diesen gilt

$$(1) \quad C_j = c_j + \sum_{r=1}^{j-1} c_r \quad \text{und} \quad (2) \quad C_n = \sum_{r=1}^n c_r = 1.$$

- C_j ist bereits ein einfaches Konzentrationsmaß. Es gibt den Anteil der Merkmalsträger mit den j größten Ausprägungen an der gesamten Merkmalssumme an. $C_1 = 1$ bedeutet maximale Konzentration.
- Nachteilig ist, dass die Wahl von j willkürlich ist.
- Für jedes j erhält man ein C_j . Damit können die sich ergebenden n Zahlenpaare (j, C_j) in ein Koordinatensystem übertragen werden. Die Verbindung der Punkte, beginnend mit dem Ursprung, nennt man **Konzentrationskurve**.

Absolute Konzentration

Konzentrationsrate und Konzentrationskurve

Beispiel 6.4: Umsatzkonzentration auf einem Markt mit 5 Unternehmen.

Fünf Unternehmen teilen sich einen Markt und weisen folgende Umsätze in Mio. € auf: 20, 15, 40, 20, 5.

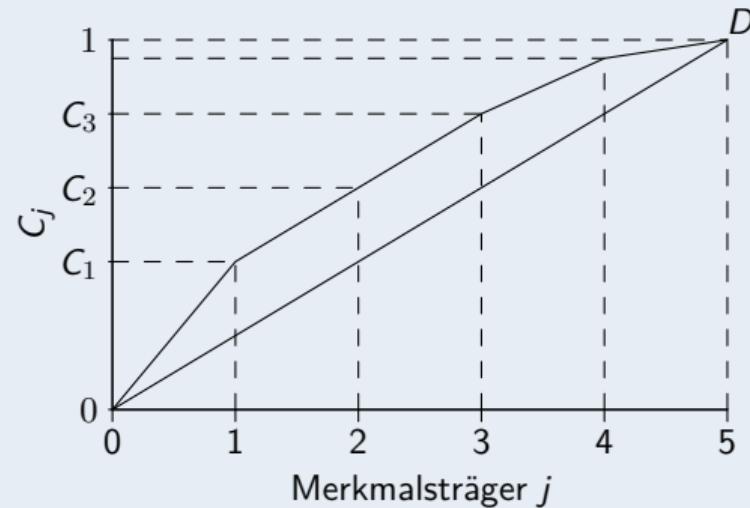
j	x_j	c_j	C_j
1	40	0,40	0,40
2	20	0,20	0,60
3	20	0,20	0,80
4	15	0,15	0,95
5	5	0,05	1,00

$n\bar{x} = 100$

Auf die drei anteilsgrößten Merkmalsträger entfallen 80% ($C_3 = 0,8$) der Merkmalssumme.

Beispiel 6.5: Umsatzkonzentration auf einem Markt mit 5 Unternehmen.

Die resultierende Konzentrationskurve:



Konzentrationsrate und Konzentrationskurve

- Wegen der abnehmenden Ordnung der Beobachtungen liegt die Konzentrationskurve stets oberhalb der Diagonalen $0D$.
- Sind alle x_j gleich $x > 0$, beträgt ihr Anteil $c = \frac{x}{nx} = \frac{1}{n}$. Dann gilt

$$C_j = \sum_{r=1}^j \frac{1}{n} = \frac{j}{n} \quad \text{für} \quad j = 1, \dots, n.$$

- Da konstante c_j bei großem n keine Konzentration bedeuten, ist bei Nichtkonzentration die Konzentrationskurve immer gleich $0D$ („**Gleichverteilungsgerade**“).
- Je weiter nach oben die Konzentrationskurve von der Gleichverteilungsgeraden abweicht, desto größer die absolute Konzentration.
- Als (nicht normiertes) Maß für die Konzentration könnte daher die Fläche zwischen Konzentrationskurve und $[0D]$ herangezogen werden.

Absolute Konzentration

Konzentrationskurve bei häufigkeitsverteilten Daten

- Bei häufigkeitsverteilten Daten kann nach Transformation in Einzelbeobachtungen genauso wie oben vorgegangen werden.
- Bei wenigen x_i , $i = 1, \dots, m$ ist es aber einfacher die Ausprägungen abnehmend zu sortieren: $x_1 > x_2 > \dots > x_m$. Dann ist $c_1 = \frac{n_1 x_1}{n \bar{x}}$ der Anteil, der auf die n_1 Merkmalsträger mit der größten Ausprägung x_1 entfällt.
- Die Konzentrationsrate $C_{s(i)}$ ist dann

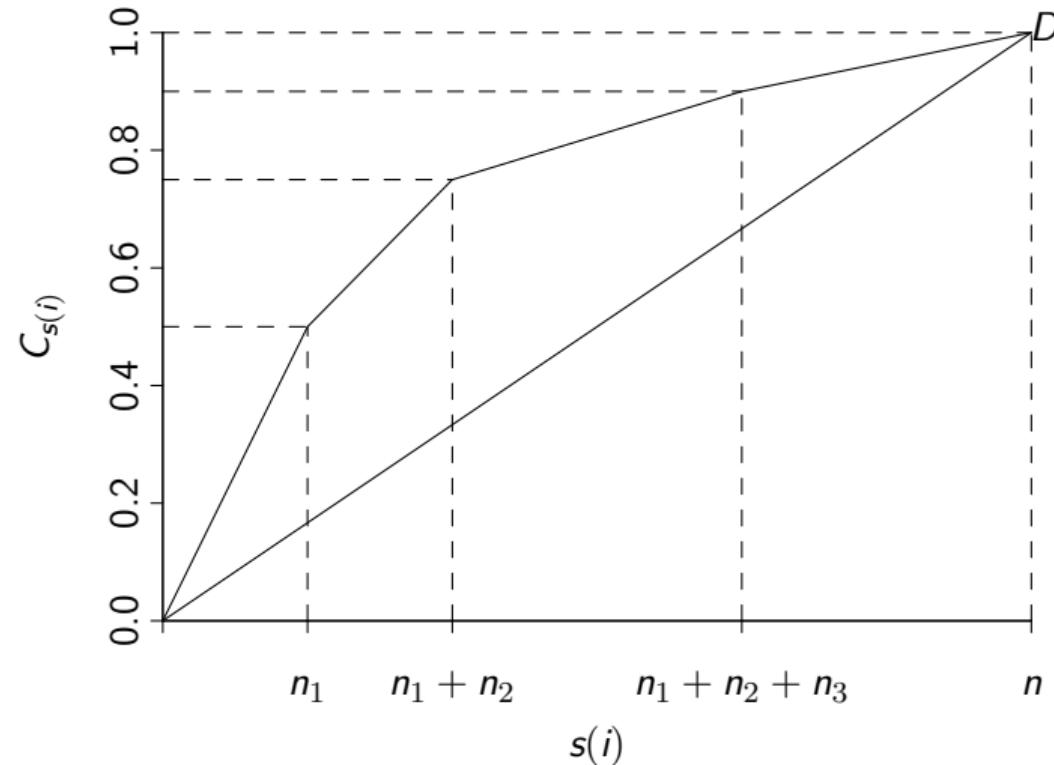
$$C_{s(i)} = \frac{\sum_{r=1}^i n_r x_r}{n \bar{x}}$$

mit $s(i) = \sum_{r=1}^i n_r$, $i = 1, \dots, m$. Die Paare $[s(i), C_{s(i)}]$ liefern die Konzentrationskurve für häufigkeitsverteilte Daten. (Siehe nächste Folie.)

- Die Funktion $s(i)$ ordnet die Konzentrationsraten bei häufigkeitsverteilten Daten der entsprechenden Anzahl an Merkmalsträgern zu.

Absolute Konzentration

Konzentrationskurve bei häufigkeitsverteilten Daten



Absolute Konzentration

Herfindahl-Index C_H

Definition 6.6: Herfindahl-Index.

Ein einfacher, absoluter **Konzentrationsparameter** ist der **Herfindahl-Index** C_H , definiert als Summe der quadrierten Anteilswerte c_j :

$$C_H = \sum_{j=1}^n c_j^2.$$

Bei maximaler Konzentration ist $C_H = 1$, da dann gilt: $c_1 = 1$ und $c_j = 0$ für $j = 2, \dots, n$. Bei Gleichverteilung ($c_j = \frac{1}{n}$) erhält man

$$C_H = \sum_{j=1}^n \frac{1}{n^2} = \frac{n}{n^2} = \frac{1}{n}.$$

Damit ist das Wertebereich $\frac{1}{n} \leq C_H \leq 1$.

Absolute Konzentration

Herfindahl-Index C_H

- Geht die Anzahl der Merkmalsträger gegen unendlich, wird bei Gleichverteilung die Konzentration immer kleiner; C_H strebt gegen null.
- Schreibt man das Quadrat in C_H als $c_j c_j$, wird wegen $0 \leq c_j \leq 1$ deutlich, dass der Herfindahl-Index ein gewogenes arithmetisches Mittel der Anteile c_j ist, wobei die Gewichte gleich den Daten sind.
- Bei großem n gilt die Konzentration für $C_H < 0,10$ als gering, bei $C_H > 0,18$ bereits als hoch. Diese Faustregel gilt auch bei wenigen Merkmalsträgern mit gleichen Anteilen.

Absolute Konzentration

Herfindahl-Index C_H

- Der Herfindahl-Index kann in Abhängigkeit des Variationskoeffizienten v geschrieben werden.
Einsetzen der Anteilswerte in C_H ergibt

$$C_H = \sum_{j=1}^n c_j^2 = \sum_{j=1}^n \left(\frac{x_j}{n\bar{x}} \right)^2 = \frac{\sum_{j=1}^n x_j^2}{n^2 \bar{x}^2}.$$

- Aus dem speziellen Verschiebungssatz folgt $\sum_{j=1}^n x_j^2 = n(s^2 + \bar{x}^2)$. Also gilt

$$C_H = \frac{n(s^2 + \bar{x}^2)}{n^2 \bar{x}^2} = \frac{\frac{s^2}{\bar{x}^2} + 1}{n}$$

oder $C_H = \frac{v^2+1}{n}$.

- Dieser Zusammenhang ist praktisch, da \bar{x} und s^2 oft bereits vorliegen.

```
library(ineq)

x <- c(40,20,20,15,5)
c.j <- sort(x/sum(x), decreasing = T)

(C.H <- sum(c.j^2))

## [1] 0.265

Herfindahl(x)

## [1] 0.265

(moments::moment(x, 2, T)/mean(x)^2+1)/length(x)

## [1] 0.265
```

Konzentration 1

Kahoot!

Relative Konzentration

Lorenzkurve

- Relative Konzentration setzt Anteile der Merkmalssumme zu Anteilen der Merkmalsträger in Beziehung.
- Nun werden alle Daten aufsteigend geordnet.
- Die **Lorenzkurve** wird für Einzelbeobachtungen entwickelt. Die geordneten Beobachtungen $x_1 \leq x_2 \leq \dots \leq x_n$ werden in Anteile c_j an der Merkmalssumme $n\bar{x}$ überführt.
- $C_j = \sum_{r=1}^j c_r, j = 1, \dots, n$ gibt jetzt den kumulierten Anteil der j Merkmalsträger mit den kleinsten Merkmalssummenanteilen wieder. Der kumulierte Anteil H_j dieser j Merkmalsträger beträgt $H_j = \frac{j}{n}, \quad j = 1, \dots, n$.

- Damit erhält man wieder Paare (H_j, C_j) mit $H_n = C_n = 1$.
- Trägt man H_j an der Abszisse und C_j an der Ordinate ein und verbindet beginnend mit dem Ursprung die Punkte, entsteht die Lorenzkurve.

Relative Konzentration

Lorenzkurve

Beispiel 6.7: Umsatzkonzentration auf einem Markt mit 5 Unternehmen.

Ausgangspunkt bildet das schon behandelte Beispiel. Die Arbeitstabelle ist nun:

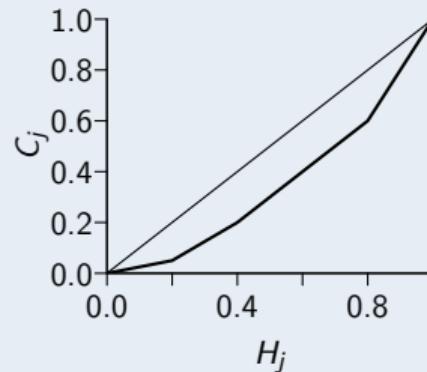
j	x_j	c_j	C_j	$H_j = \frac{j}{n}$
1	5	0,05	0,05	0,2
2	15	0,15	0,20	0,4
3	20	0,20	0,40	0,6
4	20	0,20	0,60	0,8
5	40	0,40	1,00	1,0

Relative Konzentration

Lorenzkurve

Beispiel 6.7: Fortsetzung.

Es folgt die Lorenzkurve, die sich aus den Koordinaten (H_j, C_j) ergibt:



```
library(DescTools)
x <- c(40, 20, 20, 15, 5)
plot(Lc(x), xlab = "$H_j$", ylab = "$C_j$", main = "")
```

Relative Konzentration

Lorenzkurve

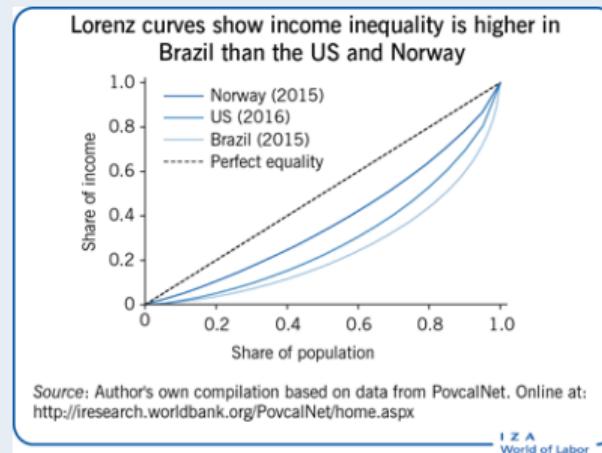
Beispiel 6.7: Fortsetzung.

Da sich die Anzahl der Merkmalsträger diskret verändert, können strenggenommen nur die Punkte (H_j, C_j) mit $j = 1, \dots, n$ interpretiert werden. Der Punkt $(0,40; 0,20)$ bedeutet, dass auf 40% der kleinsten Merkmalsträger nur 20%, auf die übrigen 60% hingegen 80% der Merkmalssumme entfallen.

Relative Konzentration

Lorenzkurve

Beispiel 6.8: Vermögensverteilungen im Vergleich.



Quelle: Trapeznikova, I., Measuring income inequality. IZA World of Labor 2019

Relative Konzentration

Lorenzkurve bei Häufigkeitsverteilten Daten

- Liegen die Daten als Häufigkeitsverteilung vor, entsteht die Lorenzkurve analog zur Konzentrationskurve für diese Datenlage (einiger Unterschied: Merkmalsausprägungen aufsteigend ordnen!). Die C_i erhält man erneut als

$$C_i = \sum_{r=1}^i n_r x_r / n \bar{x};$$

die kumulierten Anteile der Merkmalssumme.

- Die kumulierten Anteile der Merkmalsträger an ihrer Gesamtzahl werden nun berechnet als:
 $H_i = \sum_{r=1}^i n_r / n.$
- Dies liefert die Punkte (H_i, C_i) der Lorenzkurve. Die weitere Vorgehensweise entspricht der für Einzelbeobachtungen.

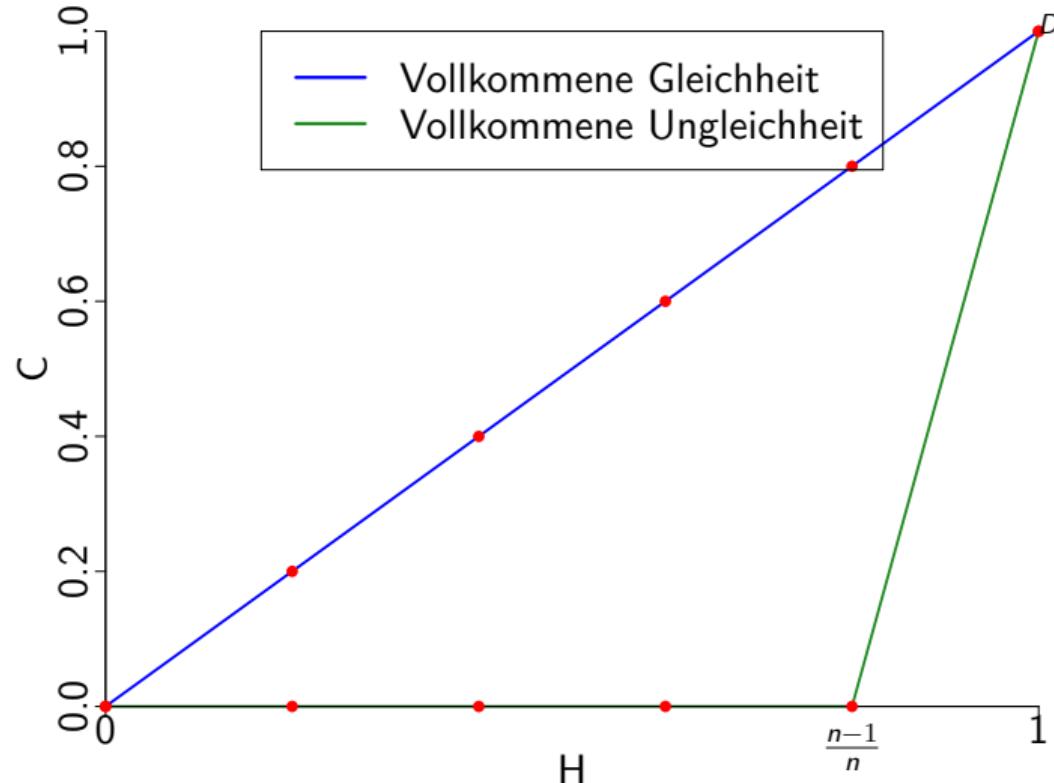
Relative Konzentration

Grenzlagen der Lorenzkurve

- Liegt keine Gleichverteilung vor, verläuft die Lorenzkurve wegen der aufsteigenden Ordnung der Daten konvex zur Abszisse.
- Bei vollkommener Ungleichheit gibt es nur einen Merkmalsträger, der die gesamte Merkmalssumme auf sich vereint. Es gilt dann: $C_j = 0$ für $j = 1, \dots, n - 1$ und $C_n = 1$.
- Die Lorenzkurve verläuft bis zur Stelle $\frac{n-1}{n}$ auf der Abszisse und von da zum Punkt D . Die nächste Folie vergleicht die beiden Grenzlagen bei Gleichverteilung (durchgezogene Diagonale) und vollkommener Ungleichheit (grün); man nutzt sie bei der Konstruktion von relativen Konzentrationsmaßen.

Relative Konzentration

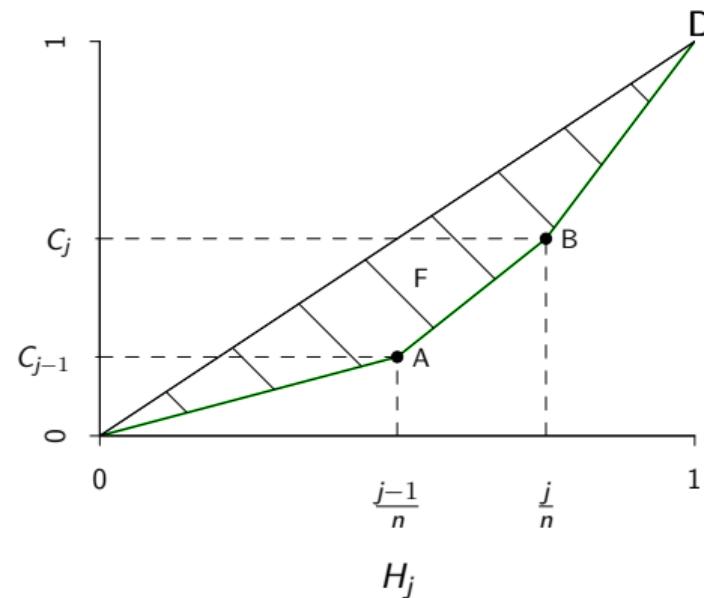
Grenzlagen der Lorenzkurve



Relative Konzentration

Gini-Koeffizient D_G

Das bekannteste relative Konzentrationsmaß ist der **Gini-Koeffizient**, der wegen seines Bezugs zur Lorenzkurve auch Lorenz'sches Konzentrationsmaß heißt. Je ungleicher sich eine Merkmalssumme auf die Merkmalsträger verteilt, desto größer ist die schraffierte Fläche F zwischen Lorenzkurve und der 45° -Gerade OD („Konzentrationsfläche“).



Relative Konzentration

Gini-Koeffizient D_G

- Ihr maximaler Wert F_{max} lässt sich leicht berechnen. Die Fläche des Dreiecks $(0, 1, D)$ ist $\frac{1}{2}$; die des Dreiecks $(\frac{n-1}{n}, 1, D)$ ist $\frac{1}{2n}$. F_{max} ist dann

$$F_{max} = \frac{1}{2} - \frac{1}{2n} = \frac{1}{2} \left(1 - \frac{1}{n}\right) = \frac{1}{2} \frac{n-1}{n} < \frac{1}{2}.$$

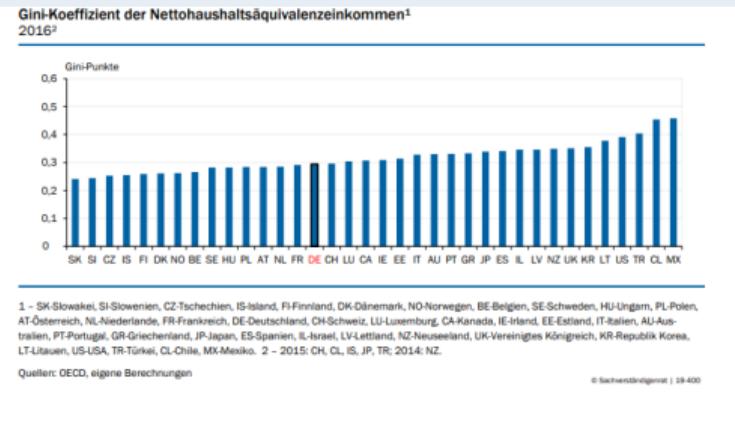
- Für die Konstruktion eines Konzentrationsmaßes mit Werten in $[0, 1]$ gibt es zwei Möglichkeiten:
 - ① Beziehe F auf $\frac{1}{2}$ (Fläche des Dreiecks $(0, 1, D)$) oder
 - ② auf F_{max} .

Beide Quotienten bezeichnet man als Gini-Koeffizienten. [GiniFussballigen.R](#)

Relative Konzentration

Gini-Koeffizient D_G

Beispiel 6.9: OECD.

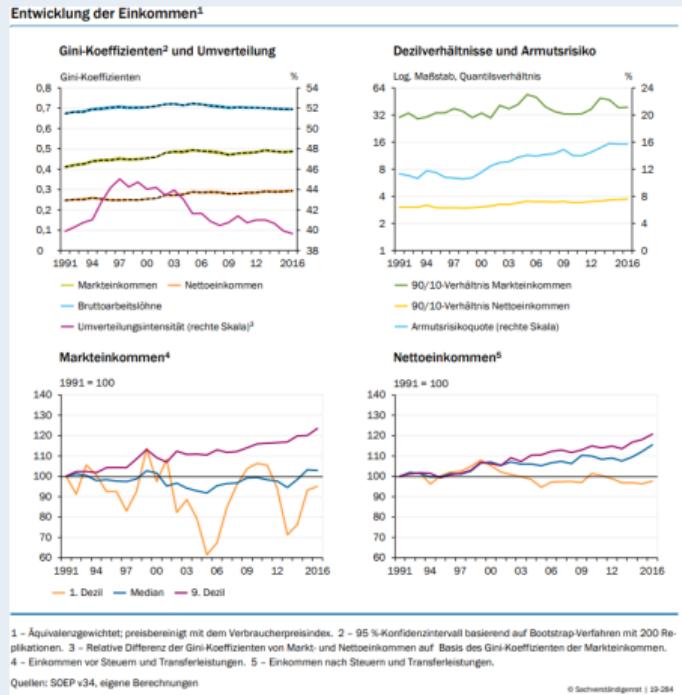


Quelle: OECD (2020), How's Life? 2020:Measuring Well-being, OECD Publishing

Relative Konzentration

Gini-Koeffizient, D_c

Beispiel 6.10: Entwicklung der Einkommen in Deutschland.



Quelle: Sachverständigenrat (2020), Jahresgutachten 2019/20

- „Äquivalenzgewichtet“ bedeutet, dass ein Vierpersonenhaushalt kein vier Mal so hohes Einkommen für den gleichen Lebensstandard benötigt wie ein Single-Haushalt.
- Nach Steuern sind Einkommen gleicher als vor Steuern.
- Hat sich die Ungleichheit nach Steuern deutlich geändert?
- Was ist insbesondere seit der Agenda 2010 passiert?

Relative Konzentration

Gini-Koeffizient D_G

Definition 6.11: Gini-Koeffizient.

Die erste Methode liefert $D_G = 2F$. Sein maximaler Wert für $F = F_{max}$ ist

$$D_{G,max} = 1 - \frac{1}{n} = \frac{n-1}{n} < 1.$$

Methode 2 führt zu D_G^* , welcher proportional zu D_G ist:

$$D_G^* = \frac{F}{F_{max}} = \frac{2F}{1 - \frac{1}{n}} = \frac{n}{n-1} D_G.$$

- Da (im Gegensatz zu D_G) $D_G^* = 1$ bei vollkommener relativer Konzentration ($F = F_{max}$), heißt D_G^* auch **normierter Gini-Koeffizient**.
- Zur Berechnung ist die Konzentrationsfläche F zu bestimmen. Die Vorgehensweise erfolgt für Einzelbeobachtungen, häufigkeitsverteilte und klassierte Daten getrennt. In allen Fällen wird zunächst die über der Lorenzkurve liegende Fläche in Trapeze zerlegt.
- Bei Einzelbeobachtungen ergibt sich die Fläche F_j des Trapezes (ABC_jC_{j-1}) als

$$F_j = \frac{1}{2}(\overrightarrow{C_{j-1}A} + \overrightarrow{C_jB})(\overrightarrow{C_{j-1}C_j})$$

Relative Konzentration

Gini-Koeffizient bei Einzelbeobachtungen

- Nach Substitution folgt

$$F_j = \frac{1}{2} \left(\frac{j-1}{n} + \frac{j}{n} \right) c_j = \frac{2j-1}{2n} c_j, \quad c_j = C_j - C_{j-1}.$$

- Addiert man alle Trapezflächen F_j und subtrahiert hiervon den Flächeninhalt des Dreiecks über der Geraden $0D$, erhält man F

$$F = \sum_{j=1}^n F_j - \frac{1}{2}.$$

Substitution von F_j durch die obige Beziehung ergibt wegen $\sum_{j=1}^n c_j = 1$

$$F = \sum_{j=1}^n \frac{2j-1}{2n} c_j - \frac{1}{2} = \frac{2 \sum_{j=1}^n j c_j - \sum_{j=1}^n c_j}{2n} - \frac{1}{2} = \frac{2 \sum_{j=1}^n j c_j - 1}{2n} - \frac{1}{2}.$$

Relative Konzentration

Gini-Koeffizient bei Einzelbeobachtungen

- Der Gini-Koeffizient $D_G = 2F$ ist dann

$$D_G = \frac{2 \sum_{j=1}^n j c_j - 1}{n} - 1.$$

- Obige Gleichung kann so umgeformt werden, dass D_G direkt aus den Einzelbeobachtungen x_j folgt. Dies ist dann von Vorteil, wenn die Lorenzkurve nicht erstellt werden soll.
- Schreibt man $1 = \frac{n}{n}$ und $c_j = \frac{x_j}{n\bar{x}}$ folgt

$$D_G = \frac{\frac{2 \sum_{j=1}^n j x_j}{n\bar{x}} - (1 + n)}{n} = \frac{2 \sum_{j=1}^n j x_j - (1 + n) \sum_{j=1}^n x_j}{n \sum_{j=1}^n x_j},$$

da $n\bar{x} = \sum_{j=1}^n x_j$.

Relative Konzentration

Gini-Koeffizient bei Einzelbeobachtungen - in 

```
library(ineq)
x    <- c(40, 20, 20, 15, 5)
x    <- sort(x)                      # aufsteigend sortieren
c.j  <- x / sum(x)                  # Anteil an der Merkmalssumme
n    <- length(x)                   # Anzahl Beobachtungen
(D.G <- (2*sum(c.j*(1:n)) - 1)/n - 1) # Variante D_G
## [1] 0.3

Gini(x)                         # einfacher
## [1] 0.3

n / (n - 1) * D.G               # D_G*
## [1] 0.375

Gini(x, corr = TRUE)             # einfacher
## [1] 0.375
```

Konzentration 2

Kahoot!

Absolute vs. relative Konzentration

```
library(ineq)

(x0 <- c(rep(0.2, 4), rep(0.05, 4)))      # "4 große, 4 kleine Unternehmen am Markt"
## [1] 0.20 0.20 0.20 0.20 0.05 0.05 0.05 0.05

(x1 <- rep(0.25, 4))                      # "die 4 Großen schlucken die 4 Kleinen"
## [1] 0.25 0.25 0.25 0.25

# Was passiert nun mit "der" Konzentration?
```

Absolute vs. relative Konzentration

```
# relativ
Gini(x0)                                # vorher bereits etwas rel. Konzentration
## [1] 0.3

Gini(x1)                                # da die 4 großen nun gleich groß sind,
# zeigt sich rel. Konzentration von null
## [1] 0

# absolut
Herfindahl(x0)                            # abs. Konzentration vorher
## [1] 0.17

Herfindahl(x1)                            # abs. Konzentration ist nun *gewachsen*,
# da nun geringere Zahl an Anbietern
# berücksichtigt wird
## [1] 0.25
```

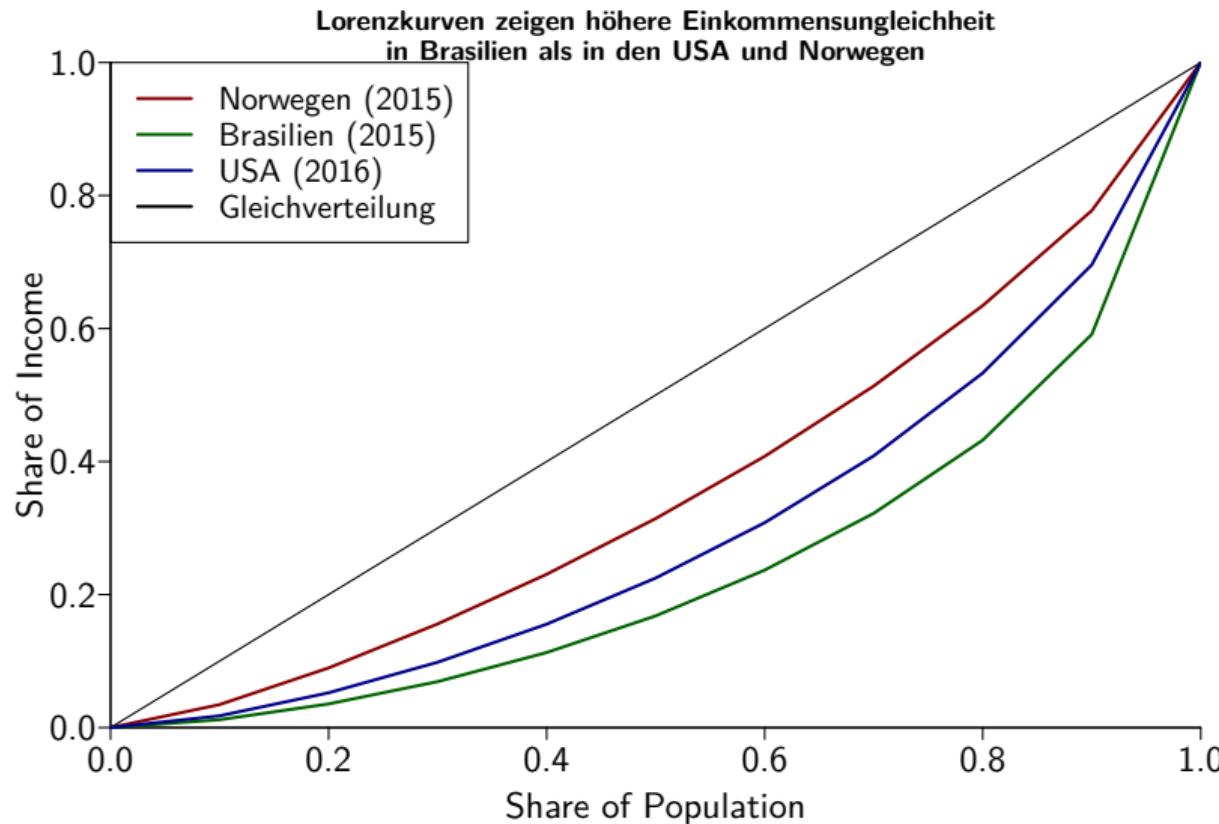
- Gegenstand dieser Vorlesung war das Thema „Konzentration“. Hiermit kann nun auch die Verteilung der Merkmalsträger auf die Merkmalssumme betrachtet werden, was z.B. bei der Beurteilung von Marktmacht eine große Rolle spielt.
- Nachbereitung: Kapitel 4.5 des Buches von Prof. Assenmacher.
- Das nächste Kapitel beginnt mit der Analyse zweidimensionaler Häufigkeitsverteilungen. Dies ist sehr praxisrelevant, da man sich häufig für den Zusammenhang zweier (oder mehr) Variablen interessiert.
- Vorbereitung: Kapitel 5.1 und 5.2 des Buches von Prof. Assenmacher.

Anhang

```
library(povcalnetR)
library(ineq)

Datensatz <- rbind(povcalnet(country = c("NOR", "BRA"), year = 2015),
                     povcalnet(country = "USA", year = 2016))
Datensatz <- Datensatz[, 22:31]
Datensatz <- as.data.frame(t(Datensatz))
colnames(Datensatz) <- c("BRA", "NOR", "USA")

Farben <- c("darkred", "darkgreen", "darkblue", "black")
plot(Lc(Datensatz$NOR), col = Farben[1], xlab = "Share of Population",
     ylab = "Share of Income",
     main = 'Lorenzkurven zeigen höhere Einkommensungleichheit
     in Brasilien als in den USA und Norwegen',
     lwd = 2, lty = 1)
lines(Lc(Datensatz$BRA), col = Farben[2], lwd = 2, lty = 1)
lines(Lc(Datensatz$USA), col = Farben[3], lwd = 2, lty = 1)
legend("topleft",
       legend = c("Norwegen (2015)", "Brasilien (2015)",
                 "USA (2016)", "Gleichverteilung"),
       col = Farben, lty = 1, lwd = 2)
```



- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze**
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

- Erfassen von zwei Variablen X und Y für n Merkmalsträger $\omega_1, \dots, \omega_n$ liefert einen **bivariaten Datensatz**. Die Urliste besteht hier aus Zahlenpaaren $(x_1, y_1), \dots, (x_n, y_n)$. **Beobachtungsmatrix**:

Merkmalsträger	X	Y
ω_1	x_1	y_1
ω_2	x_2	y_2
\vdots	\vdots	\vdots
ω_n	x_n	y_n

- Aus der Beobachtungsmatrix kann man eine **zweidimensionale Häufigkeitstabelle** gewinnen. Der Index i kennzeichnet die Ausprägungen von X : $i = 1, \dots, m$; j bezeichnet jetzt nicht mehr die Beobachtungen, sondern die Ausprägungen von Y : $j = 1, \dots, l$.

Häufigkeitstabellen bivariater Datensätze

Die n_{ij} geben an, wie oft (x_i, y_j) in der Urliste vorkommt, d.h. wie viele Merkmalsträger sowohl die Ausprägung x_i als auch y_j aufweisen. Es gilt $n_{ij} \geq 0$ und $\sum_{j=1}^l \sum_{i=1}^m n_{ij} = n$.

X/Y	y_1	y_2	\cdots	y_l	$n_{i\cdot}(h_{i\cdot})$
x_1	$n_{11}(h_{11})$	n_{12}	\cdots	n_{1l}	$n_{1\cdot}$
x_2	$n_{21}(h_{21})$	n_{22}	\cdots	n_{2l}	$n_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_m	$n_{m1}(h_{m1})$	n_{m2}	\cdots	n_{ml}	$n_{m\cdot}$
$n_{\cdot j}(h_{\cdot j})$	$n_{\cdot 1}$	$n_{\cdot 2}$	\cdots	$n_{\cdot l}$	n

$$n_{\cdot j} = \sum_{i=1}^m n_{ij} \quad n_{i\cdot} = \sum_{j=1}^l n_{ij}$$

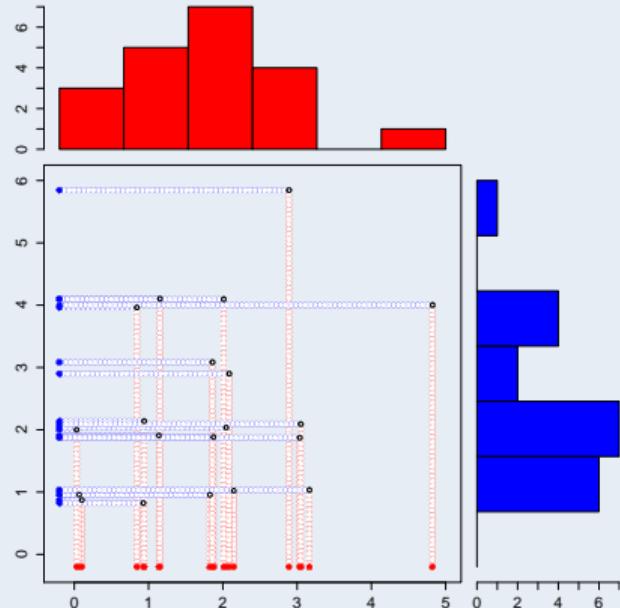
- Die Tabelle heißt auch $(m \times l)$ -Feldertafel, bzw. bei kardinalen Merkmalen X und Y oft Korrelationstabelle; sind X und Y ordinal skaliert, heißt sie Kontingenztabelle (manchmal auch für kardinale).
- Die n_{ij} bilden die **absolute bivariate Häufigkeitsverteilung**.
- Man kann auch die **relativen bivariaten Häufigkeiten** $h_{ij} = \frac{n_{ij}}{n}$ angeben.
- Analog sprechen wir von absoluten bzw. relativen bivariaten Häufigkeitsfunktionen:

$$n(X = x_i, Y = y_j) = n_{ij} \quad \text{bzw.} \quad h(X = x_i, Y = y_j) = h_{ij}$$

mit $i = 1, \dots, m$ und $j = 1, \dots, l$.

Häufigkeitstabelle

Beispiel 7.1: Graphisch.



Dynamisch wird dies eventuell noch deutlicher: `MarginalDistributions.R`

Häufigkeitstabelle

Beispiel 7.2: Anzahl an Wählerstimmen bei der Bundestagswahl 2021 nach Partei und Geschlecht (in Tsd.).

X/Y	$y_1 = \text{Frauen}$	$y_2 = \text{Männer}$	$n_i.$
$x_1 = \text{CDU}$	5469.2	4982.4	10451.6
$x_2 = \text{SPD}$	6497.7	5737.0	12234.7
$x_3 = \text{AfD}$	1841.0	2854.6	4695.6
$x_4 = \text{FDP}$	1932.6	2110.4	4043.0
$x_5 = \text{Linke}$	1199.8	1107.7	2307.5
$x_6 = \text{Grüne}$	3638.1	2831.0	6469.1
$x_7 = \text{CSU}$	1443.4	1344.6	2788.0
$x_8 = \text{Sonstige}$	1825.9	1546.7	3372.6
$n_{\cdot j}$	23847.7	22514.4	46362.1

Quelle: Informationen des Bundeswahlleiters | Statistisches Bundesamt, Bundestagswahl 2021, Heft 4: Repräsentative Wahlstatistik.
Die Gesamtanzahl an Wählerstimmen bezieht sich ausschließlich auf gültige Stimmabgaben.

Häufigkeitstabelle

Beispiel 7.3: Anteil an Wählerstimmen bei der Bundestagswahl 2021 nach Partei und Geschlecht.

X/Y	$y_1 = \text{Frauen}$	$y_2 = \text{Männer}$	h_i
$x_1 = \text{CDU}$	11.8	10.7	22.5
$x_2 = \text{SPD}$	14.0	12.4	26.4
$x_3 = \text{AfD}$	4.0	6.2	10.2
$x_4 = \text{FDP}$	4.2	4.6	8.8
$x_5 = \text{Linke}$	2.6	2.4	5.0
$x_6 = \text{Grüne}$	7.8	6.1	13.9
$x_7 = \text{CSU}$	3.1	2.9	6.0
$x_8 = \text{Sonstige}$	3.9	3.3	7.2
h_j	51.4	48.6	100

Quelle: Informationen des Bundeswahlleiters | Statistisches Bundesamt, Bundestagswahl 2021, Heft 4: Repräsentative Wahlstatistik.
Die Gesamtanzahl an Wählerstimmen bezieht sich ausschließlich auf gültige Stimmabgaben.
Relative bivariate Häufigkeiten sind in Prozent angegeben.

Kumulation der bivariaten Häufigkeitsfunktionen liefert **absolute** und **relative gemeinsame** bzw. bivariate **Häufigkeitssummenfunktionen**:

$$N(X \leq x, Y \leq y) = N(x, y) = \sum_{j \text{ mit } y_j \leq y} \sum_{i \text{ mit } x_i \leq x} n_{ij} \quad \text{bzw.}$$

$$H(X \leq x, Y \leq y) = H(x, y) = \sum_{j \text{ mit } y_j \leq y} \sum_{i \text{ mit } x_i \leq x} h_{ij},$$

Die relative Häufigkeitssummenfunktion wird auch **empirische Verteilungsfunktion** genannt.

Häufigkeitssummenfunktion

Insgesamt ergibt sich

$$N(x, y) = \begin{cases} 0 & \text{für } x < x_{i=1} \text{ oder } y < y_{j=1} \\ N(x_i, y_j) & \text{für } x_i \leq x < x_{i+1}, y_j \leq y < y_{j+1} \\ n & \text{für } x \geq x_m \text{ und } y \geq y_l \end{cases}$$

und

$$H(x, y) = \begin{cases} 0 & \text{für } x < x_{i=1} \text{ oder } y < y_{j=1} \\ H(x_i, y_j) & \text{für } x_i \leq x < x_{i+1}, y_j \leq y < y_{j+1} \\ 1 & \text{für } x \geq x_m \text{ und } y \geq y_l \end{cases}$$

Häufigkeitssummenfunktion

Randhäufigkeiten

- Aus der Kontingenztabelle lassen sich die Häufigkeiten der Ausprägungen von X und Y auch getrennt gewinnen. Addiert man n_{ij} bzw. h_{ij} für festes i über j , erhält man die Häufigkeit der Merkmalsausprägung x_i .
- Die so gebildete Summe heißt absolute bzw. relative Randhäufigkeit, geschrieben $n_{i\cdot}$ bzw. $h_{i\cdot}$:

$$n_{i\cdot} = \sum_{j=1}^I n_{ij} \quad \text{und} \quad h_{i\cdot} = \sum_{j=1}^I h_{ij}$$

- Alle $n_{i\cdot}$ bzw. $h_{i\cdot}$ bilden die **absolute bzw. relative Randhäufigkeitsverteilung** von X . Entsprechend für Y :

$$n_{\cdot j} = \sum_{i=1}^m n_{ij} \quad \text{und} \quad h_{\cdot j} = \sum_{i=1}^m h_{ij}$$

Kontingenztabelle

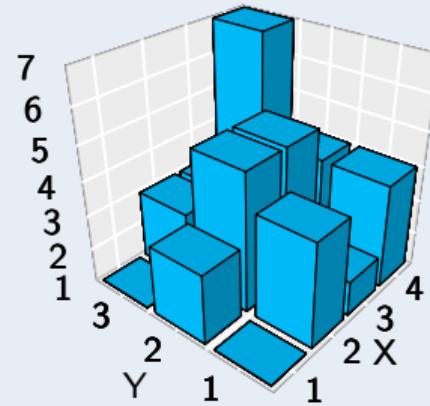
Beispiel 7.4: Einfache Kontingenztabelle.

Die beiden diskreten Variablen X und Y haben die Ausprägungen $x_1 < x_2 < x_3 < x_4$ und $y_1 < y_2 < y_3$. Die Tabelle liefert n_{ij} , die absolute bivariate Häufigkeitsverteilung findet sich auf der nächsten Folie.

X/Y	y_1	y_2	y_3	$n_{i\cdot}$
x_1	1	3	1	5
x_2	4	5	3	12
x_3	2	5	3	10
x_4	4	4	7	15
$n_{\cdot j}$	11	17	14	$n = 42$

Beispiel 7.4: Fortsetzung.

Lediglich aus optischen Gründen sind die Abstände zwischen den Ausprägungen gleich groß.



Beispiel 7.5: Gemeinsame Häufigkeitssummenfunktion.

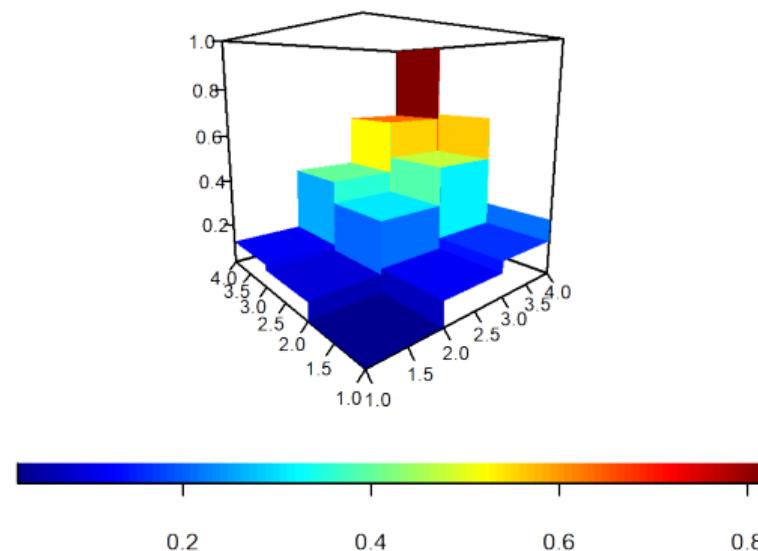
Die Berechnung der absoluten gemeinsamen Häufigkeitssummen mit z.B. $X = x_3$ und $Y = y_2$ ergibt

$$\begin{aligned}N(x_3, y_2) &= \sum_{j=1}^2 \sum_{i=1}^3 n_{ij} \\&= n_{11} + n_{21} + n_{31} + n_{12} + n_{22} + n_{32} = 1 + 4 + 2 + 3 + 5 + 5 = 20.\end{aligned}$$

Für alle Werte x und y mit $x_3 \leq x < x_4$ und $y_2 \leq y < y_3$ bleibt N auf dem Wert 20. Daher ist der Graph von N über dem Rechteck $x_3 \leq x < x_4$ und $y_2 \leq y < y_3$ eine zur (X, Y) -Ebene im Abstand von 20 liegende parallele Fläche.

Gemeinsame Häufigkeitssummen

Graphisch:



Häufigkeitstabelle - in R

```
Feldertafel <- as.table(matrix(c(1, 3, 1, 4, 5, 3,
                                2, 5, 3, 4, 4, 7),
                                nrow = 4, byrow = T))

rownames(Feldertafel) <- 1:nrow(Feldertafel)
colnames(Feldertafel) <- 1:ncol(Feldertafel)

Feldertafel

##   1 2 3
## 1 1 3 1
## 2 4 5 3
## 3 2 5 3
## 4 4 4 7

n <- sum(Feldertafel)
```

```
hij <- prop.table(Feldertafel)
round(hij, 3)

##      1     2     3
## 1 0.024 0.071 0.024
## 2 0.095 0.119 0.071
## 3 0.048 0.119 0.071
## 4 0.095 0.095 0.167
```

```
# Randverteilungen
(ni. <- margin.table(Feldertafel, 1))

##  1  2  3  4
##  5 12 10 15

(n.j <- margin.table(Feldertafel, 2))

##  1  2  3
## 11 17 14

(hi. <- margin.table(hij, 1))

##          1          2          3          4
## 0.1190476 0.2857143 0.2380952 0.3571429

(h.j <- margin.table(hij, 2))

##          1          2          3
## 0.2619048 0.4047619 0.3333333
```

```
x <- 1:nrow(Feldertafel)
y <- 1:ncol(Feldertafel)

# Momente der Randverteilung
(mean.x <- sum(x * hi.))
## [1] 2.833333

(s2X <- sum(x^2 * hi.) - mean.x^2)
## [1] 1.09127
```

- Es existieren noch weitere Verteilungen, die wichtige Information über die Struktur bivariater Datensätze liefern.
- Die Häufigkeiten einer beliebigen Zeile i der Tabelle geben etwa an, wie häufig die einzelnen Ausprägungen von Y bei den Merkmalsträgern vorkommen, die gleichzeitig für X die Ausprägung x_i aufweisen.
- Analog geben die Häufigkeiten einer beliebigen Spalte j an, wie sich die Merkmalsträger mit Ausprägung y_j auf die Ausprägungen von X aufteilen.
- Diese Verteilungen, die von der Vorgabe x_i bzw. y_j abhängen, heißen **bedingte Verteilungen**. Die Bedingung wird rechts von einem senkrechten Strich aufgeführt.
- Die weiteren Ausführungen beziehen sich auf die bedingten Verteilungen von Y ; sie gelten natürlich analog auch für X .

- Die Funktion $n_y(y_j \mid X = x_i)$ gibt für jede Merkmalsausprägung y_j die Anzahl der Merkmalsträger an, bei denen die Ausprägung x_i vorliegt (bedingte absolute Häufigkeiten). Schreibe kurz:
 $n_y(y_j \mid X = x_i) = n_{j|i}.$
- $n_{j|i}$ liegt mit der Kontingenztabelle vor.
- Lautet die Bedingung z.B. $X = x_2$, erhält man die bedingten absoluten Häufigkeiten aus der zweiten Zeile als

$$n_{1|2} = n_{21}, n_{2|2} = n_{22}, \dots, n_{l|2} = n_{2l}.$$

- Die Paare $(y_j, n_{j|i})$, $j = 1, \dots, l$ stellen die bedingte Häufigkeitsverteilung für $X = x_i$ dar.

- Jede Bedingung x_i liefert eine bedingte Verteilung; da jede Ausprägung x_i mit ihrer Randhäufigkeit $n_{i\cdot}$ vorkommt, gilt für jede bedingte Verteilung: $\sum_{j=1}^I n_{j|i} = n_{i\cdot}, i = 1, \dots, m.$
- Nicht alle $n_{i\cdot}$ gleich \Rightarrow Überführung bedingter absoluter in bedingte relative Häufigkeiten.
- Dividiere hierzu die $n_{j|i}$ durch die Anzahl ihrer Beobachtungen, also durch $n_{i\cdot}$ (und nicht durch $n!$):

$$h_{j|i} = \frac{n_{j|i}}{n_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}} \quad \text{für } i \text{ fest und } j \text{ variabel.}$$

- $h_{j|i}$ ist die relative Häufigkeit von y_j unter der Bedingung $X = x_i$; man schreibt auch $h_y(y_j | X = x_i) = h_{j|i}$. Für $h_{j|i}$ gilt

$$h_{j|i} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\cdot}}{n}} = \frac{h_{ij}}{h_{i\cdot}} \quad \text{und} \quad \sum_{j=1}^I h_{j|i} = 1.$$

```
# bedingte Verteilungen
```

```
h_j_bar_i <- prop.table(Feldertafel, 1)
h_i_bar_j <- prop.table(Feldertafel, 2)
round(h_j_bar_i, 3)
```

```
##      1     2     3
## 1 0.200 0.600 0.200
## 2 0.333 0.417 0.250
## 3 0.200 0.500 0.300
## 4 0.267 0.267 0.467
```

```
round(h_i_bar_j, 3)
```

```
##      1     2     3
## 1 0.091 0.176 0.071
## 2 0.364 0.294 0.214
## 3 0.182 0.294 0.214
## 4 0.364 0.235 0.500
```

```
#...summieren sich zu 1
rowSums(h_j_bar_i)

## 1 2 3 4
## 1 1 1 1
```

```
colSums(h_i_bar_j)

## 1 2 3
## 1 1 1
```

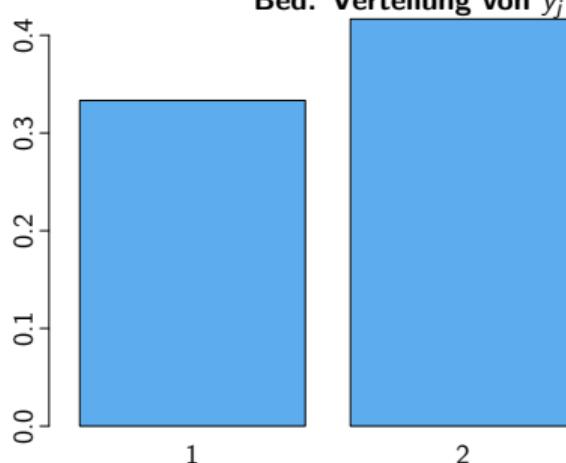
Bedingte Verteilungen und empirische Unabhängigkeit

- in 

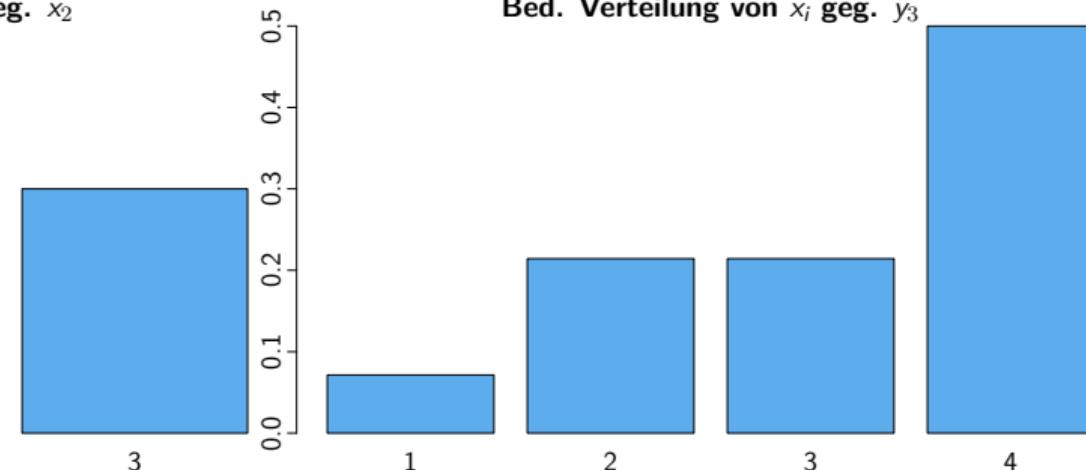
```
barplot(h_j_bar_i[2,], main = "Bed. Verteilung von $y_j$ geg. $x_2$",
        col = "steelblue2")
```

```
barplot(h_i_bar_j[,3], main = "Bed. Verteilung von $x_i$ geg. $y_3$",
        col = "steelblue2")
```

Bed. Verteilung von y_j geg. x_2



Bed. Verteilung von x_i geg. y_3



- Die Zahlenpaare $\{(y_j, h_{j|i}), j = 1, \dots, l\}$ ergeben die bedingte relative Häufigkeitsverteilung von Y für $X = x_i$.
- Analog ergeben sich bedingte Verteilungen für X . Bezeichne die bedingten Häufigkeiten jetzt mit $n_{i|j}$ und $h_{i|j}$, die Häufigkeitsfunktionen mit $n_x(x_i | Y = y_j)$ und $h_x(x_i | Y = y_j)$, für $i = 1, \dots, m$.
- Das arithmetische Mittel der bedingten Verteilungen (bedingtes arithmetisches Mittel) erhält man als

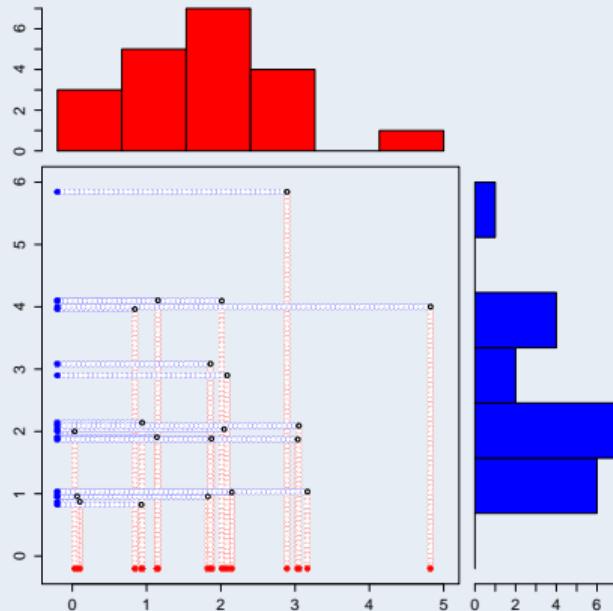
$$\bar{x} | y_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^m x_i n_{i|j} = \sum_{i=1}^m x_i h_{i|j}$$

und

$$\bar{y} | x_i = \frac{1}{n_{i \cdot}} \sum_{j=1}^l y_j n_{j|i} = \sum_{j=1}^l y_j h_{j|i}$$

```
# ein bedingter Mittelwert
(bedingter.Mittelwert.x.gegeben.y1 <- sum(x* h_i_bar_j[,1]))
## [1] 2.818182
```

Beispiel 7.6: Graphisch.



Auch dies kann man sich anhand obiger Abbildung plausibel machen.

Beispiel 7.7: Bundestagswahl.

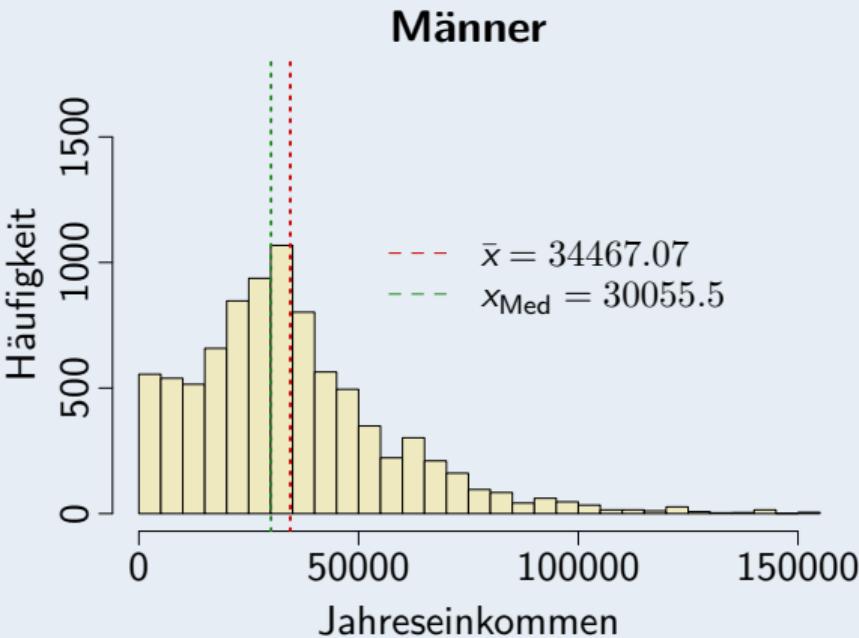
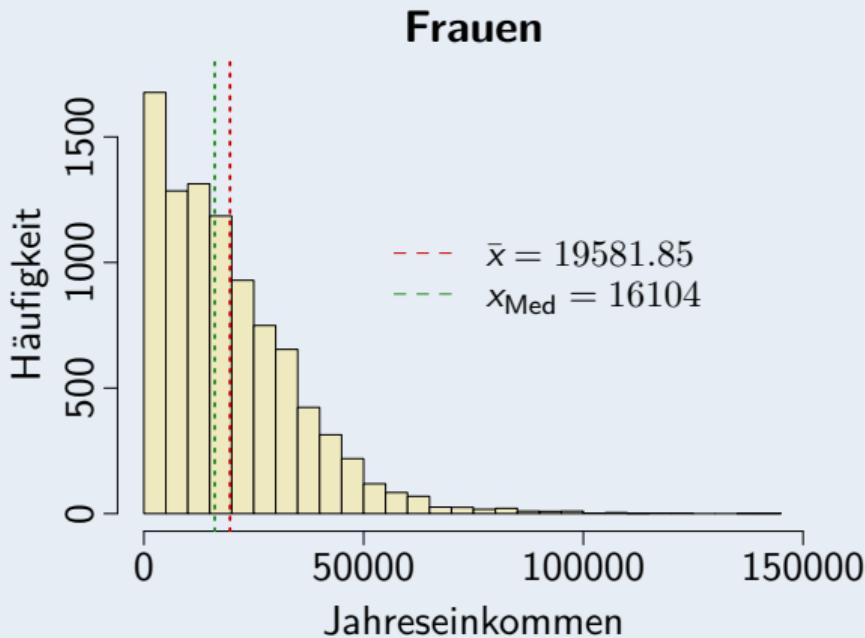
Wie wäre die Wahl ausgegangen, wenn...

Zweidimensionale Datensätze 1

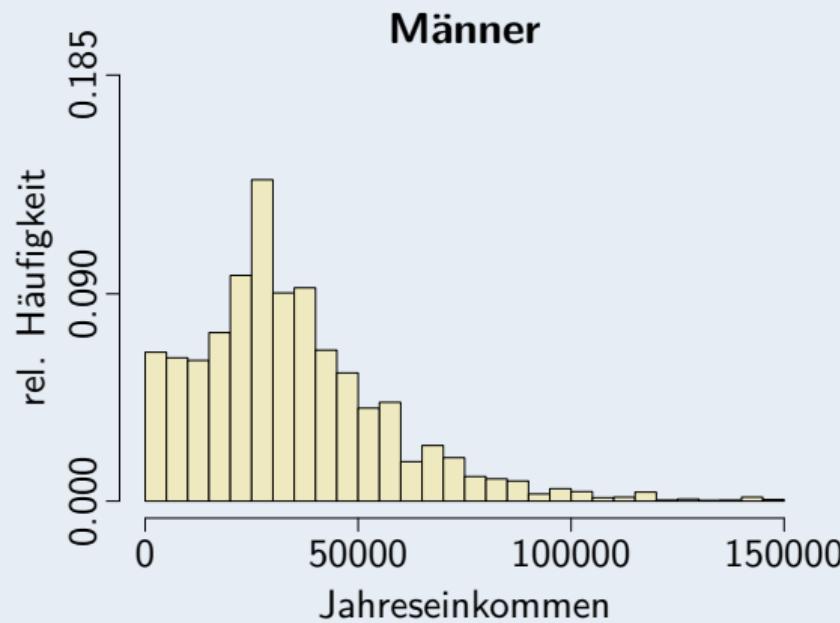
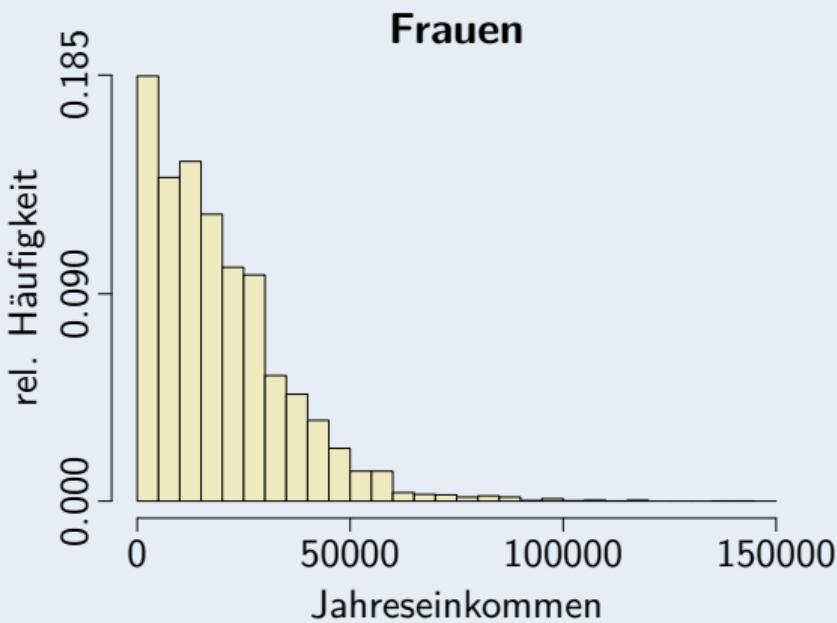
Kahoot!

Beispiel 7.8:

Jahresgehälter in Deutschland im Jahr 2013 (Quelle: SOEP)



Beispiel 7.8: Fortsetzung.



- Sind die m bedingten relativen Häufigkeitsfunktionen $h_y(y_j | X = x_i)$ für alle $i = 1, \dots, m$ gleich, dann ist die Verteilung von Y unabhängig von der Bedingung für X .
- Die m bedingten relativen Verteilungen stimmen dann auch mit der relativen Randverteilung $h_{\cdot j}$ von Y überein:

$$h_{j|1} = h_{j|2} = \dots = h_{j|m} = h_{\cdot j} \quad \text{für } j = 1, \dots, l$$

Denn: Schreibe $h_{\cdot j} = \sum_{i=1}^m h_{ij} = \sum_{i=1}^m h_{j|i} h_{i\cdot}$. Sei $h_{j|i} \equiv h_j$, unabhängig von i . Dann folgt $h_{\cdot j} = \sum_{i=1}^m h_j h_{i\cdot} = h_j \sum_{i=1}^m h_{i\cdot} = h_j$.

- Aus $h_{j|i} = h_{ij}/h_{i\cdot}$ folgt bei Unabhängigkeit der Variablen Y von X :

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

und hieraus: $h_{ij} = h_{i\cdot} h_{\cdot j}$.

- Ist Y von X unabhängig, gilt auch die Umkehrung. Die bedingten relativen Häufigkeiten von X stimmen für jede Bedingung y_j überein:

$$h_{i|1} = h_{i|2} = \dots = h_{i|l} = h_i. \quad \text{für } i = 1, \dots, m.$$

Denn: $\frac{h_{ij}}{h_{\cdot j}} = h_i$. Da $\frac{h_{ij}}{h_{\cdot j}} = h_{i|j}$ ist, folgt: $h_{i|j} = h_i$ für alle $i = 1, \dots, m$.

- X und Y beeinflussen sich dann nicht, man bezeichnet sie daher als **(empirisch) unabhängig**. Bei Unabhängigkeit enthalten die Randverteilungen bereits die Information aus der Kontingenztabelle.
- Lassen sich alle h_{ij} als Produkt der entsprechenden Randhäufigkeiten darstellen, gilt also: $h_{ij} = h_{i\cdot} \cdot h_{\cdot j}$, dann sind X und Y empirisch unabhängig. Gilt dies für mindestens ein h_{ij} nicht, liegt Abhängigkeit vor.
- Alternativ kann geprüft werden, ob $n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$ gilt.

```
# unabhängig?  
round(hij, 3)  
  
##      1     2     3  
## 1 0.024 0.071 0.024  
## 2 0.095 0.119 0.071  
## 3 0.048 0.119 0.071  
## 4 0.095 0.095 0.167  
  
round(outer(hi., h.j, "*"), 3)  
  
##      1     2     3  
## 1 0.031 0.048 0.040  
## 2 0.075 0.116 0.095  
## 3 0.062 0.096 0.079  
## 4 0.094 0.145 0.119
```

```
# alternativ: keine identischen Zeilen
```

```
round(h_j_bar_i,3)
```

```
##      1     2     3
```

```
## 1 0.200 0.600 0.200
```

```
## 2 0.333 0.417 0.250
```

```
## 3 0.200 0.500 0.300
```

```
## 4 0.267 0.267 0.467
```

```
# alternativ: keine identischen Spalten
```

```
round(h_i_bar_j,3)
```

```
##      1     2     3
```

```
## 1 0.091 0.176 0.071
```

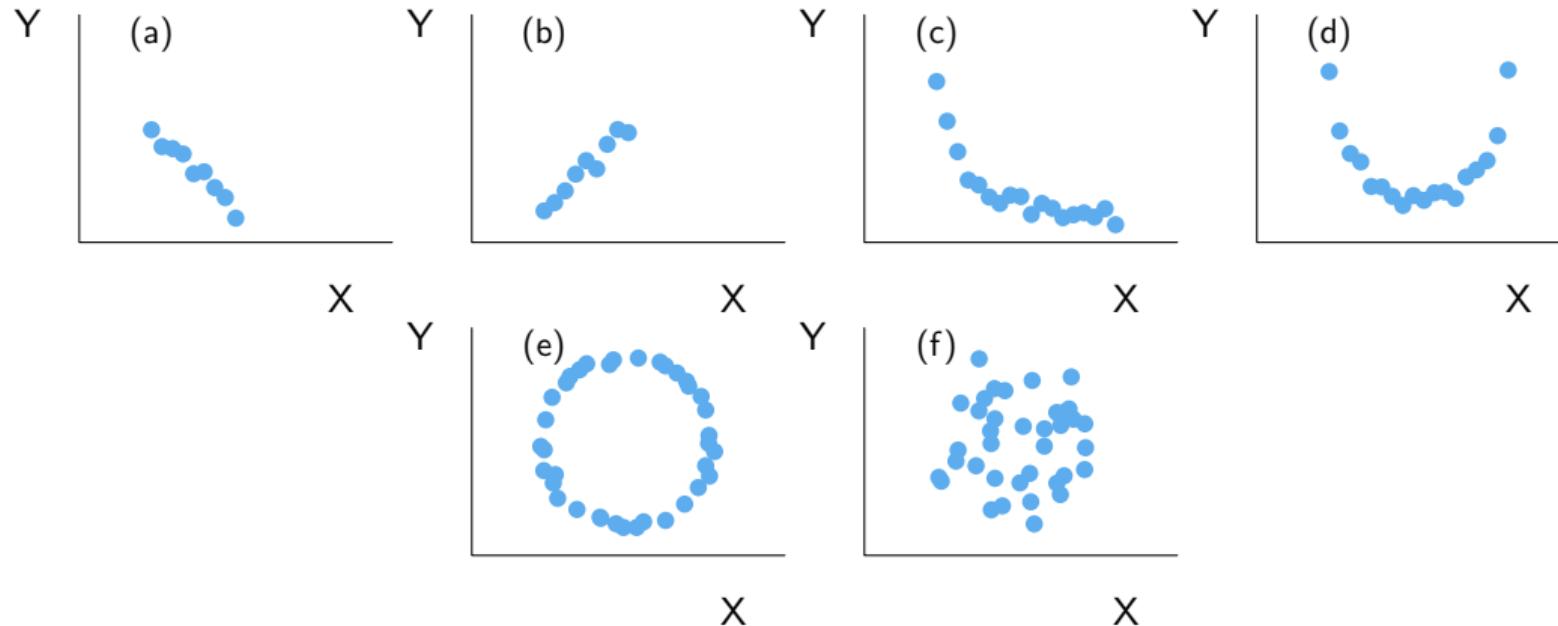
```
## 2 0.364 0.294 0.214
```

```
## 3 0.182 0.294 0.214
```

```
## 4 0.364 0.235 0.500
```

- Wir müssen klären, (1) welche Art des Zusammenhangs zwischen X und Y vorliegt und benötigen (2) Maßzahlen zur Messung seiner Stärke.
- Einen ersten Eindruck vermittelt ein **Streudiagramm** (scatter plot), das entsteht, indem man die Beobachtungen (x_r, y_r) für $r = 1, \dots, n$ als Punkte in ein Koordinatensystem überträgt.
- Die Punkte bilden eine Punktfolge, in der sie mehr oder weniger stark streuen. Typische Zusammenhänge sind auf der nächsten Folie abgebildet.
- Jedes Kreuzchen entspricht einer Beobachtung (x_r, y_r) ; nur aus Platzgründen liegen alle Punktfolgen im ersten Quadranten.

Empirische Formen des Zusammenhangs

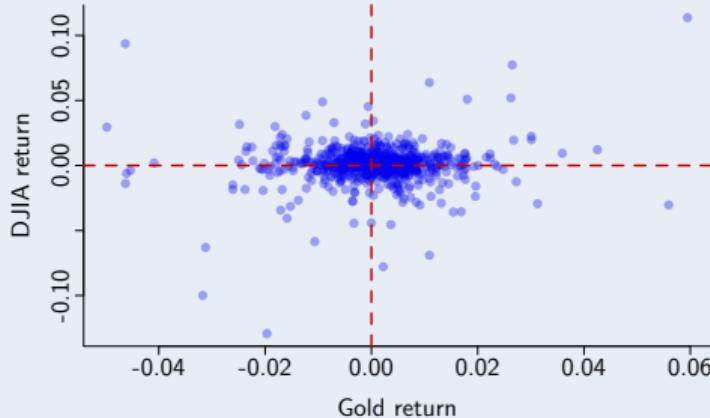
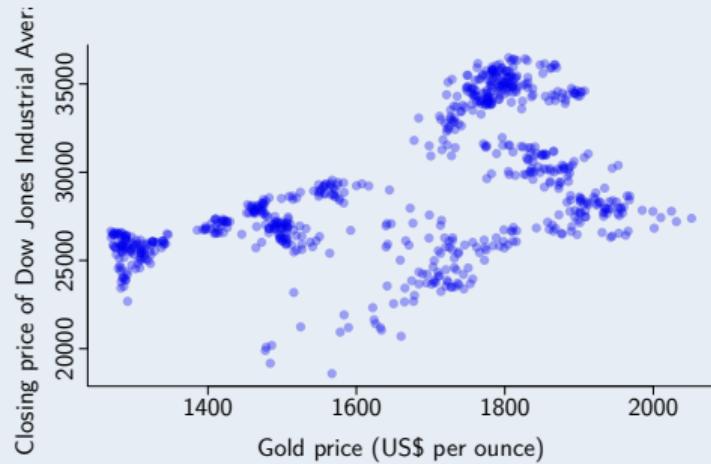


- Bis auf Diagramm f) deuten die übrigen auf einen Zusammenhang hin. In a) und b) scheint der Zusammenhang ca. linear zu sein, in c) und d) ist er hyperbelartig bzw. parabolisch, und in e) kreisförmig.

- Maßzahlen für lineare Zusammenhänge reichen meist aus, da viele nichtlineare Zusammenhänge zwischen X und Y durch eine Transformation linearisiert werden können.
- Lägen z.B. alle Beobachtungen (x_r, y_r) auf einer Hyperbel, gilt: $y_r = 1/x_r$. Nimmt x_r die Werte $\frac{1}{2}, 1, 2, 3$ an, ergibt sich für $y_r : 2, 1, \frac{1}{2}, \frac{1}{3}$. Verwendet man anstelle der x_r die transformierten Werte $x_r^* = \frac{1}{x_r}$, ergibt sich $x_r^* = y_r$: Der Zusammenhang zwischen x_r^* und y_r ist linear.
- Exakte Abhängigkeiten zweier Merkmale stellen die Ausnahme dar.
- Nimmt Y tendenziell mit X zu, spricht man von positiver, im umgekehrten Fall von negativer **Korrelation**.
- Zur Messung der Stärke solcher Zusammenhänge sind Maßzahlen entwickelt worden. Diese hängen von der Skalierung beider Merkmale ab.

Beispiel 7.9: Streudiagramm.

Aktienkurse und Goldpreise 2019-2021



Quelle: Quandl

Beispiel 7.10: Daten visualisieren.

Alles, was wir hier an Statistik machen können, ist natürlich leider nur ein Anfang.

Wenn Sie mal sehen wollen, was man noch so alles mit Daten machen kann, sehen Sie sich das hier mal an:

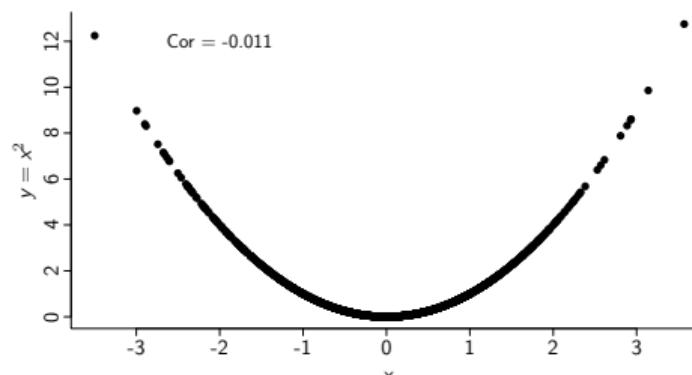
TED

Einiges von dem, was wir hier besprechen – und viel mehr – sehen Sie hier in Aktion:

Gapminder bzw. [direkt in R](#).

- Bei nominal skalierten Merkmalen nutzt man **Assoziationsmaße**, bei allen anderen Skalierungen **Korrelationskoeffizienten**.
- Ein Problem tritt auf, wenn die Merkmale verschiedene Skalen haben. Ordnet man die Skalen „aufsteigend“ nach den zu erfüllenden Anforderungen, ergibt sich: Nominal-, Ordinal- und Kardinalskala.
- Bei unterschiedlichen Skalen für X und Y werten wir i.d.R. bei der Wahl des Maßes die höhere Skala ab.
- Maßzahlen, die für eine bestimmte Skalierung gültig sind, lassen sich auf alle ranghöheren Skalenniveaus anwenden, nicht aber umgekehrt.
- So können z.B. Assoziationsmaße auch für metrische Merkmale berechnet werden, nicht aber Korrelationskoeffizienten für nominal skalierte Variablen.

- Für die Vergleichbarkeit von Resultaten ist es hilfreich, Maße auf das Intervall $[-1, 1]$ zu normieren.
- Da diese Maße den linearen funktionalen Zusammenhang erfassen, nehmen sie den Wert 1 bei positivem und exakt linearem, den Wert -1 bei negativem und exakt linearem Zusammenhang an.
- Je geringer der lineare Zusammenhang in den Daten ausfällt, desto näher liegt die Maßzahl bei null.
- Ein Wert von null bedeutet jedoch nicht, dass zwischen X und Y kein Zusammenhang existiert, sondern nur, dass dieser nicht linear ist.

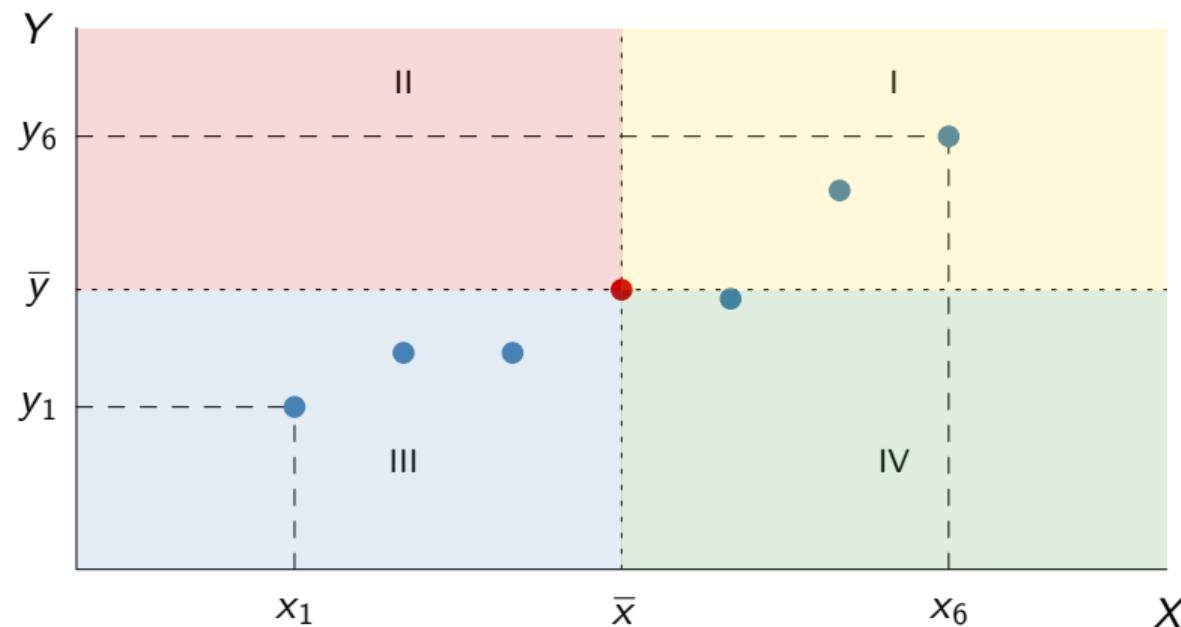


Kovarianz s_{xy}

- Für metrische Merkmale X und Y können Richtung und Stärke ihrer Abhängigkeit über Korrelationen quantifiziert werden.
- Betrachte hierzu zunächst, ob die Beobachtungen (x_r, y_r) , $r = 1, \dots, n$ zusammen variieren. Verwende hierzu die Abweichungen $(x_r - \bar{x})$ und $(y_r - \bar{y})$ jedes Beobachtungspaares (x_r, y_r) zu deren arithmetischen Mitteln.
- Folgendes Streudiagramm von sechs Beobachtungen $(x_1, y_1), \dots, (x_6, y_6)$ verdeutlicht die Vorgehensweise.

Korrelations- und Assoziationsmaße

Kovarianz s_{xy}



Korrelations- und Assoziationsmaße

Kovarianz s_{xy}

- Korrelieren X und Y positiv, liegen die Beobachtungen (x_r, y_r) überwiegend im I. und III. Quadranten. D.h. die Abweichungen $(x_r - \bar{x})$ und $(y_r - \bar{y})$ haben gleiches Vorzeichen; und ihr Produkt ist positiv (bspw. (x_1, y_1) und (x_6, y_6)).
- Bei negativer Korrelation liegen die Beobachtungen meist im II. und IV. Quadranten; die Abweichungen $(x_r - \bar{x})$ und $(y_r - \bar{y})$ haben verschiedene Vorzeichen, ihr Produkt ist daher negativ. Produkte von Abweichungen für Punkte auf den Achsen des Hilfskoordinatensystems sind null.

Korrelations- und Assoziationsmaße

Kovarianz s_{xy}

Definition 7.11: Kovarianz.

Diesen Zusammenhang nutzt die **empirische Kovarianz**, definiert als:

$$s_{xy} = \frac{1}{n} \sum_{r=1}^n (x_r - \bar{x})(y_r - \bar{y}).$$

Gleichen sich positive und negative Abweichungsprodukte aus oder liegen alle Beobachtungspunkte auf den Achsen des Hilfskoordinatensystems, ist $s_{xy} = 0$.

s_{xy} wird umso größer (kleiner), je mehr Wertepaare mit großen x -Werten und großen (kleinen) y -Werten vorliegen.

Korrelations- und Assoziationsmaße

Kovarianz s_{xy}

- Bei positiver (negativer) Kovarianz heißen X und Y **positiv (negativ) korreliert**; nimmt s_{xy} den Wert null an, sind beide Merkmale unkorreliert.
- Liegt eine Kontingenztabelle vor, berechnen wir s_{xy} unter Verwendung der absoluten oder relativen bivariaten Häufigkeiten:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^l (x_i - \bar{x})(y_j - \bar{y}) n_{ij} = \sum_{i=1}^m \sum_{j=1}^l (x_i - \bar{x})(y_j - \bar{y}) h_{ij}.$$

Korrelations- und Assoziationsmaße

Berechnung von Kovarianzen in R

```
n <- 10
x <- 1:10
y <- 1:10
cov(x, y) * (n-1) / n
## [1] 8.25

x <- 1:10
y <- 10:1
cov(x, y) * (n-1) / n
## [1] -8.25

# Kovarianz zwischen PS und Meilen pro Gallone
head(mtcars, n = 3)
##          mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6 160 110 3.90 2.620 16.46  0  1     4     4
## Mazda RX4 Wag 21.0   6 160 110 3.90 2.875 17.02  0  1     4     4
## Datsun 710    22.8   4 108  93 3.85 2.320 18.61  1  1     4     1
```

```
n <- nrow(mtcars)
cov(mtcars$hp, mtcars$mpg) * (n-1) / n
## [1] -310.7092
```

Korrelations- und Assoziationsmaße

Zerlegungsformel der Kovarianz s_{xy}

- Für die Berechnung ist die nun herzuleitende Formel praktischer:

$$\begin{aligned}s_{xy} &= \frac{1}{n} \sum_{r=1}^n (x_r - \bar{x})(y_r - \bar{y}) = \frac{1}{n} \sum_{r=1}^n [x_r(y_r - \bar{y}) - \bar{x}(y_r - \bar{y})] \\ &= \frac{1}{n} \sum_{r=1}^n x_r(y_r - \bar{y}) - \bar{x} \frac{1}{n} \sum_{r=1}^n (y_r - \bar{y}).\end{aligned}$$

- Wegen der Schwerpunkteigenschaft gilt $\sum_{r=1}^n (y_r - \bar{y}) = 0$ und daher:

$$s_{xy} = \frac{1}{n} \sum_{r=1}^n x_r y_r - \bar{y} \frac{1}{n} \sum_{r=1}^n x_r = \frac{1}{n} \sum_{r=1}^n x_r y_r - \bar{x} \bar{y}.$$

- Analog erhält man für eine Kontingenztabelle:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^l x_i y_j n_{ij} - \bar{x} \bar{y} = \sum_{i=1}^m \sum_{j=1}^l x_i y_j h_{ij} - \bar{x} \bar{y}.$$

Zweidimensionale Datensätze 2

Kahoot!

Korrelations- und Assoziationsmaße

Kovarianz s_{xy}

- Sind X und Y (empirisch) unabhängig, gilt $n_{ij} = (n_{i\cdot} \cdot n_{\cdot j})/n$. Einsetzen in die Formel der Kovarianz bei gegebener Kontingenztabelle liefert

$$s_{xy} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^l (x_i - \bar{x})(y_j - \bar{y}) \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} = \frac{1}{n^2} \sum_{i=1}^m (x_i - \bar{x}) n_{i\cdot} \sum_{j=1}^l (y_j - \bar{y}) n_{\cdot j} = 0,$$

weil die beiden Summen der letzten Gleichung wegen der Schwerpunkteigenschaft null sind.

Korrelations- und Assoziationsmaße

Kovarianz s_{xy}

- Diese Eigenschaft der Kovarianz bei **Unabhängigkeit** ist wichtig bei der Berechnung der Varianz einer Variablen Z , wenn $Z = X + Y$, z.B. aus der Addition des Konsums X und der Nettoinvestitionen Y zum Nettoinlandsprodukt Z .
- Die empirische Varianz der Variablen $Z = X + Y$ erhält man als:

$$\begin{aligned}s_z^2 &= \frac{1}{n} \sum_{r=1}^n (z_r - \bar{z})^2 = \frac{1}{n} \sum_{r=1}^n [(x_r - \bar{x}) + (y_r - \bar{y})]^2 \\&= \frac{1}{n} \sum_{r=1}^n (x_r - \bar{x})^2 + \frac{2}{n} \sum_{r=1}^n (x_r - \bar{x})(y_r - \bar{y}) + \frac{1}{n} \sum_{r=1}^n (y_r - \bar{y})^2 \\&= s_x^2 + s_y^2 + 2s_{xy}\end{aligned}$$

- Dies ist der **Additionssatz für Varianzen abhängiger Variablen**. Bei Unabhängigkeit reduziert er sich wegen $s_{xy} = 0$ zu $s_z^2 = s_x^2 + s_y^2$.

Korrelations- und Assoziationsmaße

Korrelationskoeffizient nach Bravais-Pearson r_{xy}

- Dividiert man s_{xy} durch $s_x s_y$, erhält man den **Korrelationskoeffizienten** r_{xy} (nach Bravais-Pearson).

Definition 7.12: Korrelationskoeffizient.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{r=1}^n (x_r - \bar{x})(y_r - \bar{y})}{\sqrt{\sum_{r=1}^n (x_r - \bar{x})^2} \sqrt{\sum_{r=1}^n (y_r - \bar{y})^2}}, \quad -1 \leq r_{xy} \leq 1.$$

Das Vorzeichen des Korrelationskoeffizienten gibt die Richtung, sein Betrag die Stärke des Zusammenhangs an.

- Bei Unabhängigkeit folgt wegen $s_{xy} = 0$ immer auch $r_{xy} = 0$. Jedoch impliziert umgekehrt $r_{xy} = 0$ bzw. $s_{xy} = 0$ nicht automatisch Unabhängigkeit (vgl. 7-41).
- Liegen alle Punkte (x_r, y_r) auf einer Geraden mit positiver (negativer) Steigung, so ist $r_{xy} = 1$ (-1).

Korrelations- und Assoziationsmaße

Korrelationskoeffizient nach Bravais-Pearson r_{xy}

- Es existieren noch weitere Formeln zur Berechnung von r_{xy} . Bei Einzelbeobachtungen ist es günstig, die Varianzen und die Kovarianz nach dem speziellen Verschiebungssatz zu ermitteln. Man erhält

$$r_{xy} = \frac{\sum_{r=1}^n x_r y_r - n\bar{x}\bar{y}}{\sqrt{(\sum_{r=1}^n x_r^2 - n\bar{x}^2)(\sum_{r=1}^n y_r^2 - n\bar{y}^2)}}.$$

- Liegen die Daten in einer Kontingenztabelle, nutzen wir

$$r_{xy} = \frac{\sum_{i=1}^m \sum_{j=1}^l x_i y_j n_{ij} - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^m x_i^2 n_{i\cdot} - n\bar{x}^2)(\sum_{j=1}^l y_j^2 n_{\cdot j} - n\bar{y}^2)}}.$$

Beispiel 7.13: Berechnung von r_{xy} .

Eine Arbeitstabelle ist hilfreich für die Berechnung von r_{xy} . Spalten (2) und (3) der folgenden Tabelle geben die Anzahl eingeschriebener Studierender X (in Tsd) und die durchschnittliche Studiendauer Y (in Semester) an zehn unterschiedlichen Fachbereichen wieder.

Da beide Merkmale metrisch skaliert sind, kann der Zusammenhang mit r_{xy} beschrieben werden.

Korrelations- und Assoziationsmaße

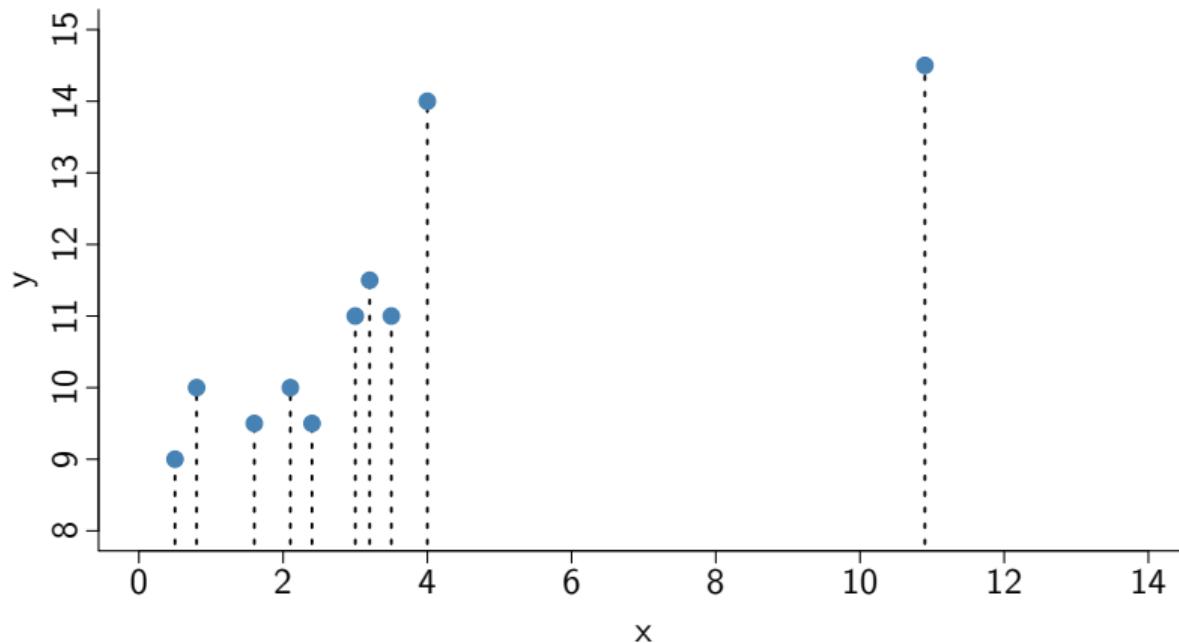
Korrelationskoeffizient nach Bravais-Pearson r_{xy}

Beispiel 7.13: Fortsetzung.

r (1)	x_r (2)	y_r (3)	$x_r - \bar{x}$ (4)	$y_r - \bar{y}$ (5)	$(4) \cdot (5)$ (6)	$(x_r - \bar{x})^2$ (7)	$(y_r - \bar{y})^2$ (8)
1	3,5	11,0	0,3	0	0,0	0,09	0,00
2	2,4	9,5	-0,8	-1,5	1,2	0,64	2,25
3	3,0	11,0	-0,2	0	0,0	0,04	0,00
4	2,1	10,0	-1,1	-1,0	1,1	1,21	1,00
5	1,6	9,5	-1,6	-1,5	2,4	2,56	2,25
6	3,2	11,5	0	0,5	0,0	0,00	0,25
7	0,8	10,0	-2,4	-1,0	2,4	5,76	1,00
8	4,0	14,0	0,8	3,0	2,4	0,64	9,00
9	0,5	9,0	-2,7	-2,0	5,4	7,29	4,00
10	10,9	14,5	7,7	2,5	26,95	59,29	12,25
\sum	32	110,0	0	0	41,85	77,52	32,00
$\bar{x} =$	3,2	$\bar{y} = 11$					

Korrelations- und Assoziationsmaße

Streudiagramm



Korrelations- und Assoziationsmaße

Korrelationskoeffizient nach Bravais-Pearson r_{xy}

Beispiel 7.13: Fortsetzung.

Auch ohne Berücksichtigung des Ausreißers (x_{10}, y_{10}) zeigt das Streudiagramm positive Korrelation zwischen Studierendenzahl und durchschnittlicher Studiendauer.

Mit den Summen aus den Spalten (6), (7) und (8) ergibt sich

$$r_{xy} = \frac{41,85}{\sqrt{77,52}\sqrt{32}} = 0,8403.$$

Der hohe Wert für r_{xy} weist auf einen stark ausgeprägten linearen Zusammenhang zwischen X und Y hin. Ohne zusätzliche Information sollte eine weitere - insbesondere kausale - Interpretation der Korrelation unterbleiben.

Eine kausale Interpretation der Abhängigkeit ist nur mit einer Theorie über die Beziehung der Variablen möglich. Ohne Begründung wird u.U. eine „Nonsense-Korrelation“ ermittelt. [Beispiele](#)

Korrelations- und Assoziationsmaße

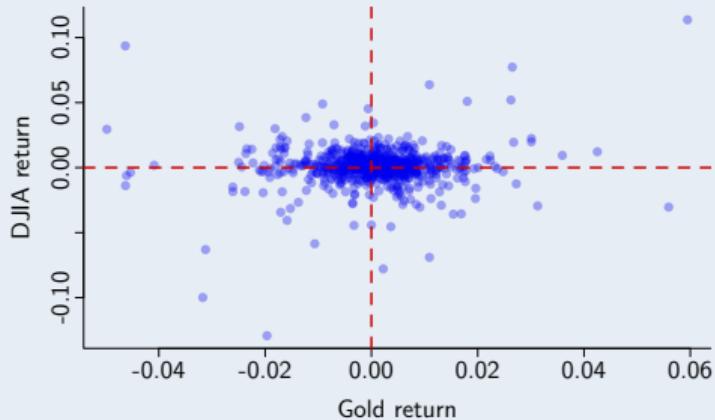
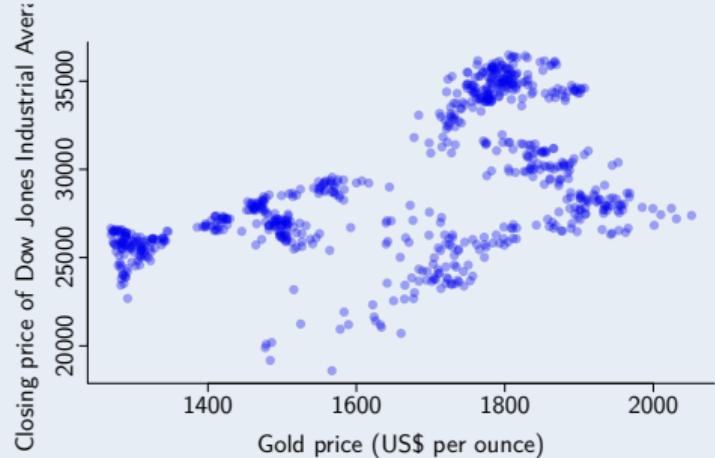
Korrelationskoeffizient nach Bravais-Pearson r_{xy} - in 

```
# Die Daten von Folie 7-56
x <- c(3.5, 2.4, 3, 2.1, 1.6, 3.2, 0.8, 4, 0.5, 10.9)
y <- c(11, 9.5, 11, 10, 9.5, 11.5, 10, 14, 9, 14.5)
cor(x,y)
## [1] 0.8402598
```

Korrelations- und Assoziationsmaße

Korrelationskoeffizient nach Bravais-Pearson r_{xy}

Beispiel 7.14: Aktienkurse und Goldpreise 2019-2021.



$$s_{xy} = 446948.363 \quad r_{xy} = 0.5371$$

$$s_{xy} \approx 0 \quad r_{xy} = 0.0806$$

Quelle: Yahoo Finance

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

- Für ordinal skalierte Merkmale sollte r_{xy} nicht berechnet werden.
- Bei solchen Merkmalen sind nur die Ränge, nicht aber der Abstände der Ausprägungen relevant. Daher weist man x_r und y_r aufsteigende Ränge zu, eins für den kleinsten und n für den größten Wert: x_r erhält Rang i , falls x_r an i -ter Stelle der Rangordnung liegt, $R(x_r) = i$, analog für y_r .
- Auch abnehmende Ränge sind möglich. Wichtig ist nur, dass beide Merkmale gleich behandelt werden.
- Kommen Werte mehrmals vor, spricht man von Bindung. Vergebe dann mittlere Ränge. Sind z.B. die Ränge 5 und 6 mit denselben Werten besetzt, erhalten beide die Rangnummer 5,5.
- Dies liefert Rangpaare $[R(x_1), R(y_1)], \dots, [R(x_n), R(y_n)]$.

Pay for performance?

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

Definition 7.15: Rangkorrelationskoeffizient.

Für $R(x_r)$ kann man den Korrelationskoeffizient nach Bravais-Pearson berechnen, der jetzt **Rangkorrelationskoeffizient** nach Spearman heißt:

$$r_s = \frac{\sum_{r=1}^n [R(x_r) - \bar{R}(x)][R(y_r) - \bar{R}(y)]}{\sqrt{\sum_{r=1}^n [R(x_r) - \bar{R}(x)]^2 \sum_{r=1}^n [R(y_r) - \bar{R}(y)]^2}}$$

wobei $\bar{R}(\cdot)$ das arithmetische Mittel der Ränge von X bzw. Y ist. Liegen keine Bindungen vor, vereinfacht sich die Berechnung von r_s durch Ausnutzung bestimmter Summeneigenschaften der natürlichen Zahlen, mit $d_r = R(x_r) - R(y_r)$ (siehe Buch, S. 171-172):

$$r_s = 1 - \frac{6 \sum_{r=1}^n d_r^2}{n(n^2 - 1)},$$

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

- r_s ist also bereits dann eins, wenn die $R(x_r) = R(y_r)$ für alle r , also x_r und y_r streng monoton steigen. Damit gilt nicht nur bei positivem linearem Zusammenhang $r_s = 1$.
- Wegen seiner einfachen Berechnung wird die Formel für r_s ohne Bindung oft auch bei Bindung angewendet. Der sich einstellende Fehler ist bei wenigen Bindungen vernachlässigbar. Liegen viele Bindungen vor, sollte r_s mit der ursprünglichen Gleichung berechnet werden.
- Die Formeln für r_s können analog zur Vorgehensweise bei r_{xy} der speziellen Datenlagen angepasst werden. Gilt in einer Kontingenztabelle $n_{ij} > 1$, liegen Bindungen vor.
- Das folgende Beispiel zeigt, wie die Kontingenztabelle in eine Arbeitstabelle überführt werden kann.

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

Beispiel 7.16: Befragung zu Lernintensität und erzielter Note in einer Statistikklausur.

Vgl. die Kontingenztabelle auf der nächsten Folie.

Die Note (X) hat die Ausprägungen 1 bis 5, die Lernintensität (Y) die Ausprägungen „intensiv“, „durchschnittlich“ und „schwach“.

Da „sehr gut“ zwar die kleinste Ausprägung, jedoch die größte Leistung darstellt, werden X und Y abnehmend geordnet.

Zwei Studierende erreichten eine 1; sie erhalten jeweils Rang 1,5. $n_2 = 4$ Studierende erreichten die Note 2; sie haben die Platzierungen 3 bis 6 und erhalten den Rang 4,5; entsprechend für die übrigen Ränge. Die Ränge für Y stehen in der ersten Zeile.

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

Beispiel 7.16: Fortsetzung.

$R(x_r)$	$R(y_r)$ X/Y	3 intensiv	10,5 durchschnittl.	18 schwach	n_i	$\sum d_{ij}^2 n_{ij}$
1,5	1	2 (4,5)			2	4,5
4,5	2	2 (4,5)	2 (72)		4	76,5
10	3	1 (49)	5 (1,25)	1 (64)	7	114,25
15,5	4		2 (50)	2 (12,5)	4	62,5
19	5		1 (72,25)	2 (2)	3	74,25
	n_j	5	10	5	20	332

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

Beispiel 7.16: Fortsetzung.

Die Differenz $d_1 = R(x_1) - R(y_1)$ für $r = 1$ liefert $d_1 = 1,5 - 3 = -1,5$; Quadrieren und Multiplizieren mit $n_{11} = 2$ ergibt 4,5. Diese Werte stehen in Klammern neben den gemeinsamen Häufigkeiten; ihre Zeilensumme steht in der letzten Spalte. Obwohl Bindung vorliegt, wird r_s zunächst mit der vereinfachten Gleichung berechnet:

$$r_s = 1 - \frac{6 \cdot 332}{20 \cdot 399} = 0,7504.$$

Die bei Bindung angemessene, aber aufwändiger Formel ergibt

$$r_s = 0,7213.$$

Beide Ergebnisse zeigen, dass zwischen Lernintensität und Klausurleistung eine große positive Korrelation besteht.

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

```
Intensitaet <- rep(c("intensiv", "Durchschnitt", "schwach"),
                    c(5, 10, 5))
n <- length(Intensitaet)
Intensitaet.ordered <- factor(Intensitaet, ordered = TRUE,
                                levels = c("intensiv", "Durchschnitt", "schwach"))

Noten <- c("sehr gut", "sehr gut", "gut", "gut", "befriedigend", "gut", "gut",
          "befriedigend", "befriedigend", "befriedigend", "befriedigend",
          "befriedigend", "ausreichend", "ausreichend", "mangelhaft",
          "befriedigend", "ausreichend", "ausreichend",
          "mangelhaft", "mangelhaft")

Noten.ordered <- factor(Noten, ordered = TRUE,
                        levels = c("sehr gut", "gut", "befriedigend",
                                  "ausreichend", "mangelhaft"))

head(data.frame(Intensitaet.ordered, Noten.ordered), 4)

##   Intensitaet.ordered Noten.ordered
## 1      intensiv       sehr gut
## 2      intensiv       sehr gut
## 3      intensiv           gut
## 4      intensiv           gut
```

Korrelations- und Assoziationsmaße

Rangkorrelationskoeffizient von Spearman r_s

```
ranks.Intensitaet <- rank(Intensitaet.ordered)
ranks.Noten <- rank(Noten.ordered)
cor(ranks.Intensitaet, ranks.Noten, method = "spearman")
## [1] 0.7212879

RangDifferenz <- ranks.Intensitaet - ranks.Noten
(Spearman vereinfacht <- 1 - 6 * sum(RangDifferenz^2)/(n*(n^2-1)))
## [1] 0.7503759
```

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

- Auch für nominal skalierte Merkmale existieren Zusammenhangsmaße. Häufig verwendet wird der **Kontingenzkoeffizient von Pearson**.
- Aus den Randverteilungen einer Kontingenztabelle errechnet sich die theoretische Feldbesetzung \tilde{n}_{ij} bei Unabhängigkeit von X und Y als

$$\tilde{n}_{ij} = (n_{i\cdot} \cdot n_{\cdot j}) / n.$$

Die Stärke des Zusammenhangs zwischen X und Y wird quantifiziert über die relativen Differenzen zwischen den empirischen, absoluten bivariaten Häufigkeiten n_{ij} und den theoretischen Werten \tilde{n}_{ij} bei Unabhängigkeit:

$$(n_{ij} - \tilde{n}_{ij}) / \tilde{n}_{ij}.$$

Je größer der Unterschied zwischen n_{ij} und \tilde{n}_{ij} , desto stärker der Zusammenhang zwischen X und Y .

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

Definition 7.17: quadratische Kontingenz.

Damit sich positive und negative relative Differenzen nicht kompensieren, werden die $(n_{ij} - \tilde{n}_{ij})$ zunächst quadriert. Die so gebildete Summe heißt quadratische Kontingenz χ^2 (gelesen: chi-quadrat):

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^l \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = n \sum_{i=1}^m \sum_{j=1}^l \frac{(n_{ij} - \tilde{n}_{ij})^2}{n_{i \cdot} \cdot n_{\cdot j}}.$$

Sind X und Y unabhängig, gilt $n_{ij} = \tilde{n}_{ij}$ und daher: $\chi^2 = 0$. Bei Abhängigkeit ist $\chi^2 > 0$. Die letzte Umformung zeigt, dass χ^2 mit n unbegrenzt wächst. Division von χ^2 durch n , χ^2/n , eliminiert dies („mittlere quadratische Kontingenz“).

Kontingenzkoeffizient von Pearson K

- Den maximalen Wert erreicht χ^2 bei gegebenem n wenn bei Kontigenztabellen mit
 - (a) $m = l$ zu jeder Ausprägung von X genau eine Ausprägung von Y gehört,
 - (b) $m < l$ in jeder Spalte l genau eine bivariate Häufigkeit ungleich null ist, d.h. zu jeder Ausprägung von X gehört mindestens eine Ausprägung von Y (aber zu jeder Ausprägung von Y gehört genau eine Ausprägung von X),
 - (c) $m > l$ die Zuordnung der bivariaten Häufigkeiten genau umgekehrt zu (b) ist.
- Es liegt dann die größte Abhängigkeit zwischen X und Y vor und das Maximum von χ^2 ist

$$\chi^2_{\max} = n \min(l - 1, m - 1).$$

Korrelations- und Assoziationsmaße

Maximum von χ^2

O.B.d.A. betrachten wir Fall (b). Das bedeutet, dass $n_{ij} = n_{\cdot j}$ für ein i und 0 sonst, da nur eine Zelle pro Spalte besetzt ist. Nenne dieses i , wegen seiner Abhängigkeit von der Spalte j , i_j^* , so dass $n_{ij} = n_{i_j^* j}$ wenn $n_{ij} \neq 0$. Schreibe

$$\begin{aligned}\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} &= \left(\frac{n_{ij} - \tilde{n}_{ij}}{\tilde{n}_{ij}} \right)^2 \tilde{n}_{ij} \\ &= \left(\frac{n_{ij}}{\tilde{n}_{ij}} - 1 \right)^2 \tilde{n}_{ij} \\ &= \left(\frac{n_{ij}^2}{\tilde{n}_{ij}^2} - 2 \frac{n_{ij}}{\tilde{n}_{ij}} + 1 \right) \tilde{n}_{ij} \\ &= \frac{n_{ij}^2}{\tilde{n}_{ij}} - 2n_{ij} + \tilde{n}_{ij} \\ &= \frac{n_{ij}^2}{\frac{n_{i \cdot} \cdot n_{\cdot j}}{n}} - 2n_{ij} + \frac{n_{i \cdot} \cdot n_{\cdot j}}{n}\end{aligned}$$

Also ist

$$\begin{aligned}\chi^2 &= \sum_{i=1}^m \sum_{j=1}^l \left(\frac{n_{ij}^2}{\frac{n_{i \cdot} \cdot n_{\cdot j}}{n}} - 2n_{ij} + \frac{n_{i \cdot} \cdot n_{\cdot j}}{n} \right) \\ &= n \sum_{i=1}^m \sum_{j=1}^l \frac{n_{ij}^2}{n_{i \cdot} \cdot n_{\cdot j}} - 2 \sum_{i=1}^m \sum_{j=1}^l n_{ij} + \sum_{i=1}^m \sum_{j=1}^l \frac{n_{i \cdot} \cdot n_{\cdot j}}{n} \\ &= n \sum_{i=1}^m \sum_{j=1}^l \frac{n_{ij}^2}{n_{i \cdot} \cdot n_{\cdot j}} - 2n + \frac{n^2}{n} \\ &= n \sum_{i=1}^m \sum_{j=1}^l \frac{n_{ij}^2}{n_{i \cdot} \cdot n_{\cdot j}} - n\end{aligned}$$

Korrelations- und Assoziationsmaße

Maximum von χ^2

Wegen $n_{i^*j} = n_{\cdot j}$ folgt

$$\begin{aligned}\chi^2 &= n \sum_{i=1}^m \sum_{j=1}^l \frac{n_{i^*j}^2}{n_{i\cdot} \cdot n_{j^*j}} - n \\ &= n \sum_{i=1}^m \sum_{j=1}^l \frac{n_{i^*j}}{n_{i\cdot}} - n \\ &= n \sum_{i=1}^m \frac{1}{n_{i\cdot}} \sum_{j=1}^l n_{i^*j} - n \\ &= n \sum_{i=1}^m \frac{1}{n_{i\cdot}} n_{i\cdot} - n = n \sum_{i=1}^m 1 - n \\ &= n \cdot m - n = n(m-1)\end{aligned}$$

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

- Die maximale quadratische Kontingenz hängt somit ab von der Zeilenzahl m oder der Spaltenzahl l und n .

Definition 7.18: Kontingenzkoeffizient.

Man normiert daher χ^2 durch Division mit $(\chi^2 + n)$. Die Wurzel dieses Quotienten liefert den Kontingenzkoeffizienten K von Pearson:

$$K = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

K ist bei Unabhängigkeit wie gewünscht null; mit zunehmender Abhängigkeit strebt K gegen eins.

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

Sein maximaler Wert K_{\max} ist durch χ^2_{\max} determiniert, hängt aber wegen der Normierung nicht mehr von n ab:

$$\begin{aligned} K_{\max} &= \sqrt{\frac{\chi^2_{\max}}{\chi^2_{\max} + n}} \\ &= \sqrt{\frac{n \min(l-1, m-1)}{n \min(l-1, m-1) + n}} \\ &= \sqrt{\frac{\min(l-1, m-1)}{\min(l-1, m-1) + 1}} \\ &= \sqrt{\frac{\lambda - 1}{\lambda}} < 1, \end{aligned}$$

mit $\lambda := \min(l, m)$, da $\min(l-1, m-1) + 1 = \min(l, m)$.

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

- Soll der Kontingenzkoeffizient bei größter Abhängigkeit gleich eins sein, ist K einfach durch K_{\max} zu dividieren. Man bezeichnet diesen Quotienten als **korrigierten Kontingenzkoeffizienten**

$$K^* = \frac{K}{K_{\max}}.$$

- Die Berechnung von K^* besteht aus drei Arbeitsschritten:
 - Berechnung des χ^2 -Wertes,
 - Berechnung von K und K_{\max} ,
 - Berechnung des Quotienten

$$K^* = \frac{K}{K_{\max}}.$$

Beispiel 7.19: Berechnung des Kontingenzkoeffizienten.

Die Arbeitsschritte werden verdeutlicht mit Daten aus einer Befragung von 100 Frauen und 100 Männern über ihr Haustier. Das nominale Merkmal X ist Geschlecht, das nominale Merkmal Y ist Haustier.

X/Y	$y_1 = \text{Hund}$	$y_2 = \text{Katze}$	$n_i.$
$x_1 = \text{weiblich}$	7	93	100
$x_2 = \text{männlich}$	28	72	100
n_j	35	165	200

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

Beispiel 7.19: Fortsetzung.

Für den ersten Schritt benötigt man die theoretischen bivariaten Häufigkeiten \tilde{n}_{ij} :

$\tilde{n}_{11} = \tilde{n}_{21} = 17,5 (= 100 \cdot 35/200)$ und $\tilde{n}_{12} = \tilde{n}_{22} = 82,5$. Damit ist

$$\chi^2 = \frac{(7 - 17,5)^2}{17,5} + \frac{(93 - 82,5)^2}{82,5} + \frac{(28 - 17,5)^2}{17,5} + \frac{(72 - 82,5)^2}{82,5} = 15,27.$$

Der Kontingenzkoeffizient beträgt: $K = \sqrt{\frac{15,27}{215,27}} = 0,2664$;

sein Maximum ist, mit $\lambda = 2$, $K_{\max} = 1/\sqrt{2} = 0,7071$.

Damit ist $K^* = \frac{0,2664}{0,7071} = 0,3768$.

Dieser Wert zeigt, dass zwischen Geschlecht und Haustier ein Zusammenhang besteht, der jedoch nicht „sehr“ ausgeprägt ist. Beachte bei der Interpretation, dass K bzw. K^* nur Werte in $[0, 1]$ annimmt.

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

Beispiel 7.20: Digital distractions in the classroom.

How much of a distraction is caused by other students who use digital devices during class for non-class activities?

X/Y	No distraction	A little distraction	More than a little distraction	Big distraction	Very big distraction	n_i
Female	146	267	43	17	9	482
Male	105	131	34	7	2	279
n_j	251	398	77	24	11	761

Quelle: Journal of Media Education, Vol. 4(4), October 2013

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K

Beispiel 7.20: Fortsetzung.

\tilde{n}_{ij} :

X/Y	No distraction	A little distraction	More than a little distraction	Big distraction	Very big distraction
Female	158,9777	252,0841	48,77	15,2011	6,9671
Male	92,0223	145,9159	28,23	8,7989	4,0329

$$\chi^2 = 9,3575, K = \sqrt{\frac{9,3575}{770,3575}} = 0,1102, \lambda = 2, K_{\max} = 1/\sqrt{2} = 0,7071$$

$$K^* = \frac{0,1102}{0,7071} = 0,1559$$

Korrelations- und Assoziationsmaße

Kontingenzkoeffizient von Pearson K - in

```
library(DescTools)
# Daten von 7-77 replizieren
df <- data.frame(Geschlecht = c(rep("Mann", 100), rep("Frau", 100)),
                  Haustier    = c(rep("Hund", 28), rep("Katze", 72),
                                 rep("Hund", 7), rep("Katze", 93)))

# Häufigkeitstable erstellen
(hfg_tab <- table(df))

##             Haustier
## Geschlecht Hund Katze
##     Frau      7    93
##     Mann     28    72

# Kontingenzkoeffizient
ContCoef(hfg_tab)

## [1] 0.2663568

# korrigierter Kontingenzkoeffizient
ContCoef(hfg_tab, correct = T)

## [1] 0.3766854
```

Beispiel 7.21: Das Simpson-Paradox.

Bei der Interpretation von Kontingenztabellen muss man sich aber immer vor dem Simpson-Paradox hüten. Rauchen rettet nämlich nicht ihr Leben: siehe beispielsweise [hier](#).

Korrelations- und Assoziationsmaße

Das Simpson-Paradox

Beispiel 7.21: Fortsetzung.

		Smoker?	
		Yes	No
Dead	107	153	
Alive	174	175	
Total	281	328	
% Dying	38,10%	46,60%	

		Age Group					
		45-54		55-64		65-74	
		Smoker?		Smoker?		Smoker?	
		Yes	No	Yes	No	Yes	No
Dead	27	12	51	40	29	101	
Alive	103	66	64	81	7	28	
Total	130	78	115	121	36	129	
% Dying	20,80%	15,40%	44,35%	33,10%	80,60%	78,30%	

Quelle: http://www.stat.osu.edu/~biostat/newsletters/volume2_2/article_vol2_2.html (inaktiv)

- Bei gemischten Skalen ist immer die niedrigere Skala maßgebend.
- Diese entscheidet darüber, welches Zusammenhangsmaß zur Anwendung kommt.

X/Y	metrisch	ordinal	nominal
metrisch	K^*	K^*	K^*
	r_s	r_s	
	s_{xy}, r_{xy}		
ordinal	K^*, r_s	K^*, r_s	K^*
nominal	K^*	K^*	K^*

Zweidimensionale Datensätze 3

Kahoot!

- Dieses Kapitel führte in die Analyse des Zusammenhangs zweier Variablen ein. Das Verständnis der Zusammenhänge ist nicht nur wichtig für die deskriptive Statistik, sondern auch Grundlage für Veranstaltungen wie der Ökonometrie.
- Nachbereitung: Kapitel 5.1 und 5.2 des Buches von Prof. Assenmacher.
- Das nächste Kapitel behandelt elementare Regressionsrechnung.
- Vorbereitung: Kapitel 5.3 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 **Regressionsrechnung**
- 9 Elementare Zeitreihenanalyse

Regressionsfunktion

- Die Korrelationsanalyse beschreibt nur Stärke und Richtung eines Zusammenhangs zwischen X und Y , nicht jedoch die **kausale Abhängigkeit**. Dies ermöglicht keine Schlüsse auf die Kausalstruktur: Es bleibt offen, welche der Variablen Ursache und welche Wirkung ist oder ob beide „interdependent“ sind.
- Ziel der Empirie ist es aber oft, Kausalstrukturen abzuleiten, indem einer beobachtbaren Wirkung Ursachen zugeordnet werden. Y ist das Merkmal, das erklärt werden soll und deshalb die Wirkung darstellt und heißt zu erklärende, abhängige oder **endogene Variable**.
- Die als Ursachen aufgefassten Merkmale nennt man erklärende, unabhängige oder **exogene Variablen** X_k , $k = 1, \dots, K$, wobei hier k Ursachen unterscheidet.

- Mit Regressionen ist es u.U. möglich, zu überzeugenden Kausalaussagen zu kommen.
- Diese Umstände können wir allerdings noch nicht betrachten und nutzen die Regression erst mal als flexibles deskriptives Werkzeug.

Beispiel 8.1: Stilettos.

Hier ist ein Beispiel, das belegt, dass auch Wissenschaftler oft Mühe haben, zwischen Korrelation und Kausalität zu trennen: [Stilettos](#)

Beispiel 8.2: Zusammenhang zwischen Geldmengenwachstum und Inflation.

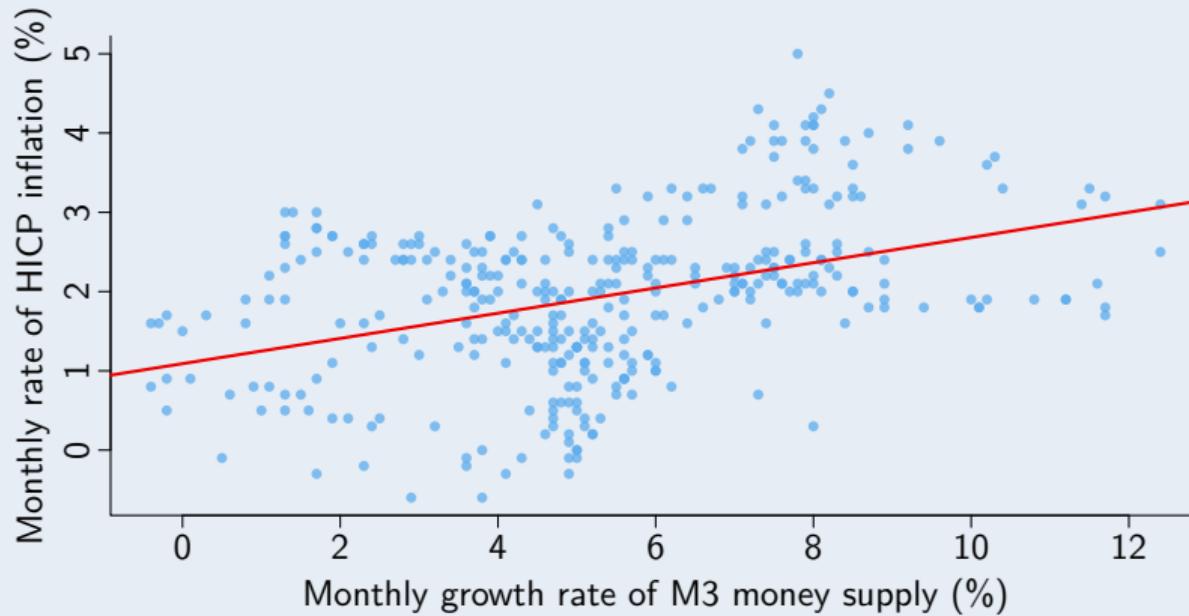


Abbildung 1: Geldmenge (M3) der Eurozone 01/1991-04/2020, (vor 2002 wurden die einzelnen nationalen Geldmengen konvertiert und aggregiert), ECB Statistical Data Warehouse

Regressionsfunktion

- Formalisiere den substanzwissenschaftlich begründeten Zusammenhang durch die Funktion $Y = f(X_1, \dots, X_K)$.
- Obwohl alle in der Funktion enthaltenen Variablen beobachtbar sind, ist sie für eine statistische Analyse noch zu allgemein. Wenn f eine lineare Funktion ist, erhält man für obige Gleichung die Linearspezifikation

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_K X_K.$$

- Die Parameter α_k , $k = 1, \dots, K$ sind unbekannt und müssen aus gegebenen Beobachtungen ermittelt werden. Die Beobachtungen lassen sich in einer $n \times (K + 1)$ -dimensionalen **Beobachtungsmatrix** darstellen.

Beobachtungsmatrix:

Y	X_1	X_2	\dots	X_K	
y_1	x_{11}	x_{12}		x_{1K}	← Beobachtungstupel
y_2	x_{21}	x_{22}		x_{2K}	
\vdots	\vdots	\vdots		\vdots	
\vdots	\vdots	\vdots		\vdots	
\vdots	\vdots	\vdots		\vdots	
y_n	x_{n1}	x_{n2}		x_{nK}	

Regressionsfunktion

- Da jede Theorie von vielen Ursachen abstrahiert, wird die Linearspezifikation nicht für jedes Beobachtungstupel $(y_r, x_{r1}, \dots, x_{rK})$, $r = 1, \dots, n$ exakt erfüllt sein.
- Um eine Gleichung für jedes r zu erreichen, wird obige Gleichung mit einer nicht beobachtbaren, **latenten Variablen** U erweitert:

$$y_r = \alpha_1 x_{r1} + \alpha_2 x_{r2} + \dots + \alpha_K x_{rK} + u_r, \quad r = 1, \dots, n.$$

- Diese Gleichung heißt multiple oder multivariate lineare Regressionsfunktion bzw. -gleichung. **Linearität** bedeutet hier, dass sowohl Variablen als auch Parameter linear in die Funktion eingehen.
- Sollte f eine nichtlineare Funktion sein, lässt sich die Ursprungsgleichung u.U. dennoch in eine lineare Regressionsfunktion überführen.

Regressionsfunktion

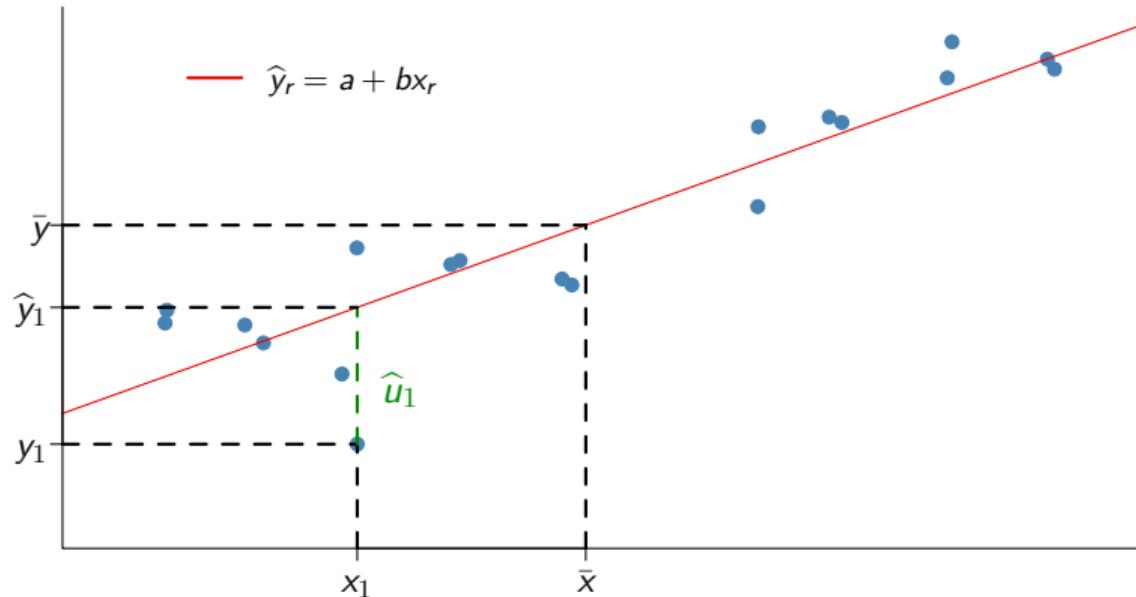
- Die Regressionsanalyse berechnet aus den Beobachtungen die unbekannten Parameter α_k .
- Die endogene Variable Y wird auch als **Ressand**, die exogenen Variablen X_k , $k = 1, \dots, K$ als **Ressoren** und die Parameter α_k als **Regressionskoeffizienten** bezeichnet.
- Soll die Regressionsfunktion einen Achsenabschnitt enthalten, also **inhomogen** sein, ist (meistens) X_1 für alle r gleich eins: $x_{11} = \dots = x_{n1} = 1$. Die ganze zweite Spalte der Beobachtungsmatrix hat dann den Wert eins, und α_1 ist der Achsenabschnitt.

Regressionsfunktion

- Die aus den Daten ermittelten Regressionskoeffizienten heißen **Regressionskoeffizientenschätzungen** (oder kurz: Schätzungen).
- Die Schätzung der Koeffizienten einer multiplen Regression ist aufwändiger. Sie wird später im Studium in der Ökonometrie weiter diskutiert.
- Die Situation vereinfacht sich, wenn eine univariate bzw. einfache lineare Regressionsfunktion vorliegt:
 $x_{r1} \equiv 1$, $r = 1, \dots, n$ und $\alpha_k = 0$ für $3 \leq k \leq K$.
- Die **einfache, inhomogene lineare Regressionsfunktion** lautet: $y_r = \alpha_1 + \alpha_2 x_{r2} + u_r$, oder mit anderer Notation

$$y_r = \alpha + \beta x_r + u_r, \quad r = 1, \dots, n.$$

- Übertragen der Zahlenpaare (x_r, y_r) in ein Koordinatensystem liefert ein **Streudiagramm**. Es soll eine Gerade $\hat{y} = a + bx$ so an die Punktfolge angepasst werden, dass sie den Zusammenhang möglichst „gut“ erfasst.



- Das zu jedem x_r gehörende \hat{y}_r heißt **Regresswert** oder „gefitteter“ Wert. Die vertikalen Abstände \hat{u}_r ergeben sich als $\hat{u}_r = y_r - \hat{y}_r$. Wähle aus der unendlichen Anzahl aller möglichen Geraden die „beste“ aus.
- Als Maß für die Abweichung nimmt man die Summe der Abstandsquadrate:

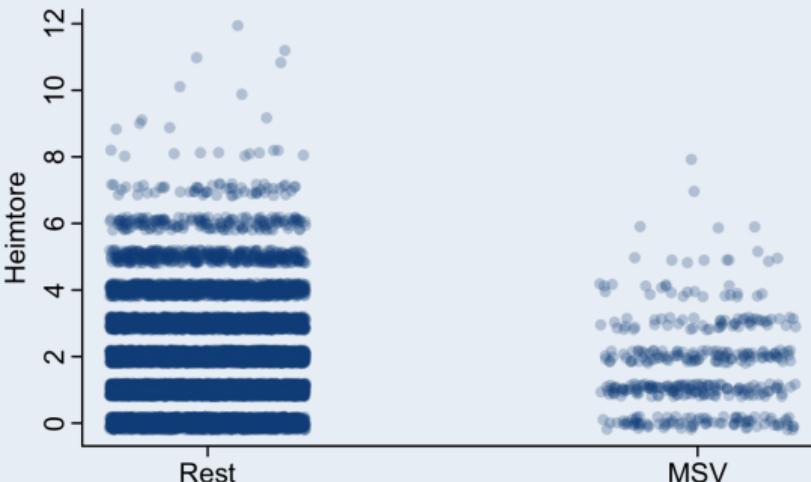
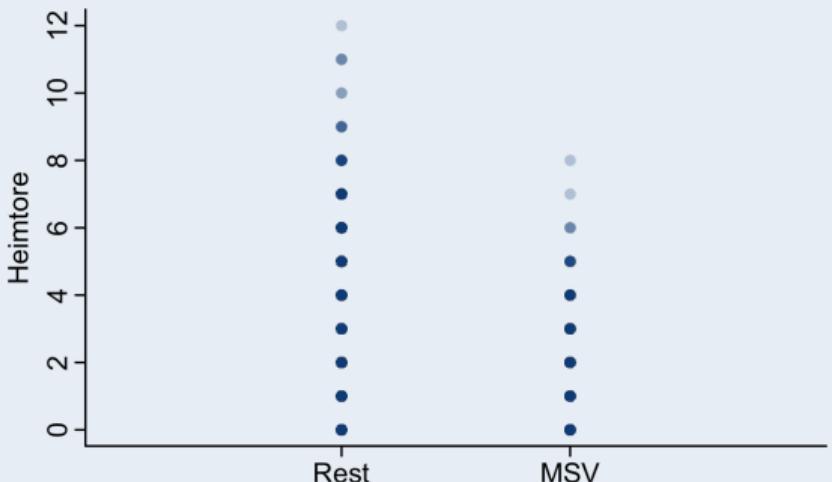
$$S = \sum_{r=1}^n \hat{u}_r^2 = \sum_{r=1}^n (y_r - a - bx_r)^2.$$

S ist bezüglich der Parameter a und b zu minimieren. Dies ist die **Methode der kleinsten Quadrate** (KQ-Methode) oder ordinary least squares (OLS)-Methode.

- Die Werte für a und b , die ein Minimum für S liefern, sind die geschätzten Regressionskoeffizienten. Da S eine nicht negative, quadratische Funktion ist, muss der Extremwert, falls er existiert, ein Minimum sein.

Beispiel 8.3: Heimspieltore des MSV Duisburg (Saison 65/66-06/07).

Graphische Analyse ist nicht immer aufschlussreich...



Quelle: Daten entnommen aus Heuer und Rubner (2009) Fitness, chance, and myths: an objective view on soccer results

Methode der kleinsten Quadrate

- Das Nullsetzen der beiden partiellen Ableitungen erster Ordnung nach a und b ist somit hier hinreichend für ein Minimum. Man kann die Ableitungen ungeachtet des Summenoperators nach der Kettenregel bilden:

$$\frac{\partial S}{\partial a} = -2 \sum_{r=1}^n (y_r - a - bx_r) \quad \text{und} \quad \frac{\partial S}{\partial b} = -2 \sum_{r=1}^n x_r(y_r - a - bx_r)$$

- Nullsetzen ergibt:

$$\sum_{r=1}^n (y_r - a - bx_r) = 0 \quad \text{und} \quad \sum_{r=1}^n x_r(y_r - a - bx_r) = 0.$$

- Hieraus folgen nach einfachen Umformungen die **Normalgleichungen**:

$$\sum_{r=1}^n y_r = na + b \sum_{r=1}^n x_r \quad \text{und} \quad \sum_{r=1}^n y_r x_r = a \sum_{r=1}^n x_r + b \sum_{r=1}^n x_r^2.$$

- Löst man die Gleichungen nach a und b auf, erhält man die **Schätzfunktionen**:

$$a = \bar{y} - b\bar{x} \quad \text{und} \quad b = \frac{\sum_{r=1}^n y_r x_r - n\bar{y}\bar{x}}{\sum_{r=1}^n x_r^2 - n\bar{x}^2} = \frac{\sum_{r=1}^n (y_r - \bar{y})(x_r - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

- Die geschätzte Regressionsgerade lautet jetzt: $\hat{y}_r = a + bx_r$; die Differenz

$$\hat{u}_r = y_r - (a + bx_r)$$

heißt (empirisches) **Residuum**.

- Dynamisch sieht das so aus: `illustrate_ls.R`

Methode der kleinsten Quadrate

Eigenschaften

- Die geschätzte Gerade verläuft immer durch den Schwerpunkt (\bar{x}, \bar{y}) der Punktfolge: Ersetzen von a in der Regressionsgleichung durch die Schätzfunktion ergibt:

$$\hat{y}_r = (\bar{y} - b\bar{x}) + bx_r \quad \text{oder} \quad \hat{y}_r = \bar{y} + b(x_r - \bar{x}).$$

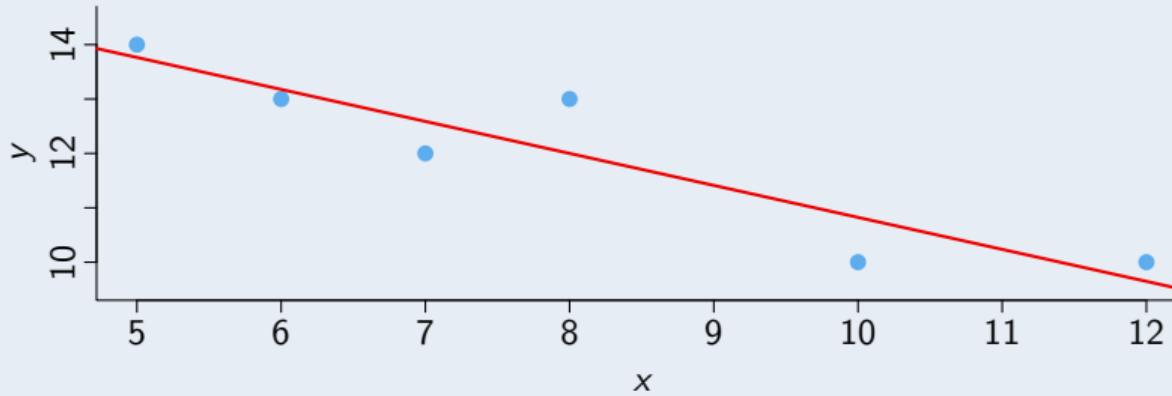
- Für $x = \bar{x}$ folgt $\hat{y}_r = \bar{y}$.
- Zudem ist die Summe aller OLS-Residuen \hat{u}_r null. Definitionsgemäß gilt:

$$\sum_{r=1}^n \hat{u}_r = \sum_{r=1}^n (y_r - \hat{y}_r) = \sum_{r=1}^n (y_r - a - bx_r) = \sum_{r=1}^n y_r - na - b \sum_{r=1}^n x_r.$$

- Substituiert man $\sum_{r=1}^n y_r$ durch die erste Normalgleichung, folgt stets auch: $\sum_{r=1}^n y_r = \sum_{r=1}^n \hat{y}_r$ und daher: $\bar{y} = \hat{\bar{y}}$.
- Ferner korrelieren die \hat{u}_r weder mit x_r noch mit \hat{y}_r : $s_{x\hat{u}} = s_{\hat{y}\hat{u}} = 0$ (siehe Übung).

Beispiel 8.4: KQ-Schätzung.

Das Streudiagramm zeigt einen negativen, linearen Zusammenhang $\hat{y} = a + bx_r$, $b < 0$.



```
x <- c(5:8,10,12)
y <- c(14, 13, 12, 13, 10, 10)
model <- lm(y ~ x)
```

Beispiel 8.4: Fortsetzung.

Die beiden arithmetischen Mittel betragen $\bar{x} = 8$ und $\bar{y} = 12$. Zur Berechnung von b kann eine Arbeitstabelle erstellt werden:

r	$x_r - \bar{x}$	$y_r - \bar{y}$	$(y_r - \bar{y})(x_r - \bar{x})$	$(x_r - \bar{x})^2$
1	-3	2	-6	9
2	0	1	0	0
3	2	-2	-4	4
4	4	-2	-8	16
5	-2	1	-2	4
6	-1	0	0	1
\sum	0	0	-20	34

Beispiel 8.4: Fortsetzung.

Für b erhält man

$$b = \frac{\sum_{r=1}^n (y_r - \bar{y})(x_r - \bar{x})}{\sum_{r=1}^n (x_r - \bar{x})^2} = -\frac{10}{17}.$$

Für a ergibt sich

$$a = 12 + \frac{10}{17} \cdot 8 = \frac{284}{17}.$$

Die geschätzte Regressionsgerade lautet somit

$$\hat{y}_r = \frac{284}{17} - \frac{10}{17}x_r.$$

Methode der kleinsten Quadrate

Beispiel 8.4 in 

```
x <- c(5, 6, 7, 8, 10, 12)
y <- c(14, 13, 12, 13, 10, 10)

(b <- cov(x, y) / var(x)) # Korrektur nicht nötig, da diese sich rauskürzt
## [1] -0.5882353

(a <- mean(y) - b * mean(x))
## [1] 16.70588

# R die Arbeit machen lassen:
(model <- lm(y ~ x))

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)          x
##       16.7059      -0.5882
```

Methode der kleinsten Quadrate

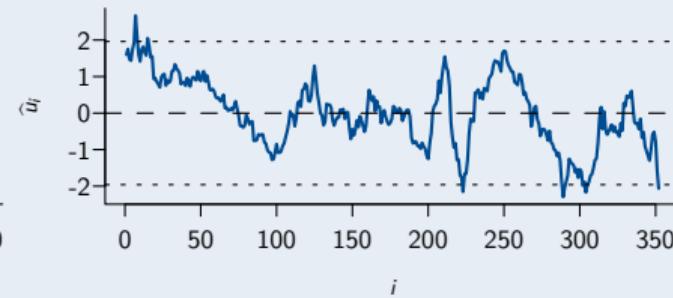
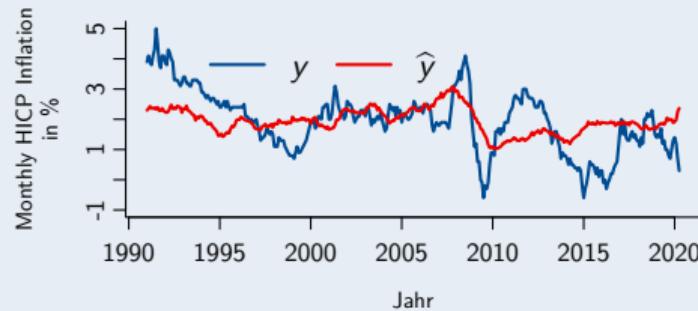
Lineare Regression in

M3Inflation.R

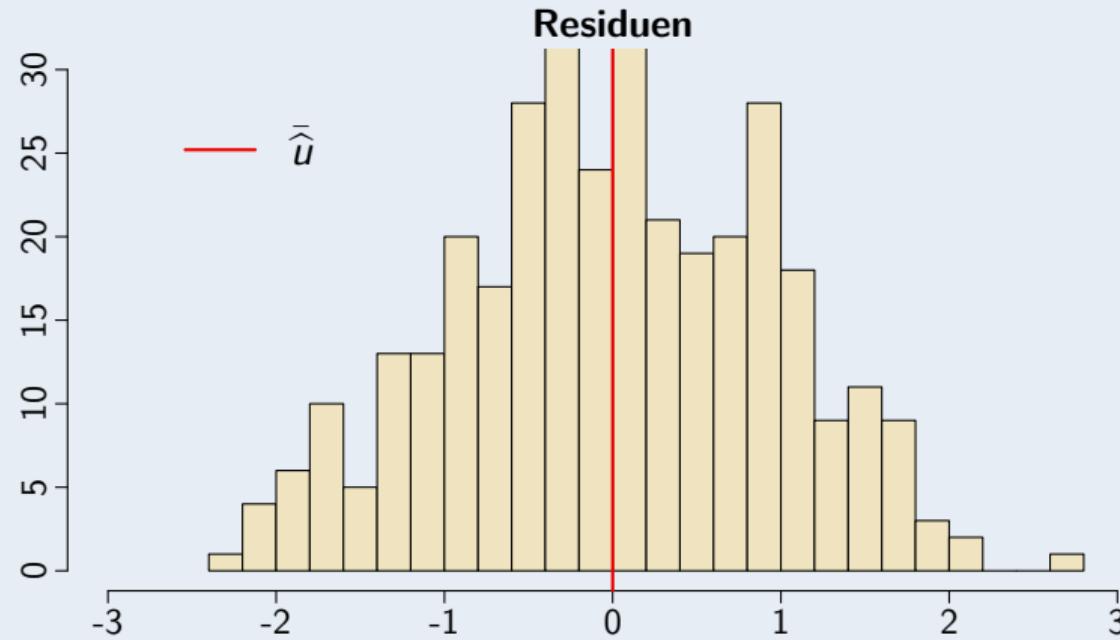
```
m3g <- window(m3g, start = c(1991, 1), end = c(2020, 4))
hicp <- window(hicp, start = c(1991, 1), end = c(2020, 4))
# Estimate lm
model <- lm(hicp ~ m3g)
summary(model)

##
## Call:
## lm(formula = hicp ~ m3g)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.29517 -0.60981  0.01805  0.73246  2.66805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.09023   0.11526   9.459 < 2e-16 ***
## m3g         0.15920   0.01945   8.186 5.08e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9382 on 350 degrees of freedom
## Multiple R-squared:  0.1607, Adjusted R-squared:  0.1583 
## F-statistic: 67.02 on 1 and 350 DF,  p-value: 5.078e-15
```

Beispiel 8.5: Zusammenhang zwischen Geldmengenwachstum und Inflation.



Beispiel 8.5: Fortsetzung.



Regressionsrechnung 1

Kahoot!

MSVHeimtore.R

```
load("Daten/msvheimtore.Rda")
# MSV = 1, Rest = 0
summary(lm(Tore ~ Team, data = msvheimtore))

##
## Call:
## lm(formula = Tore ~ Team, data = msvheimtore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.921 -0.921  0.079  1.079 10.079 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.92099   0.01335 143.850 < 2e-16 ***
## Team        -0.31805   0.07405 -4.295 1.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.471 on 12544 degrees of freedom
## Multiple R-squared:  0.001468, Adjusted R-squared:  0.001389 
## F-statistic: 18.45 on 1 and 12544 DF,  p-value: 1.761e-05
```

Varianzzerlegung und Bestimmtheitsmaß

- Wie gut erfasst die ermittelte Regressionsgerade die Variation der Beobachtungen y_r ? Hierzu verwendet man das **Bestimmtheitsmaß R^2** , auch Determinationskoeffizient genannt.
- Jede Beobachtung y_r kann zerlegt werden in $y_r = \hat{y}_r + \hat{u}_r$.
- Subtrahiert man auf beiden Seiten \bar{y} , quadriert und summiert, ergibt sich:

$$\begin{aligned}\sum_{r=1}^n (y_r - \bar{y})^2 &= \sum_{r=1}^n (\hat{y}_r - \bar{y} + \hat{u}_r)^2 \\ &= \sum_{r=1}^n (\hat{y}_r - \bar{y})^2 + \sum_{r=1}^n \hat{u}_r^2 + 2 \sum_{r=1}^n (\hat{y}_r - \bar{y})\hat{u}_r\end{aligned}$$

- Der letzte Term der Umformung ist null: $\sum_{r=1}^n (\hat{y}_r - \bar{y})\hat{u}_r = \sum_{r=1}^n \hat{y}_r \hat{u}_r = 0$.

- Damit ist $\sum_{r=1}^n (y_r - \bar{y})^2$, die Streuung der y_r um \bar{y} , zerlegt in („**Streuungszerlegungsformel**“):

$$\sum_{r=1}^n (y_r - \bar{y})^2 = \sum_{r=1}^n (\hat{y}_r - \bar{y})^2 + \sum_{r=1}^n \hat{u}_r^2.$$

- Division durch n liefert

$$s_y^2 = s_{\hat{y}}^2 + s_{\hat{u}}^2.$$

- Für R^2 sind nun zwei Komponenten relevant:
- $s_{\hat{y}}^2$ ist die Varianz, die entstünde, lägen alle Beobachtungstupel auf der Regressionsgeraden. Man nennt sie die **erklärte Varianz**.
- $s_{\hat{u}}^2$ ist die durch die Regression nicht erklärte oder **unerklärte Varianz**.

Die Streuung der (x_r, y_r) um die Regressionsgerade ist umso geringer, je größer der Anteil von $s_{\hat{y}}^2$ an s_y^2 ist.

Definition 8.6: Bestimmtheitsmaß R^2 .

Das Bestimmtheitsmaß ist daher definiert als

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{r=1}^n (\hat{y}_r - \bar{y})^2}{\sum_{r=1}^n (y_r - \bar{y})^2} \quad \text{oder} \quad R^2 = 1 - \frac{\sum_{r=1}^n \hat{u}_r^2}{\sum_{r=1}^n (y_r - \bar{y})^2}.$$

Liegen alle Beobachtungen auf der Regressionsgeraden, stimmen Gesamt- und erklärte Varianz überein:
 $R^2 = 1$. Kann kein Anteil der Varianz von Y durch die Regressionsgerade erklärt werden, gilt $s_y^2 = s_{\hat{u}}^2$ und $R^2 = 0$.

Regressionsrechnung 2

Kahoot!

- Die erklärte Varianz $s_{\hat{y}}^2$ lässt sich auf die Varianz von X zurückführen. Wegen $\hat{y}_r = a + bx_r$ und $\hat{y}_r - \bar{y} = b(x_r - \bar{x})$ folgt

$$s_{\hat{y}}^2 = \frac{1}{n} \sum_{r=1}^n (\hat{y}_r - \bar{y})^2 = b^2 \frac{1}{n} \sum_{r=1}^n (x_r - \bar{x})^2 = b^2 s_x^2.$$

- Einsetzen von $b = s_{xy}/s_x^2$ liefert $s_{\hat{y}}^2 = s_{xy}^2/s_x^2$. Folglich gilt

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2$$

- R^2 ist also bei einer linearen, inhomogenen Regressionsgleichung gleich dem quadrierten Korrelationskoeffizienten.

Beispiel 8.7: Bestimmtheitsmaß.

Berechne für die Daten aus Beispiel 8.4 zunächst die gerundeten Regresswerte \hat{y}_r , $r = 1, \dots, 6$ (Spalte 2) gemäß obiger Regressionsgeraden $\hat{y}_r = \frac{284}{17} - \frac{10}{17}x_r$. Als Bestimmtheitsmaß erhält man $R^2 = 11,75/14 = 0,8393$; d.h.: 83,93% der Varianz in Y wird durch die Regression erklärt, 16,07% bleiben unerklärt.

r	\hat{y}_r	$(\hat{y}_r - \bar{y})^2$	$(y_r - \bar{y})^2$
1	13,76	3,10	4
2	12,00	0,00	1
3	10,82	1,39	4
4	9,65	5,52	4
5	13,18	1,39	1
6	12,59	0,35	0
\sum		11,75	14

Anknüpfend an Beispiel 8.4:

```
(yhat <- a + b * x)
## [1] 13.764706 13.176471 12.588235 12.000000 10.823529 9.647059

(var(yhat) / var(y)) # durch Rundung auf der vorherigen Folie, leichte Abweichung
## [1] 0.8403361

# Natürlich muss man R^2 nicht per Hand berechnen:
model_detailed <- summary(model)
model_detailed$r.squared
## [1] 0.8403361
```

- Der Zusammenhang zwischen Y und X muss nicht linear sein. Viele Theorien suggerieren nichtlineare Abhängigkeit zwischen Variablen.
- KQ ist auch anwendbar, wenn die Normalgleichungen einer nichtlinearen Regression linear in den Parametern sind, oder wenn die nichtlineare Funktion in eine sowohl in den Parametern als auch in den Variablen lineare Gleichung überführt werden kann.

Beispiel 8.8: Transformation von Regressoren.

Bei Linearität in den Parametern, wie dies z.B. in $y_r = \alpha + \beta x_r^n + u_t$, $n \in \mathbb{R} \setminus \{0\}$ der Fall ist, schätzt man α und β , nachdem der Regressor transformiert wurde: $y_r = \alpha + \beta x_r^* + u_t$ mit $x_r^* = x_r^n$. Für $x_r : 1, 2, 3, 4, \dots$ und $n = 2$ sind die transformierten Werte $x_r^* : 1, 4, 9, 16, \dots$. Zur Berechnung der Koeffizienten verwendet man danach die bekannten Formeln.

Beispiel 8.8: Fortsetzung.

- Ist die nichtlineare Regressionsfunktion vom Typ $y_r = \alpha x_r^\beta$, wird sie durch Logarithmieren linear in den beiden jetzt logarithmierten Variablen:

$$\ln y_r = \ln \alpha + \beta \ln x_r.$$

- Man nennt sie deshalb doppelt logarithmische Funktion. Sie enthält einen nichtlinearen Koeffizienten $\ln \alpha$.
- Setzen von $y_r^* := \ln y_r, x_r^* := \ln x_r$ liefert eine in den Variablen und Koeffizienten lineare Schätzfunktion: $y_r^* = \alpha^* + \beta x_r^*$.
- Schätzungen für α^* und β gewinnt man nach Substitution von x_r und y_r durch x_r^* und y_r^* . Aus $\hat{\alpha}^*$ folgt die Schätzung für α als $e^{\hat{\alpha}^*}$.

Beispiel 8.8: Fortsetzung.

- Schließlich kann die nichtlineare Beziehung zwischen Y und X durch $y_r = \alpha e^{\beta x_r}$ gegeben werden. Solche Funktionen verwendet man z.B. bei (stetigen) Wachstumsprozessen, wobei x_r dann die Zeit darstellt. Logarithmische Transformation liefert hier $\ln y_r = \ln \alpha + \beta x_r$.
- Die linearisierte Schätzgleichung ist

$$y_r^* = \alpha^* + \beta x_r.$$

- Da nach der Transformation nur die links vom Gleichheitszeichen stehende Variable logarithmiert ist, bezeichnet man die Gleichung auch als halblogarithmisch.

Nichtlineare Regression

Nichtlineare Regression - in

Beispiel 8.9: Stundenlöhne und Alter.

Wir nehmen an, dass der Zusammenhang zwischen Stundenlöhnen (y) und Alter (x) durch

$$y_i = \alpha e^{\beta x_i}$$

erklärt werden kann. Durch Logarithmierung folgt

$$\ln(y_i) = \ln(\alpha) + \beta x_i.$$

Die Schätzgleichung lautet daher

$$y_i^* = \alpha^* + \beta x_i.$$

```
library(AER)
data("CPS1985")
CPS1985$logWage <- log(CPS1985$wage)
summary(model <- lm(logWage ~ age, data = CPS1985))
```

Nichtlineare Regression

Nichtlineare Regression - in

```
##  
## Call:  
## lm(formula = logWage ~ age, data = CPS1985)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -2.10528 -0.36523 -0.00181  0.35425  1.87755  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.730603  0.073919 23.412 < 2e-16 ***  
## age         0.008921  0.001912  4.665 3.92e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.5178 on 532 degrees of freedom  
## Multiple R-squared:  0.03929,    Adjusted R-squared:  0.03749  
## F-statistic: 21.76 on 1 and 532 DF,  p-value: 3.916e-06
```

Nichtlineare Regression

Nichtlineare Regression - in

Beispiel 8.9: Fortsetzung.

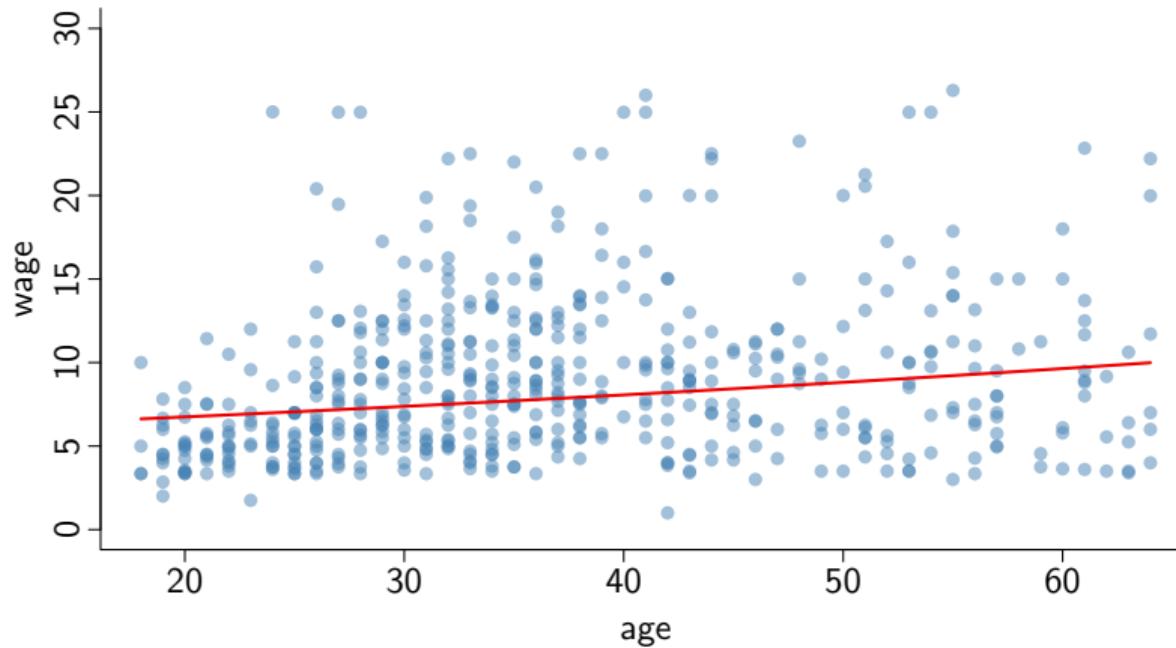
Nachdem die Koeffizienten geschätzt wurden, kann die ursprüngliche Gleichung rekonstruiert werden.
Daher gilt

$$\hat{y} \approx e^{1.7306} \cdot e^{0.0089x_i}$$

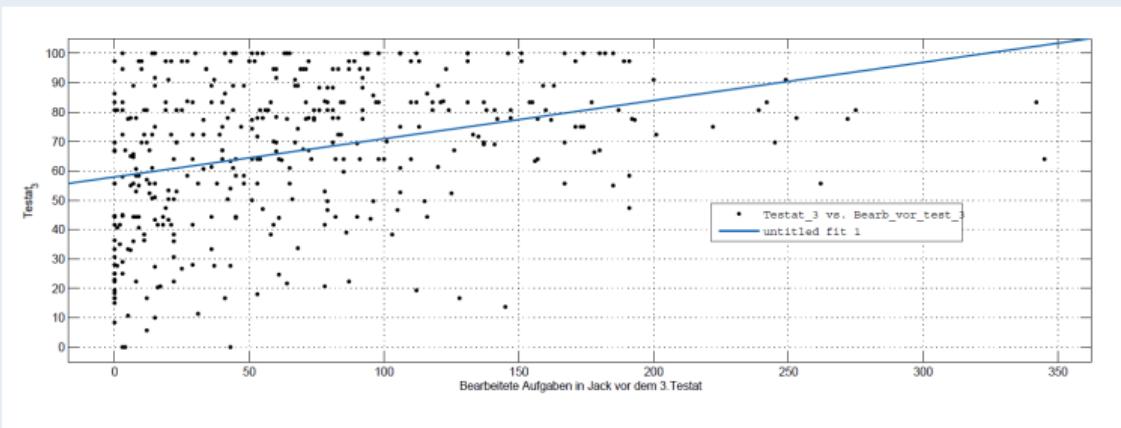
```
plot(wage ~ age, data = CPS1985, col = alpha("steelblue", 0.5), pch = 16,
      ylim = c(0,30))
grid <- seq(from = 18, to = 64, by = 0.1)
lines(x = grid, y = exp(1.730603) * exp(0.008921 * grid), lwd = 2,
      col = "red2")
```

Nichtlineare Regression

Nichtlineare Regression - in

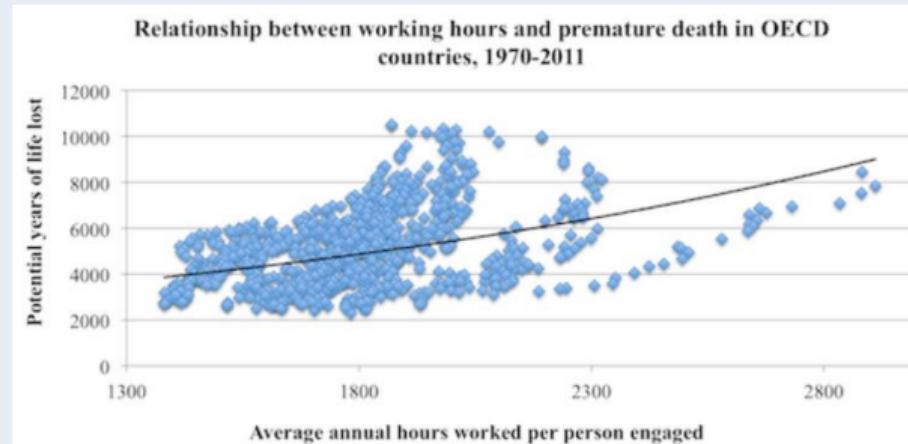


Beispiel 8.10: Testat 3.



$$\widehat{\text{Testat}_r} = 57,93 + 0,13 \text{Bearbeitete Aufgaben}_r$$

Beispiel 8.11: Arbeiten und Sterben.



Quelle: Economist.

Simpsons Paradox und Regression

```
raucher.jung.tot <- matrix(rep(c(1,1,1), 27), nrow=27, byrow = T)
nichtraucher.jung.tot <- matrix(rep(c(0,1,1), 12), nrow=12, byrow = T)
nichtraucher.jung.lebendig <- matrix(rep(c(0,1,0), 103), nrow=103, byrow = T)
raucher.jung.lebendig <- matrix(rep(c(1,1,0), 66), nrow=66, byrow = T)

raucher.mittel.tot <- matrix(rep(c(1,2,1), 51), nrow=51, byrow = T)
nichtraucher.mittel.tot <- matrix(rep(c(0,2,1), 40), nrow=40, byrow = T)
nichtraucher.mittel.lebendig <- matrix(rep(c(0,2,0), 64), nrow=64, byrow = T)
raucher.mittel.lebendig <- matrix(rep(c(1,2,0), 81), nrow=81, byrow = T)

raucher.alt.tot <- matrix(rep(c(1,3,1), 29), nrow=29, byrow = T)
nichtraucher.alt.tot <- matrix(rep(c(0,3,1), 101), nrow=101, byrow = T)
nichtraucher.alt.lebendig <- matrix(rep(c(0,3,0), 7), nrow=7, byrow = T)
raucher.alt.lebendig <- matrix(rep(c(1,3,0), 28), nrow=28, byrow = T)
```

Simpsons Paradox und Regression

```
daten <- rbind(raucher.jung.tot, nichtraucher.jung.tot,
                 nichtraucher.jung.lebendig, raucher.jung.lebendig,
                 raucher.mittel.tot, nichtraucher.mittel.tot,
                 nichtraucher.mittel.lebendig, raucher.mittel.lebendig,
                 raucher.alt.tot, nichtraucher.alt.tot,
                 nichtraucher.alt.lebendig, raucher.alt.lebendig)

raucher <- daten[, 1]
altersgruppe <- daten[, 2]
tot <- daten[, 3]

lm(tot~raucher)
##
## Call:
## lm(formula = tot ~ raucher)
##
## Coefficients:
## (Intercept)      raucher
##          0.46789     -0.08846
```

Simspons Paradox und Regression

```
lm(tot[altersgruppe==1]~raucher[altersgruppe==1])  
##  
## Call:  
## lm(formula = tot[altersgruppe == 1] ~ raucher[altersgruppe ==  
##      1])  
##  
## Coefficients:  
##                 (Intercept)  raucher[altersgruppe == 1]  
##                         0.1043          0.1860  
  
lm(tot[altersgruppe==2]~raucher[altersgruppe==2])  
##  
## Call:  
## lm(formula = tot[altersgruppe == 2] ~ raucher[altersgruppe ==  
##      2])  
##  
## Coefficients:  
##                 (Intercept)  raucher[altersgruppe == 2]  
##                         0.384615        0.001748
```

Simspons Paradox und Regression

```
lm(tot[altersgruppe==3] ~ raucher[altersgruppe==3])  
  
##  
## Call:  
## lm(formula = tot[altersgruppe == 3] ~ raucher[altersgruppe ==  
##     3])  
##  
## Coefficients:  
##                 (Intercept)  raucher[altersgruppe == 3]  
##                     0.9352          -0.4264
```

```
lm(tot ~ raucher + factor(altersgruppe) + raucher:factor(altersgruppe))  
  
##  
## Call:  
## lm(formula = tot ~ raucher + factor(altersgruppe) + raucher:factor(altersgruppe))  
##  
## Coefficients:  
##                 (Intercept)           raucher  
##                     0.1043            0.1860  
##   factor(altersgruppe)2      factor(altersgruppe)3  
##                     0.2803            0.8308  
##  raucher:factor(altersgruppe)2  raucher:factor(altersgruppe)3  
##                     -0.1842            -0.6124
```

- Dieses Kapitel behandelte Grundlagen der Regressionsanalyse. Dies ist u.a. wichtig für Veranstaltungen der Ökonometrie.
- Nachbearbeitung: Kapitel 5.3 des Buches von Prof. Assenmacher.
- Das nächste Kapitel befasst sich mit der Zeitreihenanalyse. Dort wird versucht ökonomische Phänomene anhand der Ursache „Zeit“ zu erklären.
- Vorbereitung: Kapitel 6 des Buches von Prof. Assenmacher.

- 0 Motivation
- 1 Grundzüge der Datenerhebung
- 2 Eindimensionale Häufigkeitsverteilungen
- 3 Lageparameter
- 4 Streuungsparameter
- 5 Schiefe- und Kurtosisparameter
- 6 Konzentrations- und Disparitätsmessung
- 7 Zweidimensionale Datensätze
- 8 Regressionsrechnung
- 9 Elementare Zeitreihenanalyse

- Werden Beobachtungen als **Längsschnitt** erhoben, bilden sie eine **Zeitreihe** y_1, y_2, \dots, y_T oder $y_t, t = 1, \dots, T$.
- Eine Zeitreihe ist also eine zeitlich geordnete Folge von Beobachtungen (z.B. das BIP in aufeinander folgenden Quartalen/Jahren).
- Die Zeitreihenanalyse untersucht Zeitreihen hinsichtlich typischer Bewegungsmuster. Wir verwenden dabei Methoden der deskriptiven Statistik.

Grundlagen

Etwas Selbstkritik :-)

Statistik ist nicht immer der beste Weg Information zu präsentieren.



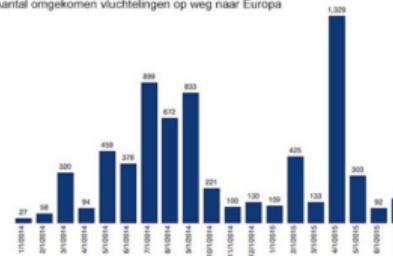
Buitenland

Dit iconische staafdiagram gaat de wereld over

Soms zegt één grafiek meer dan duizend woorden

[Printen](#) [Emailen](#)

Aantal omgekomen vluchtelingen op weg naar Europa



Door Jochum van den Berg en Diederik Smit • vrijdag 4 september 2015

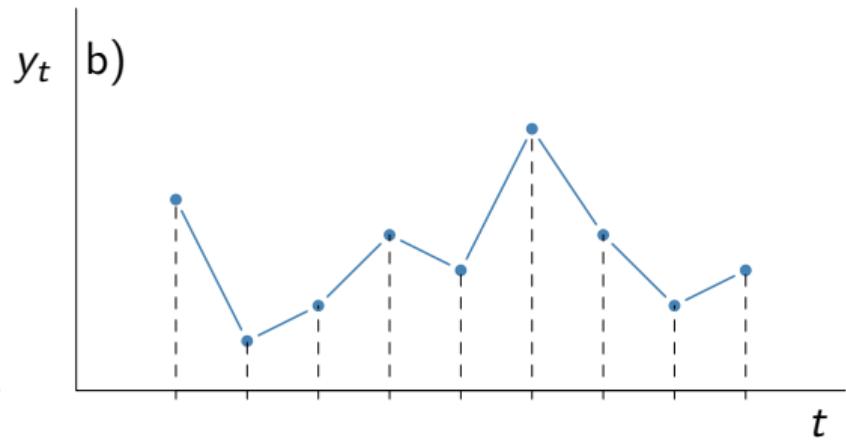
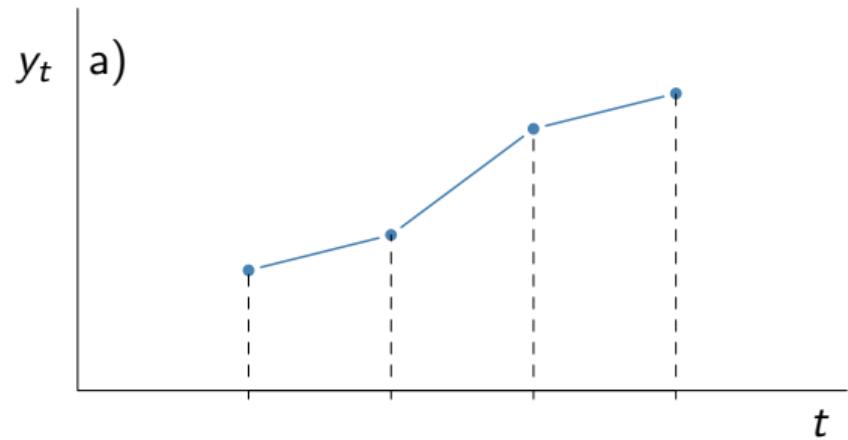
Deze grafiek toont slechts een paar van de vele cijfers die inmiddels over de vluchtelingencrisis zijn verschenen. Toch is het deze grafiek die de wereld over gaat. Omdat het de kern van de tragedie op schrijnende wijze samenvat. Soms zegt één staafdiagram meer dan duizend woorden.

De cijfers werden gisteren massaal gedeeld op sociale media en talloze kranten hebben de grafiek op hun voorpagina geplaatst. Waar het staafdiagram precies

5 ma
met e
Win t
Chek
Spar
T-Mc
brenj

...

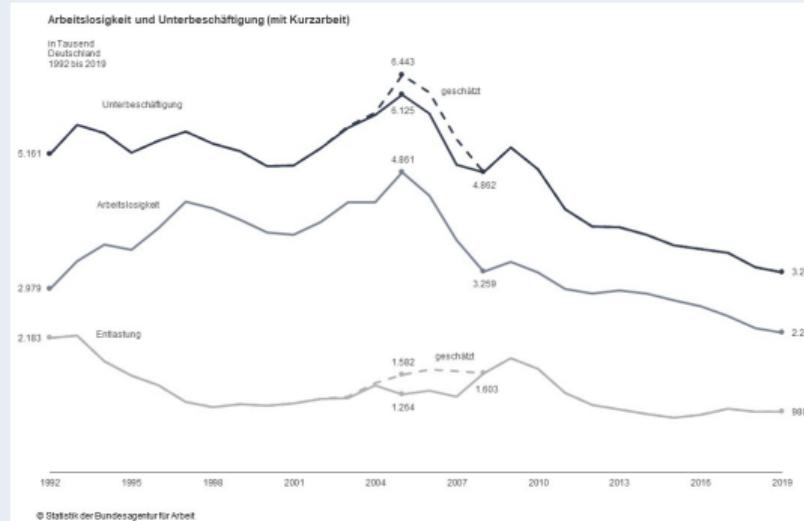
Grundlagen



- Ist das Diagramm wie in a) annähernd linear, ändert sich y_t fast konstant mit Änderungen $\Delta y_t = y_t - y_{t-1}$.
- Ist der Graph für logarithmierte Zeitreihenwerte linear, wächst die Zeitreihe mit nahezu konstanten Wachstumsraten.
- Nur selten weist die Zeitreihe eine eindeutige Form wie in a) auf; häufiger sind Graphen mit auf den ersten Blick unregelmäßigen Mustern wie in b).
- Die Stärke der Unregelmäßigkeit wird auch von der Maßeinheit der Zeit beeinflusst. Schwankungen nehmen mit kleiner werdender Zeiteinheit (Grundzeitperiode) tendenziell zu. Eine Zeitreihe mit Jahreswerten erscheint oft glatter als eine mit Monatsdaten.

Grundlagen

Beispiel 9.1: Arbeitslosenquote.



Quelle: Statistik der Bundesagentur für Arbeit, Zeitreihengrafiken

Entlastung: „zeitlich befristeter Effekt von arbeitsmarktpolitischen Instrumenten, [...] ohne die die Arbeitslosigkeit entsprechend höher ausfallen würde“.

(vgl. Bundesagentur für Arbeit (2009), Umfassende Arbeitsmarktstatistik: Arbeitslosigkeit und Unterbeschäftigung, S. 24 ff.)

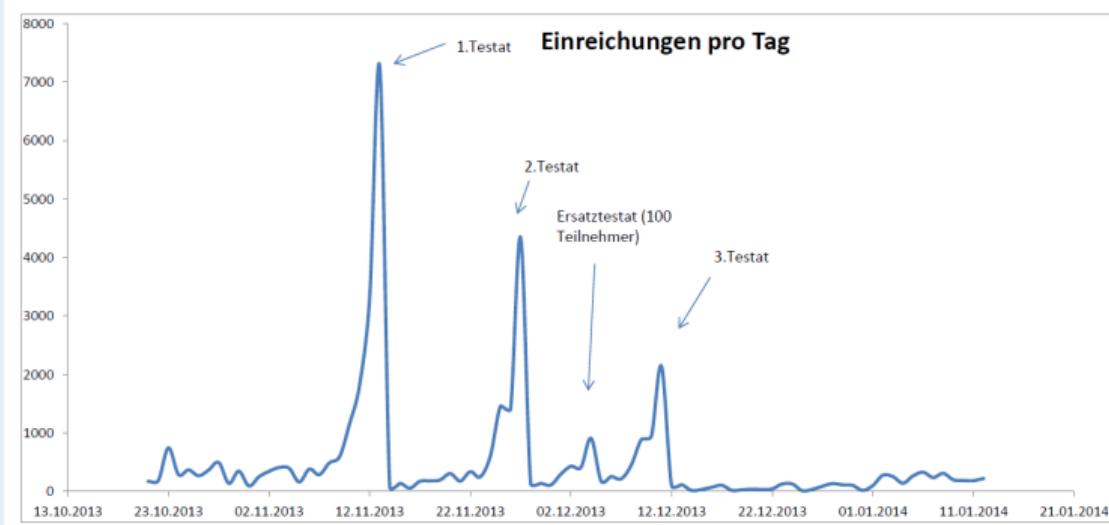
Grundlagen

Beispiel 9.2: Wechselkurs Euro - Türkische Lira.



Quelle: Finanzen.net

Beispiel 9.3: Einreichungen bei JACK.



Grundlagen

- Grafische Betrachtung sollte durch statistische Analyse ergänzt werden. Hierzu ist ein Modell hilfreich.
- Wir stellen uns vor, dass sich jede Beobachtung y_t aus bestimmten **Komponenten** zusammensetzt. Wir beschränken uns auf für ökonomische Zeitreihen relevante Komponenten. Zeitreihenanalyse wird aber auch in vielen anderen Disziplinen genutzt.
- Allgemein lässt sich y_t in **systematische und nicht systematische Komponenten** zerlegen. Systematische Komponenten beeinflussen die Zeitreihe monoton oder periodisch.
- Bei den meisten ökonomischen Zeitreihen, die für kürzere Zeiträume als ein Jahr erhoben werden, ist eine Zerlegung von y_t in drei systematische und eine nicht systematische Komponente sinnvoll.

Grundlagen

- Man unterscheidet daher:
 - ① eine **Trendkomponente** m_t , die eine langfristige, aus dem ökonomischen Wachstumsprozess resultierende Entwicklung erfasst;
 - ② eine **zyklische bzw. konjunkturelle Komponente** k_t , die mehrjährige, quasiperiodische Schwankungen um den langfristigen Trend wiedergibt;
 - ③ eine kurzfristige **saisonale Komponente** s_t , die jahreszeitlich bedingte Schwingungen in y_t zum Ausdruck bringt und
 - ④ als nicht systematische Komponente eine **Restkomponente** r_t , die alle weiteren, auch singulären Einflüsse, einschließt.
-
- Von r_t nehmen wir an, dass sie regellos um den Wert null streut. Man bezeichnet r_t daher auch als Störkomponente bzw. Störvariable.

Grundlagen

- Wir modellieren des Weiteren das Zusammenwirken der Komponenten bei der Erzeugung von y_t . Man unterscheidet drei Erklärungsansätze:

① additives Modell:

$$y_t = m_t + k_t + s_t + r_t$$

② multiplikatives Modell:

$$y_t = m_t k_t s_t r_t$$

③ gemischt additiv-multiplikatives Modell.

- Im additiven ist der Einfluss jeder Komponente auf y_t unabhängig von dem Niveau der übrigen.
- Hängt der Einfluss einer Komponente von y_t vom Niveau der anderen ab, ist das multiplikative Modell adäquat: Jedes y_t ist das Produkt der einzelnen Komponenten.

Grundlagen

- Da durch Logarithmieren das multiplikative in das additive Modell übergeht, ist eine getrennte Behandlung beider Modelle überflüssig.
- Beim gemischten Modell sind die Komponenten sowohl additiv als auch multiplikativ verknüpft. So können Saisonschwankungen vom Niveau der Trend- und konjunkturellen Komponente abhängen, die Restschwankungen aber niveauunabhängig sein: $y_t = (m_t + k_t)s_t + r_t$.
- Die Modelle sind jeweils dem Problem anzupassen. In Jahresdaten kann etwa die innerjährige saisonale Entwicklung nicht identifiziert werden.
- Ist eine Trennung in Trend- und zyklische Komponente inhaltlich oder empirisch kaum möglich, werden beide zur **glatten Komponente** $g_t = m_t + k_t$ zusammengefasst. Bei makroökonomischen Zeitreihen können Trend und Zyklus oft gut diskriminiert werden.

- Man unterscheidet je nach Annahmen zwischen globalen und lokalen Komponentenmodellen.
- **Globale Komponentenmodelle** unterstellen, dass y_t durch eine Struktur erzeugt wurde, deren Parameter über den gesamten Zeitraum konstant bleiben. Diese Modelle eignen sich besonders für Trend- und glatte Komponente.
- Bei **lokalen Komponentenmodellen** sind die Modellparameter nur für bestimmte Zeitabschnitte konstant, über den gesamten Beobachtungszeitraum aber variabel. Damit lässt sich die konjunkturelle, aber auch die glatte Komponente, falls diese Zyklen aufweist, schätzen.

- Eine Zeitreihe kann „bereinigt“ werden.
- Die um die geschätzte glatte Komponente \hat{g}_t bereinigte Zeitreihe $y_t - \hat{g}_t$ (beim additiven Modell bzw. nach Logarithmieren auch beim multiplikativen Modell) enthält jetzt nur noch Saison- und Restkomponente.
- Annahmen über das Saisongesetz erlauben eine Schätzung von s_t und r_t . Damit ist jedes y_t in seine geschätzten Komponenten zerlegbar:

$$y_t = \hat{g}_t + \hat{s}_t + \hat{r}_t$$

bzw.

$$y_t = \hat{m}_t + \hat{k}_t + \hat{s}_t + \hat{r}_t.$$

Damit kann eine Bereinigung der Originalreihe, z.B. eine **Trendelimination** durch Subtraktion der jeweiligen geschätzten Komponente erfolgen.

- Ein globales Komponentenmodell modelliert g_t als Funktion der Zeit t . Hier werden nur einfache, aber oft nützliche Spezifikationen behandelt, die mit der Kleinsten Quadrate-Methode (OLS) geschätzt werden können.
- Zur Schätzung von g_t werden Saison- und Restkomponente zu einer Variablen u_t zusammengefasst. Das additive Komponentenmodell ist dann

$$y_t = g_t + u_t, \quad t = 1, \dots, T.$$

- Die einfachste Abhängigkeit von g_t von t ist die lineare, $g_t = \alpha + \beta t$:

$$y_t = \alpha + \beta t + u_t, \quad t = 1, \dots, T.$$

- Es liegt jetzt eine einfache lineare Regression vor; α und β können daher mit OLS geschätzt werden.

- Nach Substitution von x_t durch t und n durch T erhält man aus den bekannten Schätzgleichungen die Schätzungen a und b als:

$$a = \bar{y} - b\bar{t} \quad \text{und} \quad b = \frac{\sum_{t=1}^T ty_t - T\bar{y}\bar{t}}{\sum_{t=1}^T t^2 - T(\bar{t})^2} = \frac{s_{yt}}{s_t^2}.$$

- Die Summeneigenschaften der natürlichen Zahlen vereinfachen obige Gleichung. Wegen $\sum_{t=1}^T t = \frac{T}{2}(T+1)$, $\bar{t} = \frac{1}{T} \sum_{t=1}^T t = \frac{T+1}{2}$ und $\sum_{t=1}^T t^2 = \frac{T(T+1)(2T+1)}{6}$ ist

$$b = \frac{12 \sum_{t=1}^T ty_t - 6(T+1) \sum_{t=1}^T y_t}{T(T^2 - 1)}.$$

- Mit a und b ist die glatte Komponente geschätzt: $\hat{g}_t = a + bt$.

Ermittlung der glatten Komponente

- Entwickelt sich g_t nach einer Potenz- oder Wurzelfunktion, setzt man: $g_t = \alpha t^\beta$ mit $\beta \neq 0$ und $\beta \neq 1$. Dieser Ansatz modelliert für $\beta > 1$ überproportionale, für $0 < \beta < 1$ (Wurzelfunktion) unterproportionale Veränderungen von y_t .
- Nutze nun zur Schätzung der Parameter von g_t das multiplikative Modell, das nach Einsetzen von $g_t = \alpha t^\beta$ lautet $y_t = \alpha t^\beta u_t$.
- Logarithmische Transformation, Neudefinition der Variablen und eines Parameters liefern die lineare Regression

$$y_t^* = \alpha^* + \beta t^* + u_t^*,$$

mit $y_t^* := \ln y_t$, $t^* := \ln t$, $u_t^* := \ln u_t$ und $\alpha^* := \ln \alpha$.

- Die Parameter α^* und β werden mit den bekannten Formeln mit logarithmierten Variablen geschätzt.
- Schätze β direkt mit b , sowie α durch $a = e^{a^*}$.

- Legen (z.B. makroökonomische) **Wachstumsprozesse** konstante Wachstumsraten für g_t nahe, ist eine **Exponentialfunktion** heranzuziehen: $g_t = e^{\alpha + \beta t}$, $\beta \neq 0$. Für Wachstumsprozesse gilt meist $\beta > 0$.
- $\beta < 0$ ist aber auch möglich (exponentieller Verfall).
- Da $g_0 = e^\alpha$ der Startwert aus $t = 0$ ist, schreibt man auch $g_t = g_0 e^{\beta t}$.
- Die Schätzung per OLS erfolgt über ein log-lineares Komponentenmodell:

$$y_t = e^{(\alpha + \beta t + u_t)} = \exp(\alpha + \beta t + u_t),$$

wobei $\exp(a) = e^a$. Logarithmieren der Gleichung ergibt $\ln y_t = \alpha + \beta t + u_t$.

- D.h. die logarithmierten y_t hängen linear von g_t ab, daher die Bezeichnung **log-linear**.

Beispiel 9.4: BIP.

Mal wieder etwas Gapminder

- Obige globale Modelle unterstellen für g_t Abwesenheit von **Schwankungen**. Ggf. ist eine Spezifikation zu wählen, die Schwankungen zulässt.
- Hinweise auf Schwankungen liefert das Zeitreihendiagramm, das sowohl bei Jahresdaten als auch unterjährigen Beobachtungen (Monats-, Quartalsdaten) dann Zyklen aufweisen müsste.
- Ignoriert man Zyklen durch die Modellierung eines schwankungsfreien g_t , würden diese fälschlicherweise r_t zugerechnet.
- Die Ermittlung einer zyklischen glatten Komponente erfolgt mit **lokalen Komponentenmodellen**. Wir besprechen einen einfachen Ansatz, der keine mathematische Funktion für die glatte Komponente voraussetzt.

- Die glatte Komponente wird lokal geschätzt, indem für jeweils $2\lambda + 1$, $\lambda \in \mathbb{N}$ aufeinander folgende y_t , beginnend mit y_1 , das lokale arithmetische Mittel berechnet und der mittleren Beobachtungsperiode zugeordnet wird.
- Da für jedes $\lambda \in \mathbb{N}$ der Term $2\lambda + 1$ ungerade ist, liegt ein ungerades, lokales arithmetisches Mittel vor:

$$\bar{y}_t = \frac{1}{2\lambda + 1} \sum_{\tau=t-\lambda}^{t+\lambda} y_\tau = \frac{1}{2\lambda + 1} (y_{t-\lambda} + \dots + y_t + \dots + y_{t+\lambda}), \quad \lambda \in \mathbb{N}.$$

- Die Summe verdeutlicht die Rolle von λ : Es liegen genau λ Werte vor und nach der mittleren Periode t . Für eine Zeitreihe mit T Werten berechnen wir $T - 2\lambda$ \bar{y}_t , die $t = \lambda + 1, \dots, T - \lambda$ zugeordnet werden und **einfache gleitende Durchschnitte** der Ordnung $2\lambda + 1$ heißen.

Ermittlung der glatten Komponente

- Liegt \bar{y}_t eine gerade Anzahl an Zeitreihenwerten zugrunde, existiert kein ganzzahliges t , dem \bar{y}_t mit der Ordnung 2λ zuzuordnen wäre. Man bildet dann einen ungeraden gleitenden Durchschnitt der Ordnung $2\lambda + 1$, gewichtet jedoch die beiden Randwerte jeweils nur mit 0,5:

$$\bar{y}_t = \frac{1}{2\lambda} \left(\frac{1}{2}y_{t-\lambda} + \sum_{\tau=t-(\lambda-1)}^{t+\lambda-1} y_\tau + \frac{1}{2}y_{t+\lambda} \right), \quad \lambda \in \mathbb{N}.$$

- Gleitende Durchschnitte sind **lineare Filter**. Die überführung von y_t in gleitende Durchschnitte bezeichnet man als Filtration bzw. Filtern.
- Bei ungerader Ordnung geht jedes y_t mit demselben Gewicht in \bar{y}_t ein; eine gerade Ordnung liefert einen gewogenen gleitenden Durchschnitt.

Beispiel 9.5: Ermittlung der glatten Komponente.

Für die in der Tabelle angegebenen fiktiven Jahresdaten y_t (2. Zeile) werden gleitende Durchschnitte dritter, vierter und fünfter Ordnung berechnet, $\bar{y}_t(3)$ (d.h. $\lambda = 1$), $\bar{y}_t(4)$ und $\bar{y}_t(5)$.

t	1	2	3	4	5	6	7	8	9	10
y_t	5	12	16,00	19,00	21,00	30	39,00	41,00	44,00	48
$\hat{g}_t = \bar{y}_t(3)$		11	15,67	18,67	23,33	30	36,67	41,33	44,33	
$\bar{y}_t(4)$			15,00	19,25	24,375	30	35,625	40,75		
$\hat{m}_t = \bar{y}_t(5)$				14,60	19,60	25,00	30	35,00	40,40	
\hat{k}_t					1,07	-0,93	-1,67	0	1,67	0,93

Ermittlung der glatten Komponente

Beispiel 9.5: Fortsetzung.

Der erste Wert für $\bar{y}(3)$, der Periode $t = 2$ zugeordnet, ist $\bar{y}_2(3) = \frac{1}{3}(5 + 12 + 16) = 11$. Beim gleitenden Durchschnitt 4. Ordnung gehen fünf Werte in den Durchschnitt ein. Der erste Wert, für Periode $t = 3$, ist

$$\bar{y}_3(4) = \frac{1}{4} \left(\frac{1}{2} \cdot 5 + 12 + 16 + 19 + \frac{1}{2} \cdot 21 \right) = 15,00.$$

Mit zunehmendem λ gehen offenbar zwei Effekte einher.

- ① Aktualitätsverlust: Während die Originalreihe bis $t = 10$ läuft, endet $\bar{y}_t(3)$ mit der 9., $\bar{y}_t(4)$ und $\bar{y}_t(5)$ bereits mit der achten Periode.
- ② Die Glättung steigt mit der Ordnung: obwohl die Originalreihe deutliche Zyklen aufweist, verschwinden diese bei $\bar{y}_t(5)$. Daher hat der Zyklus eine Länge von fünf Jahren.

- Der Glättungseffekt gleitender Durchschnitte lässt sich an einer Reihe, die nur aus übereinstimmenden Zyklen besteht, nachvollziehen. Die Zyklen von

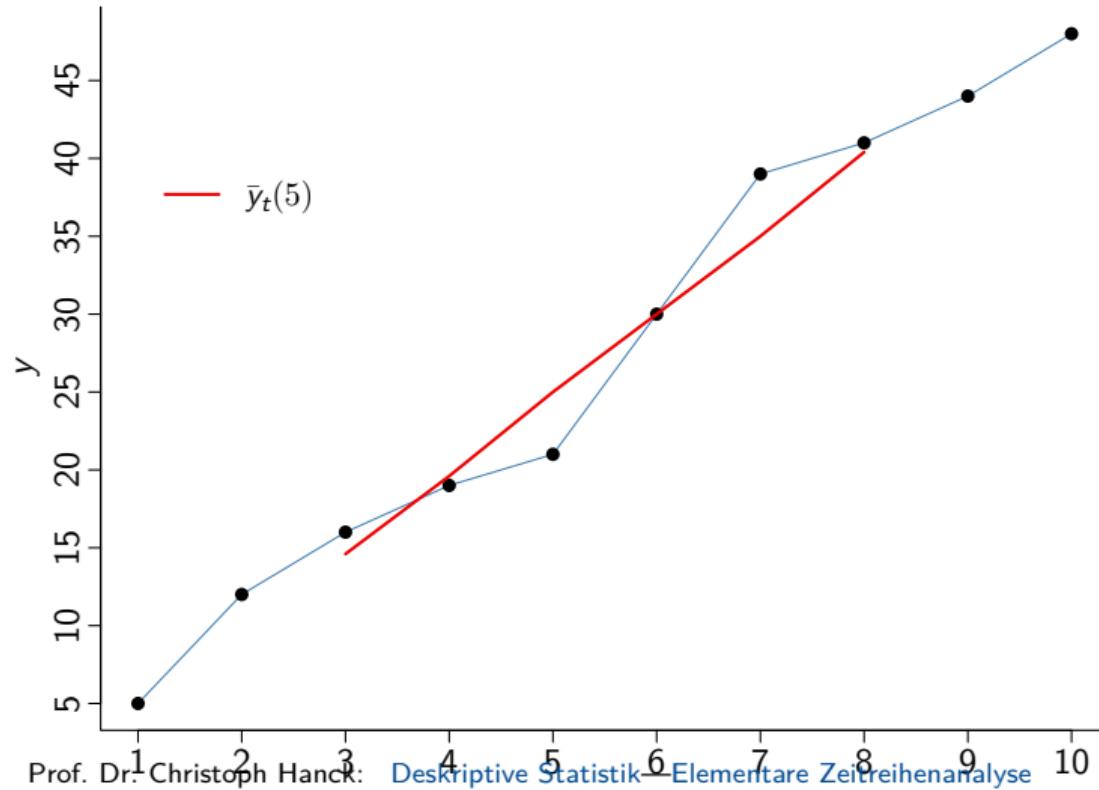
$$y_t : 5, 7, 6, 4, 3, 5, 7, 6, 4, 3$$

usw. sind alle gleich, ihre Länge beträgt 5 Perioden.

- $\bar{y}_t(5)$ ist daher eine zyklenfreie Zeitreihe: Jeder wegfallende Wert wird im Durchschnitt durch einen wertgleichen ersetzt, so dass der gleitende Durchschnitt konstant bleibt.
- Die folgende Abbildung zeigt die Zeitreihe aus Beispiel 9.5 sowie $\bar{y}_t(5)$.

Ermittlung der glatten Komponente

MovingAverage.R



Zeitreihenanalyse 1

Kahoot!

Ermittlung der glatten Komponente

- Welches λ sollte nun konkret für einen gleitenden Durchschnitt gewählt werden? Es lassen sich hier zumindest einige Orientierungshilfen angeben.
- Liegen zyklisch verlaufende Jahresdaten vor, muss λ kleiner als die Zyklendlänge sein. Bei Jahresdaten, die oft bei makroökonomischen Reihen vorliegen, sind daher geringe Ordnungen angemessen, z.B. 3.
- Bei Quartals- oder Monatsdaten ist λ so zu bestimmen, dass Saisonschwankungen von g_t ausgeschlossen bleiben.
- Hierzu muss die Ordnung mit der Länge des Saisonzyklus übereinstimmen, d.h. bei Quartalsdaten ist die Ordnung 4 (also $\lambda = 2$), bei Monatswerten die Ordnung 12 zu wählen.

- Enthält y_t keine Saisonkomponente (etwa bei Jahresdaten), ist die Summe der Restkomponente r_t fast null. Das mit $\bar{y}_t(\text{Ordnung})$ geschätzte \hat{g}_t setzt sich nur aus Trend- und konjunktureller Komponente zusammen und ermöglicht eine Zerlegung in Trend und Zyklus.
- Aus y_t schätzt man entweder mit einem globalen Komponentenmodell oder mit $\bar{y}_t(\text{Ordnung})$ die Trendkomponente. Subtrahieren der geschätzten Trendkomponente \hat{m}_t von der geschätzten glatten Komponente $\hat{g}_t = \bar{y}_t(\text{Ordnung})$ schätzt die **zyklische Komponente** als

$$\hat{k}_t = \hat{g}_t - \hat{m}_t.$$

- In der Tabelle gibt die dritte Zeile die Schätzungen der zyklischen, glatten Komponente $\hat{g}_t = \bar{y}_t(3)$ an; die Differenz von \hat{g}_t und geschätzter zyklenfreier Trendkomponente $\hat{m}_t = \bar{y}_t(5)$ (vorletzte Zeile) liefert die geschätzte konjunkturelle Komponente \hat{k}_t (letzte Zeile).

- Die Analyse saisonaler Schwankungen bei unterjährigen Daten betrachtet die Stärke saisonaler Einflüsse auf y_t .
 - Die zahlreichen Ansätze lassen sich grob in zwei Klassen unterteilen.
- ① Saisoneinflüsse sind jahresunabhängig (**konstante Saisonfigur**). Die jahreszeitlichen Einflüsse ändern sich weder in der zeitlichen Abfolge noch in ihrer Intensität. Nehme zudem an, dass sich Saisoneinflüsse über ein Jahr ausgleichen, sich somit zu null addieren (Normierungsregel).
- ② Eine über die Jahre **variable Saisonfigur**: Jahreszeitlich bedingte Einflüsse variieren über die Jahre.

- Zunächst muss g_t eliminiert werden.
- Für das additive Modell folgt: $y_t^* = y_t - \hat{g}_t \approx s_t + r_t$. (\approx bedeutet „in etwa gleich“ und ist nötig, da \hat{g}_t eine Schätzung für g_t ist.) Wegen der unterjährigen Datenerhebung kennzeichnet t hier Saisonabschnitte wie z.B. Quartale oder Monate.
- Zur Identifikation der Saisoneinflüsse ist jedoch eine Doppelindizierung nützlich, um zwischen Jahr und Saisonabschnitt unterscheiden zu können. Anstelle von y_t^* wird jetzt y_{ij}^* geschrieben, wobei der Index $i = 1, \dots, m$ die Jahre und $j = 1, \dots, n$ die Saisonabschnitte kennzeichnet.
- Die Saisonabschnittswerte ergeben sich als $s_{ij} \approx y_{ij}^* - r_{ij}$. Sie lassen sich in einer Matrix (s. nächste Folie) darstellen. Die Zeilen enthalten die n Saisonabschnittswerte pro Jahr i ; sie ergeben zusammen die Saisonfigur. Die Spalten liefern die Werte eines Saisonabschnitts über alle m Jahre.

$i \backslash j$	1	2	3	\dots	n
1	s_{11}	s_{12}	s_{13}	\dots	s_{1n}
2	s_{21}	s_{22}	s_{23}	\dots	s_{2n}
\vdots	\vdots	\vdots	\vdots	\dots	\vdots
\vdots	\vdots	\vdots	\vdots	\dots	\vdots
m	s_{m1}	s_{m2}	s_{m3}	\dots	s_{mn}
\bar{s}_j	\bar{s}_1	\bar{s}_2	\bar{s}_3	\dots	\bar{s}_n

- Bei konstanter Saisonfigur muss gelten: $s_{ij} = s_j$, d.h. die Saisonwerte sind vom Jahr i unabhängig.

Ermittlung der saisonalen Komponente

- Man bezeichnet s_j als saisontypische Abweichung oder als Saisonveränderungszahl. Schätzt diese, indem s_{ij} für festes j über die Anzahl der Jahre i gemittelt wird, wobei m_j angibt, wie oft die j -te Saison vorkommt.
- Dieser Durchschnitt heißt **roher Saisonkoeffizient** \bar{s}_j . Man erhält

$$\bar{s}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_{ij} \approx \frac{1}{m_j} \sum_{i=1}^{m_j} y_{ij}^* - \frac{1}{m_j} \sum_{i=1}^{m_j} r_{ij}.$$

- Wegen des Fehlens einer systematischen Komponente schwankt r_{ij} regellos um den Wert null, so dass $\sum_{i=1}^{m_j} r_{ij} \approx 0$. Damit gilt für $j = 1, \dots, n$, dass

$$\bar{s}_j \approx \frac{1}{m_j} \sum_{i=1}^{m_j} y_{ij}^*.$$

- Erfüllen die \bar{s}_j die Normierungsregel, gilt $\sum_{j=1}^n \bar{s}_j = 0$; \bar{s}_j ist dann ein Schätzwert für die Saisonveränderungszahl. Ist $\sum_{j=1}^n \bar{s}_j \neq 0$, wird \bar{s}_j korrigiert, indem $\frac{1}{n} \sum_{j=1}^n \bar{s}_j$ von jedem \bar{s}_j subtrahiert wird.
- Die Differenz bezeichnet man als **Saisonkoeffizient** \hat{s}_j :

$$\hat{s}_j = \bar{s}_j - \frac{1}{n} \sum_{j=1}^n \bar{s}_j.$$

Die Saisonkoeffizienten sind Schätzwerte für die Saisonveränderungszahl, die die Normierungsregel einhalten. Diese Vorgehensweise heißt **Phasendurchschnittsverfahren**.

- Um den Saisoneinfluss aus einer Zeitreihe zu eliminieren, betrachte $y_{ij} - \hat{s}_j$.
- Die Restkomponente wird geschätzt durch $\hat{r}_{ij} = y_{ij}^* - \hat{s}_j$.
- Damit kann eine Zeitreihe in ihre geschätzten Komponenten zerlegt werden:

$$y_{ij} = \hat{g}_{ij} + \hat{s}_j + \hat{r}_{ij}.$$

Ermittlung der saisonalen Komponente

- Werden Saisonauwirkungen mit g_t intensiver, muss mit einer variablen Saisonfigur gearbeitet werden. Modelliere z.B. den Saisoneinfluss als Vielfaches von g_{ij} : $s_{ij} = \lambda_{ij} g_{ij}$. Einsetzen in das additive Modell liefert

$$y_{ij} = (1 + \lambda_{ij})g_{ij} + r_{ij}.$$

- $(1 + \lambda_{ij})$ heißt **Saisonfaktor** oder Saisonmultiplikator und wird ebenfalls mit s_{ij} bezeichnet. Ist $s_{ij} > 1$, wirkt der Saisoneinfluss niveausteigernd; für $s_{ij} < 1$ entsprechend niveausenkend.
- Um s_{ij} zu schätzen, ist zunächst g_{ij} aus y_{ij} zu eliminieren. Dies geschieht durch Division der y_{ij} durch die geschätzte glatte Komponente \hat{g}_{ij} :

$$y_{ij}^* = \frac{y_{ij}}{\hat{g}_{ij}} \approx s_{ij} + \frac{r_{ij}}{\hat{g}_{ij}}, \quad y_{ij}^* : \text{bereinigter Zeitreihenwert.}$$

- Nehme an, dass die s_{ij} für gleiche Saisonabschnitte von Jahr zu Jahr konstant bleiben: $s_{ij} = s_j$. Analog zur konstanten Saisonfigur erhält man den rohen Saisonfaktor jetzt als

$$\bar{s}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_{ij} = \frac{1}{m_j} \sum_{i=1}^{m_j} \frac{y_{ij}}{\hat{g}_{ij}}.$$

- Für n Saisonabschnitte erhält man n Saisonfaktoren \bar{s}_j . Würden die Saisonphasen keinen Einfluss auf y_{ij} ausüben, müssten alle n Saisonfaktoren gleich eins und ihre Summe gleich n sein.
- Gilt dies auch für die Schätzwerte, also $\sum_{j=1}^n \bar{s}_j = n$, ist \bar{s}_j ein Schätzwert für den Saisonfaktor; wenn nicht, muss \bar{s}_j korrigiert werden. Der korrigierte Schätzwert ergibt sich als

$$\hat{s}_j = \frac{\bar{s}_j}{\frac{1}{n} \sum \bar{s}_j}.$$

- Der Saisoneinfluss ist als Anteil von g_{ij} zu interpretieren; $\hat{s}_j = 1,125$ besagt z.B. daher, dass der Saisoneinfluss eine Erhöhung des entsprechenden Zeitreihenwertes um 12,5% der glatten Komponente verursacht.

Beispiel 9.6: Ermittlung der saisonalen Komponente.

Das in Quartalsdaten vorliegende BIP der 19 Euroraum-Länder wird zerlegt in die vier Zeitreihenkomponenten (siehe folgende Tabelle). Die Zeitreihe läuft von 2013.1 bis 2016.4. (Bei Quartalsdaten notiert man erst das Jahr, dann das Quartal, siehe erste Spalte). Die Daten des BIP liegen in Mrd. € zu Preisen des Jahres 2005 vor.

Beispiel 9.6: Fortsetzung.

(1)	(2) <i>t</i>	(3) \hat{y}_t	(4) \hat{m}_t	(5) \hat{g}_t	(6) \hat{k}_t	(7) y_{ij}^*	(8) $\hat{y}_{ij} - \hat{s}_j$	(9) \hat{r}_{ij}
2013.1	1	2405,36	2437,78	-	-	-	2458,58	-
2013.2	2	2486,63	2458,41	-	-	-	2481,65	-
2013.3	3	2488,33	2479,05	2494,23	15,18	-5,90	2497,50	3,27
2013.4	4	2563,71	2499,69	2508,34	8,65	55,37	2506,30	-2,04
2014.1	5	2471,14	2520,32	2520,94	0,62	-49,81	2524,36	3,42
2014.2	6	2533,73	2540,96	2535,72	-5,24	-2,00	2528,75	-6,97
2014.3	7	2542,06	2561,60	2553,28	-8,32	-11,22	2551,23	-2,06
2014.4	8	2628,21	2582,24	2574,08	-8,16	54,13	2570,79	-3,28
2015.1	9	2547,13	2602,87	2596,92	-5,95	-49,79	2600,35	3,43
2015.2	10	2624,09	2623,51	2621,01	-2,50	3,08	2619,11	-1,90
2015.3	11	2634,45	2644,15	2643,25	-0,89	-8,80	2643,61	0,36
2015.4	12	2728,55	2664,78	2664,23	-0,55	64,32	2671,13	6,90
2016.1	13	2624,73	2685,42	2683,23	-2,19	-58,51	2677,95	-5,28
2016.2	14	2714,32	2706,06	2698,91	-7,14	15,41	2709,35	10,44
2016.3	15	2696,21	2726,69	-	-	-	2705,38	-
2016.4	16	2792,22	2747,33	-	-	-	2734,81	-

Note:

$$y_{ij}^* = y_{ij} - \hat{g}_{ij} \quad \hat{r}_{ij} = y_{ij}^* - \hat{s}_{ij}$$

Beispiel 9.6: Fortsetzung.

Die Wachstumskomponente ist als linearer Trend $m_t = \alpha + \beta t$ spezifiziert; die übrigen drei Komponenten k_t , s_t und r_t werden zusammengefasst zu u_t . Das additive Modell ist $y_t = m_t + u_t = \alpha + \beta t + u_t$. Die OLS-Schätzung lautet

$$\hat{y}_t = \hat{m}_t = 2443,22 + 19,44t, \quad t = 1, \dots, 16$$

(Spalte 4). Die glatte Komponente wird zwecks Ausschaltung saisonaler Einflüsse mit einem gleitenden Durchschnitt 4. Ordnung geschätzt (Spalte 5). Wegen der Ordnung gehen an den Rändern jeweils zwei Werte verloren; t läuft daher von 3 bis 14.

Beispiel 9.6: Fortsetzung.

Die Differenz aus glatter Komponente und Trend ergibt die konjunkturelle Komponente. Deren Schätzungen $\hat{k}_t = \hat{g}_t - \hat{m}_t$ zeigt Spalte (6). Die Bereinigung von y_t mit der glatten Komponente \hat{g}_t liefert (nach Neuindizierung) die y_{ij}^* zur Ermittlung der Saisonkomponente (Spalte (7)). Bei konstanter Saisonfigur wird hieraus der rohe Saisonkoeffizient berechnet, der eventuell zu korrigieren ist. Die y_{ij}^* werden in folgender Tabelle wiederholt. Die vorletzte Zeile enthält die \bar{s}_j , also die Durchschnitte der zugehörigen Spalten.

Beispiel 9.6: Fortsetzung.

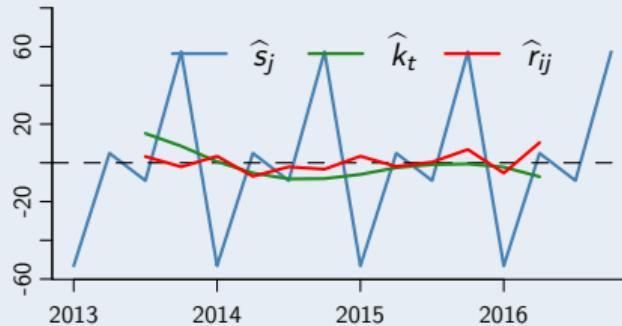
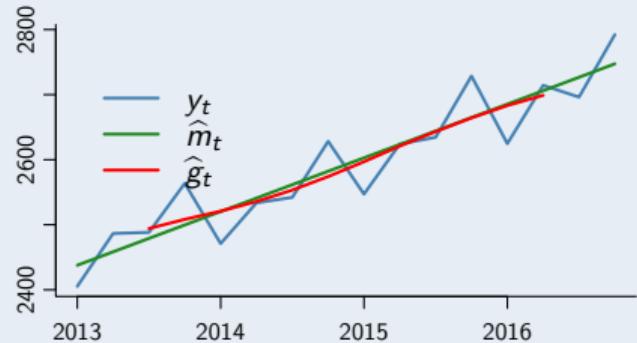
$i \backslash j$	1	2	3	4	
1	-	-	-5,9	55,37	
2	-49,81	-2	-11,22	54,13	
3	-49,79	-3,08	-8,8	64,32	
4	-58,51	15,41	-	-	
\bar{s}_j	-52,7	5,5	-8,64	57,94	$\sum_{j=1}^4 \bar{s}_j \approx 2,09$
\hat{s}_j	-53,23	4,97	-9,16	57,42	$\frac{1}{4} \sum_{j=1}^4 \bar{s}_j = \approx 0,5233$

Da die Summe der \bar{s}_j 2,09 beträgt, müssen sie mit $2,09 : 4 \approx 0,5233$ korrigiert werden. Die letzte Zeile zeigt die Saisonkoeffizienten $\hat{s}_j = \bar{s}_j - 0,5233$.

Beispiel 9.6: Fortsetzung.

Spalte (8) der Ausgangstabelle gibt das saisonbereinigte BIP an; Spalte (9) liefert die Restkomponente. Wegen der Bereinigung der rohen Saisonkoeffizienten \bar{s}_j ist die Summe der Restkomponente ungleich null. Da jedes \bar{s}_j um 0,5233 verringert wurde, muss die Summe der Restkomponente hier $12 \cdot 0,5233 \approx 6.2796$ betragen.

Beispiel 9.6: Fortsetzung.



DecomposeTimeSeries.R

```
load("Daten/GDP_euro_area.rda") # Quelle: Quandl.com
series <- window(GDP / 1000, start = c(2013, 1), end = c(2016, 4))

dc <- decompose(series) # Zeitreihe "zerlegen"

# Zyklische und Trend Komponente
reg <- lm(series ~ I(1:16))
(trend <- predict(reg))

##          1         2         3         4         5         6         7         8
## 2437.777 2458.414 2479.051 2499.688 2520.325 2540.962 2561.599 2582.236
##          9        10        11        12        13        14        15        16
## 2602.873 2623.510 2644.147 2664.784 2685.421 2706.058 2726.695 2747.332

dc$trend-trend

##          Qtr1        Qtr2        Qtr3        Qtr4
## 2013       NA       NA 15.1793153  8.6527481
## 2014  0.6196071 -5.2392501 -8.3159173 -8.1595695
## 2015 -5.9532730 -2.4990652 -0.8930024 -0.5499546
## 2016 -2.1870044 -7.1448178       NA       NA
```

```
# Gleitender Durchschnitt
dc$trend
##          Qtr1      Qtr2      Qtr3      Qtr4
## 2013       NA        NA 2494.230 2508.341
## 2014 2520.945 2535.723 2553.283 2574.076
## 2015 2596.920 2621.011 2643.254 2664.234
## 2016 2683.234 2698.913       NA        NA

# Saisonale Komponente
dc$seasonal
##          Qtr1      Qtr2      Qtr3      Qtr4
## 2013 -53.226479  4.973303 -9.163476 57.416652
## 2014 -53.226479  4.973303 -9.163476 57.416652
## 2015 -53.226479  4.973303 -9.163476 57.416652
## 2016 -53.226479  4.973303 -9.163476 57.416652
```

```
# Rest
dc$random

##           Qtr1       Qtr2       Qtr3       Qtr4
## 2013        NA        NA  3.2670614 -2.0435620
## 2014  3.4197876 -6.9683895 -2.0554849 -3.2843232
## 2015  3.4324089 -1.8975332  0.3609314  6.9003930
## 2016 -5.2796886 10.4384305        NA        NA
```

Zeitreihenanalyse 2

Kahoot!

- Diese Vorlesung behandelte Grundlagen der Zeitreihenanalyse. Hiermit können Entwicklungstendenzen von zeitlich geordneten Datensätzen untersucht werden, so dass diese Thematik ebenfalls sehr praxisrelevant für u.a. die Wirtschaftswissenschaften ist.
- Nachbearbeitung: Kapitel 6 des Buches von Prof. Assenmacher.