

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.

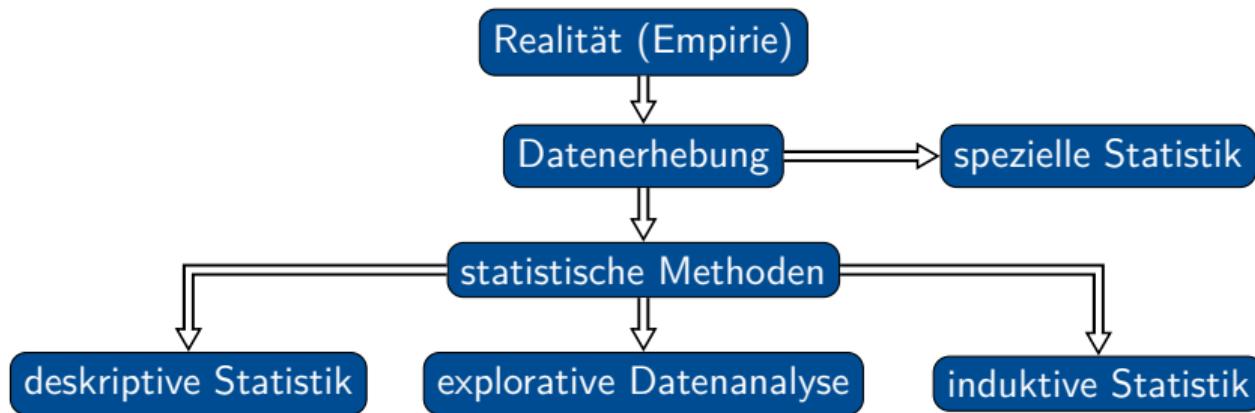
ANZEIGE

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:



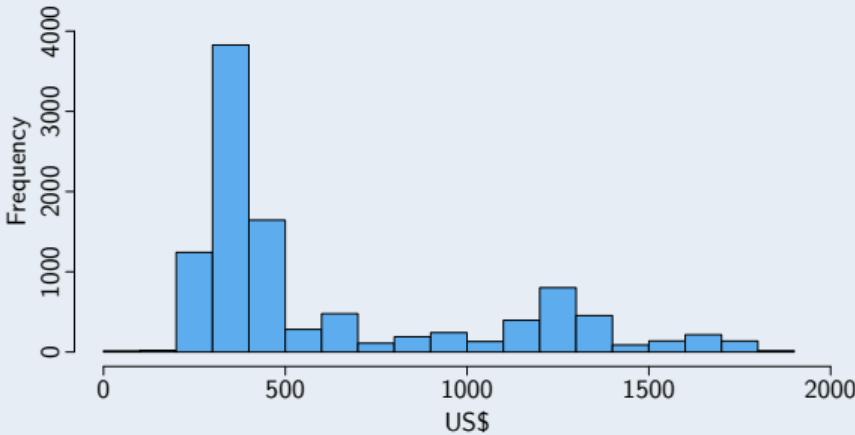
- **Deskriptive** (beschreibende) **Statistik**: Aufbereitung der statistischen Daten.
- Ziel: Übersicht mittels tabellarischer und grafischer Repräsentationen sowie geeigneter Kenngrößen.
- Vergleich von Datensätzen und Ableitung von Handlungssimplikationen.

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.

- 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 ($m \approx n$).

Datengewinnung

- 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.

Urliste und Klassierung

- 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	

Urliste und Klassierung

- 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

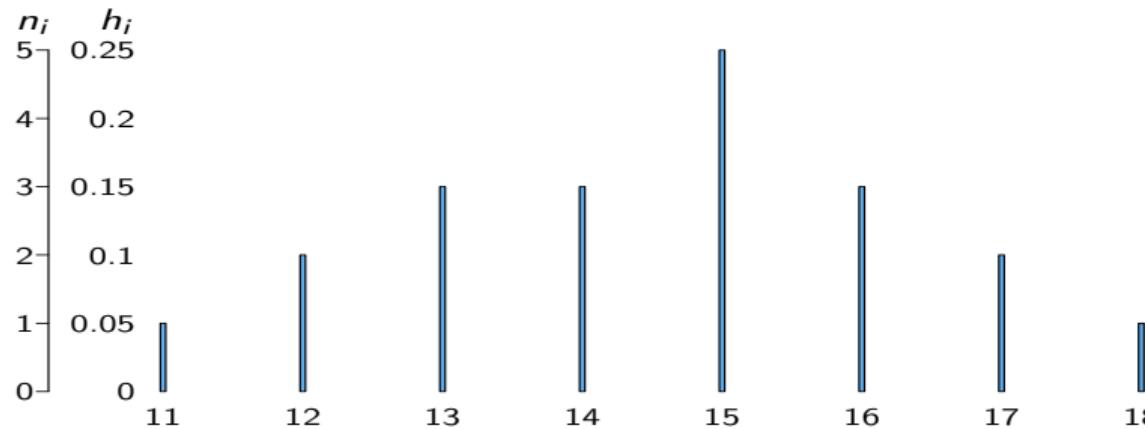
Quelle: Statistisches Bundesamt: DESTATIS

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^* .

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

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

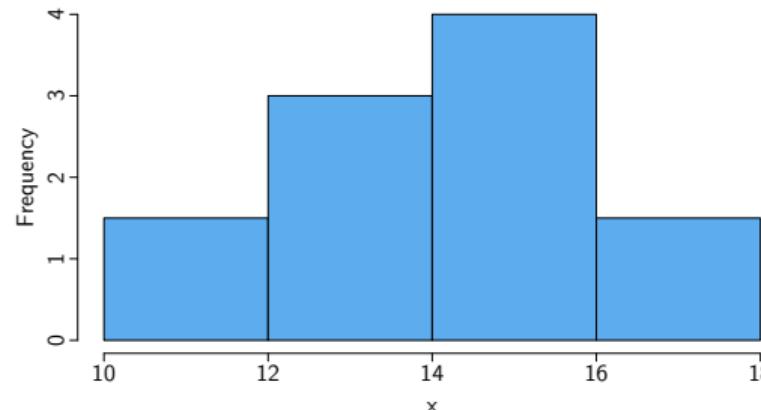
Histogramms.R

Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

Für unser Beispiel ergibt sich:

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

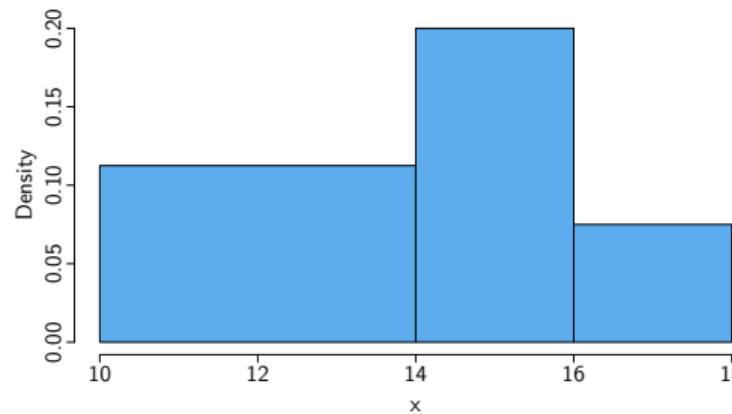


Absolute und relative Häufigkeitsverteilungen

Tabellen und Grafiken

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

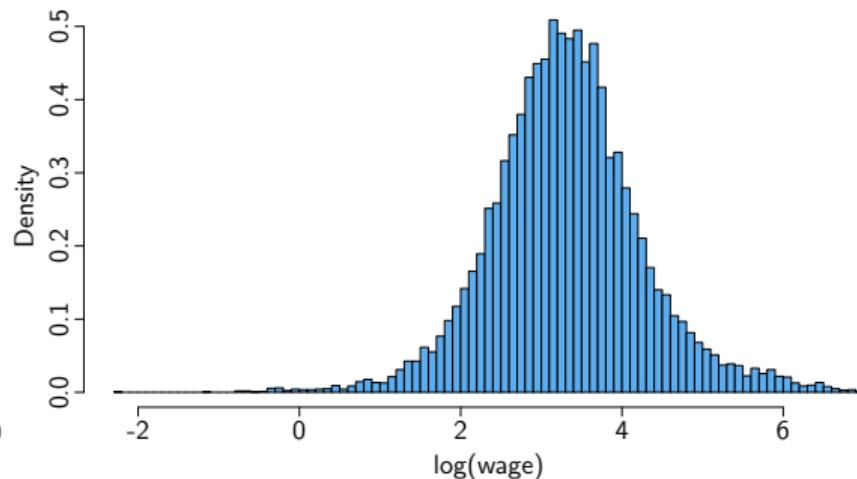
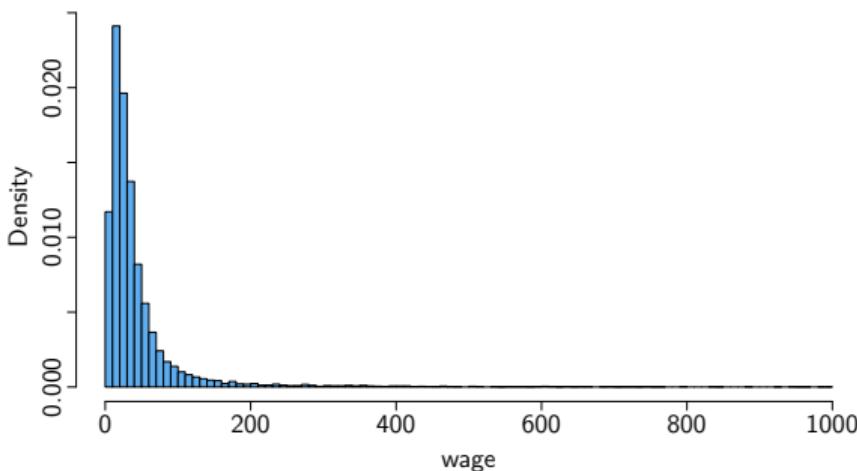
$x'_{k-1} < X \leq x'_k, k = 1, 2, 3$	n_k	n_k^*	h_k	h_k^*
(10,14]	9	2,25	0,45	0,1125
(14,16]	8	4,00	0,40	0,2000
(16,18]	3	1,50	0,15	0,0750



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

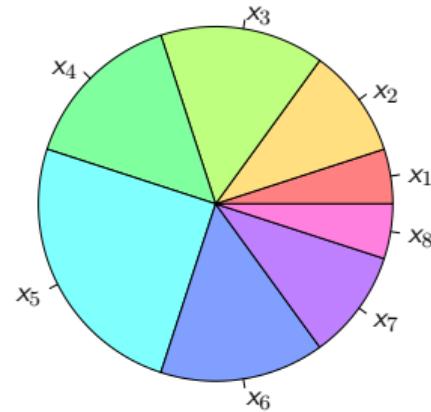
Absolute und relative Häufigkeitsverteilungen

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°

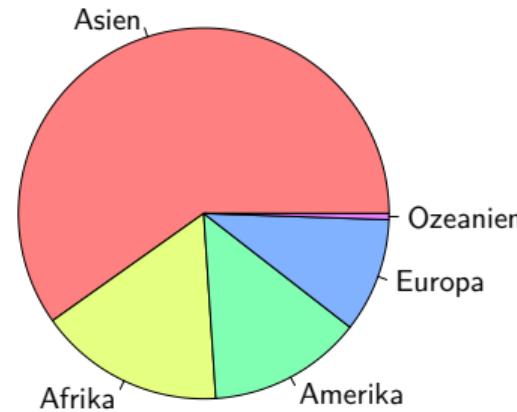
```
pie(table(x), labels = paste0("$x_",$ 1:8, "$"), col = rainbow(8, s = 0.5))
```



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))
```

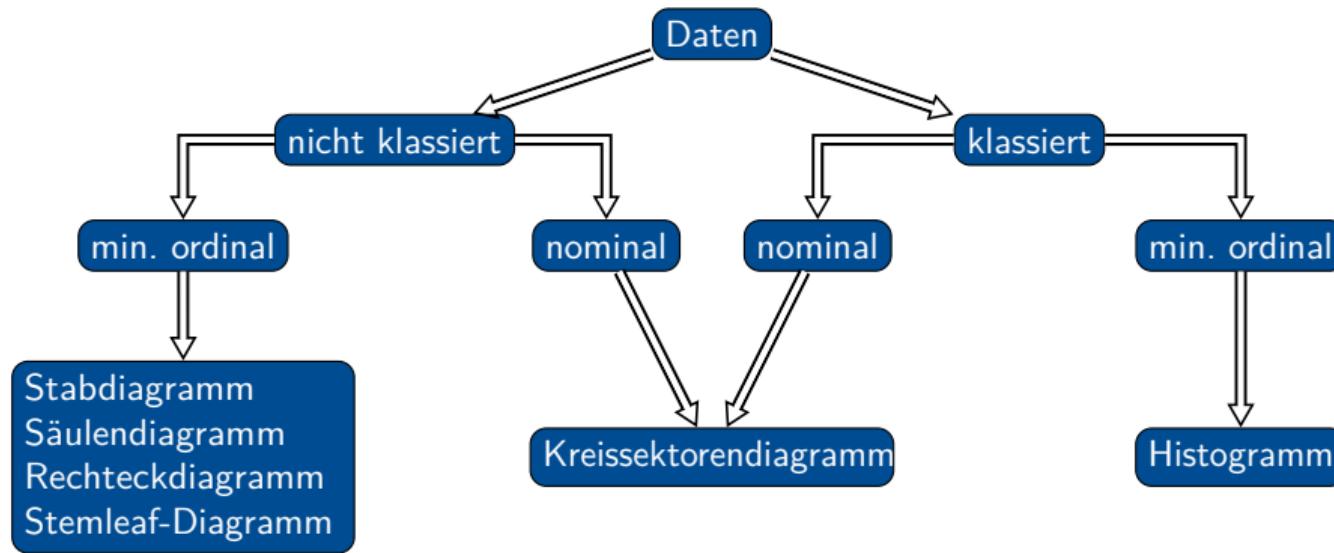


Wie auch hier ersichtlich, sind Kreisdiagramme oft suboptimal. Einige Empfehlungen und Beispiele, wie man es nicht machen sollte.

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

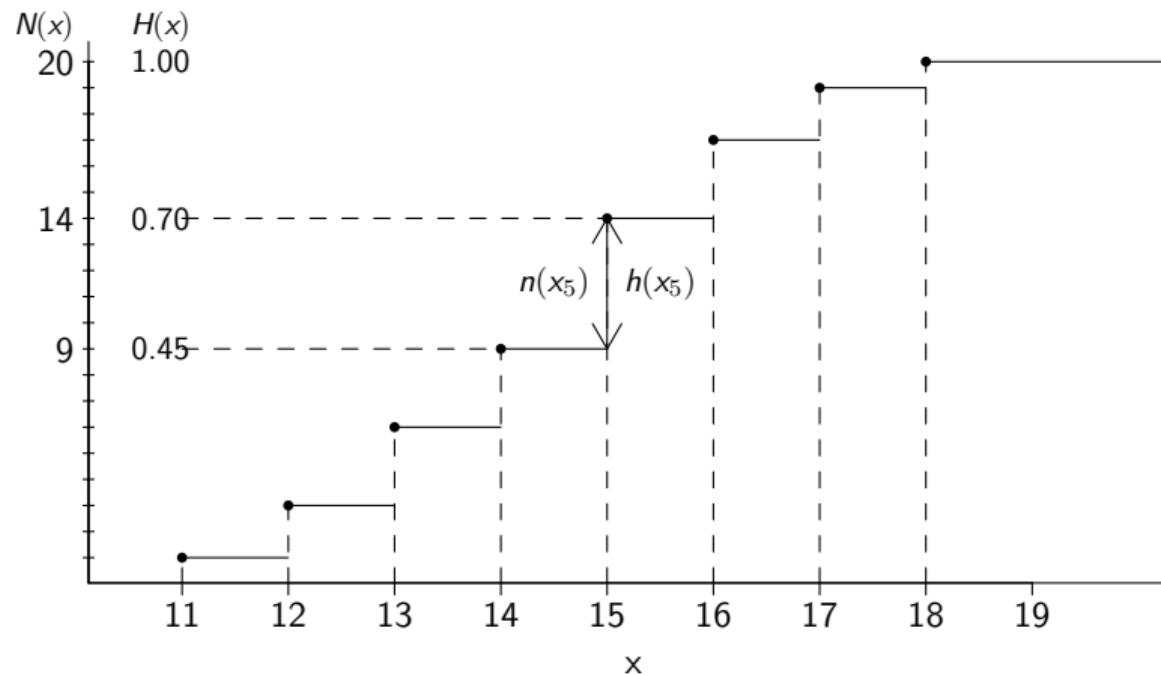
Für unser Beispiel ergeben sich folgende Häufigkeitssummen:

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

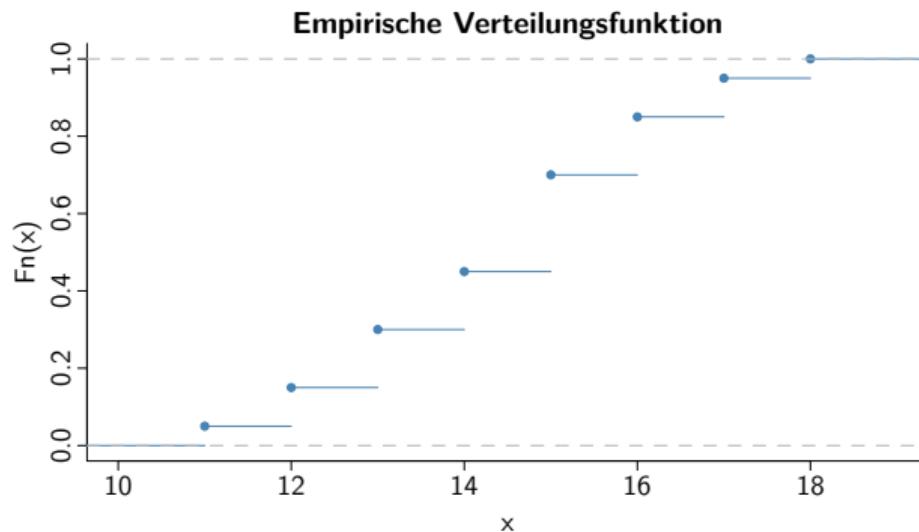
Absolute und relative Häufigkeitssummenfunktion - in **R**

```
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 Häufigkeitsverteilungen

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

$$\begin{aligned}N(a < X \leq b) &= N(b) - N(a) \quad \text{bzw.} \\H(a < X \leq b) &= H(b) - H(a).\end{aligned}$$

- 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

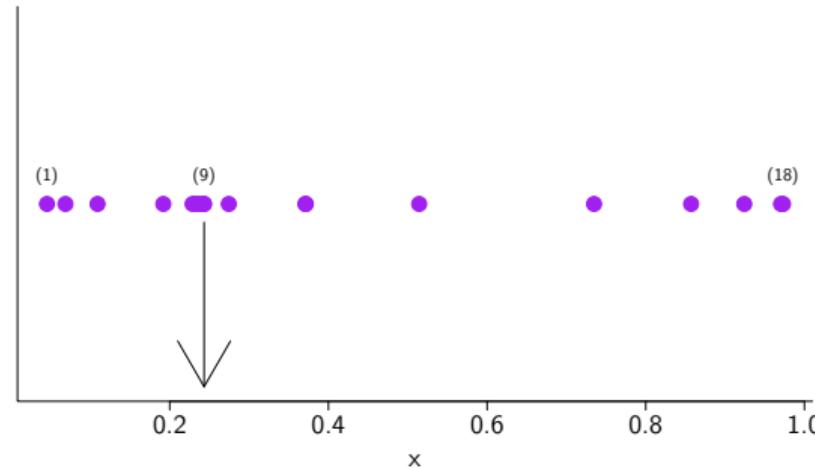
# per "Hand":
(g <- floor(n * p))
## [1] 5

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 .

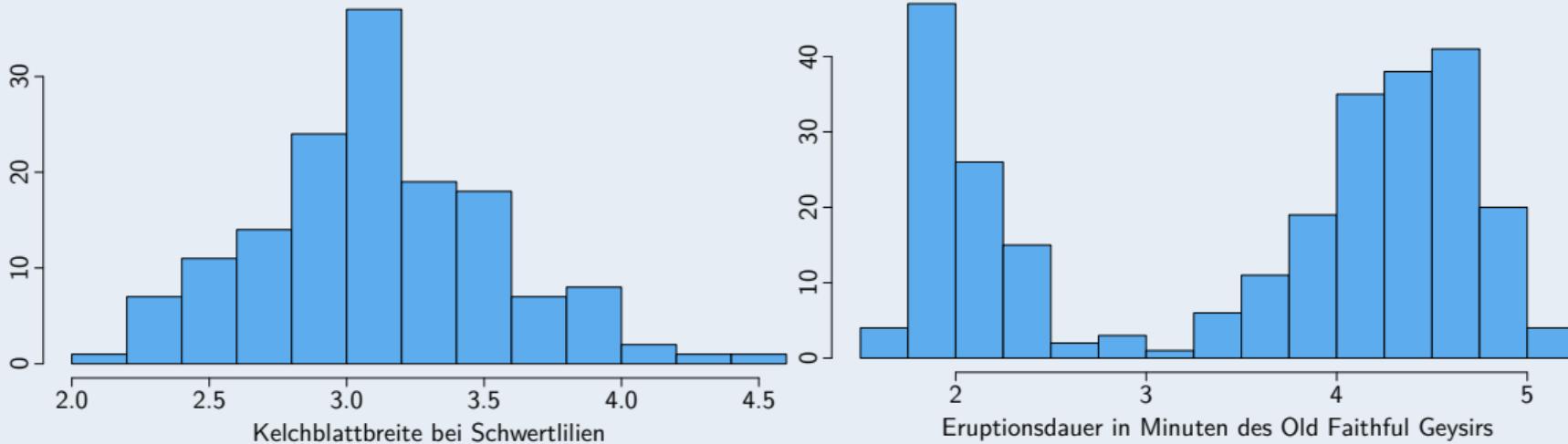
Es gilt: $x_M = x_{i^*}$, wobei i^* der Index der Ausprägungen mit der größten absoluten Häufigkeit ist.

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

Lageparameter

Modus x_M

Beispiel 3.2: Unimodale vs. bimodale Verteilung.

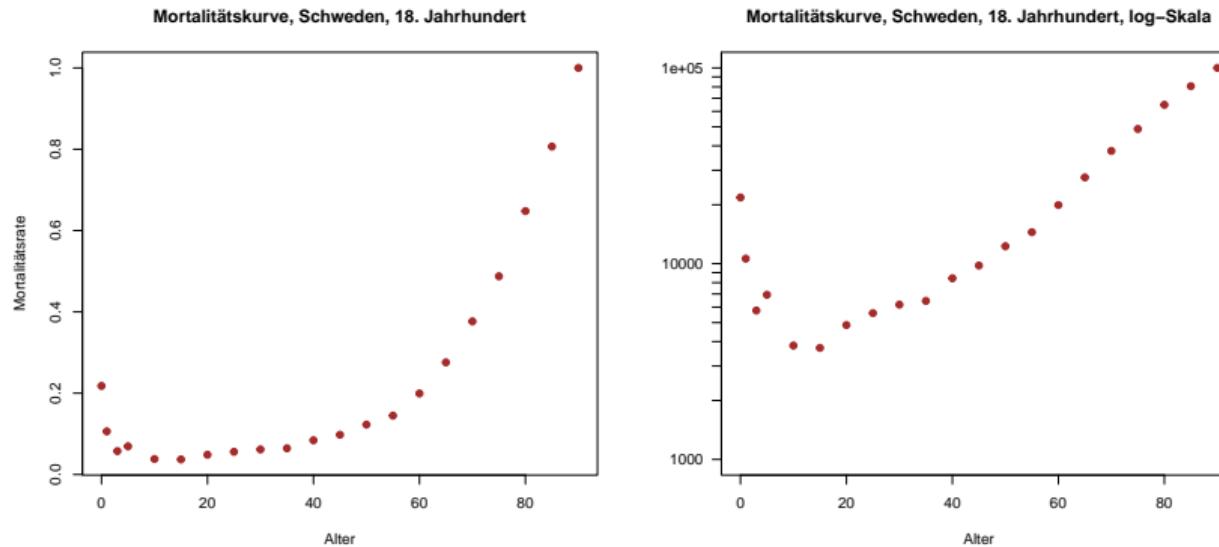


Unimodale (links) im Vergleich mit einer bimodalen Verteilung.

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

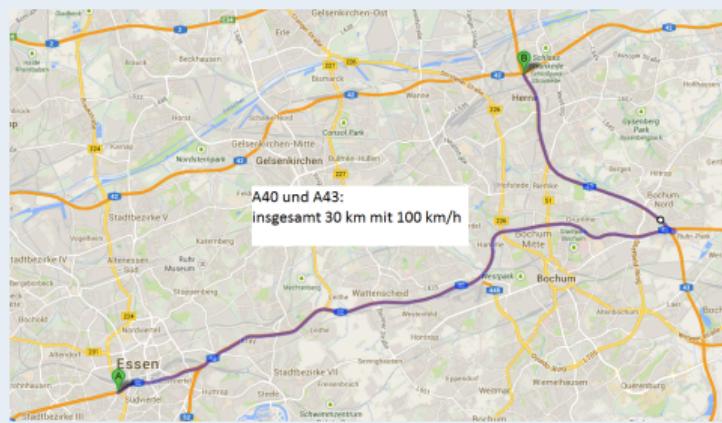
Harmonisches Mittel \bar{x}_H

- 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:

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

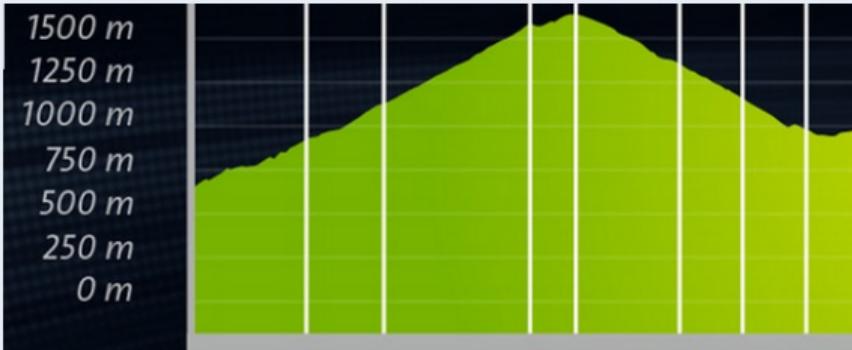
HarmonicMean.R

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

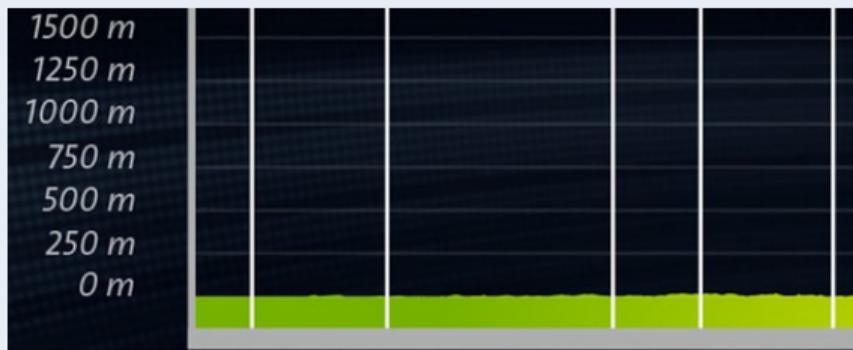
Lageparameter

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!

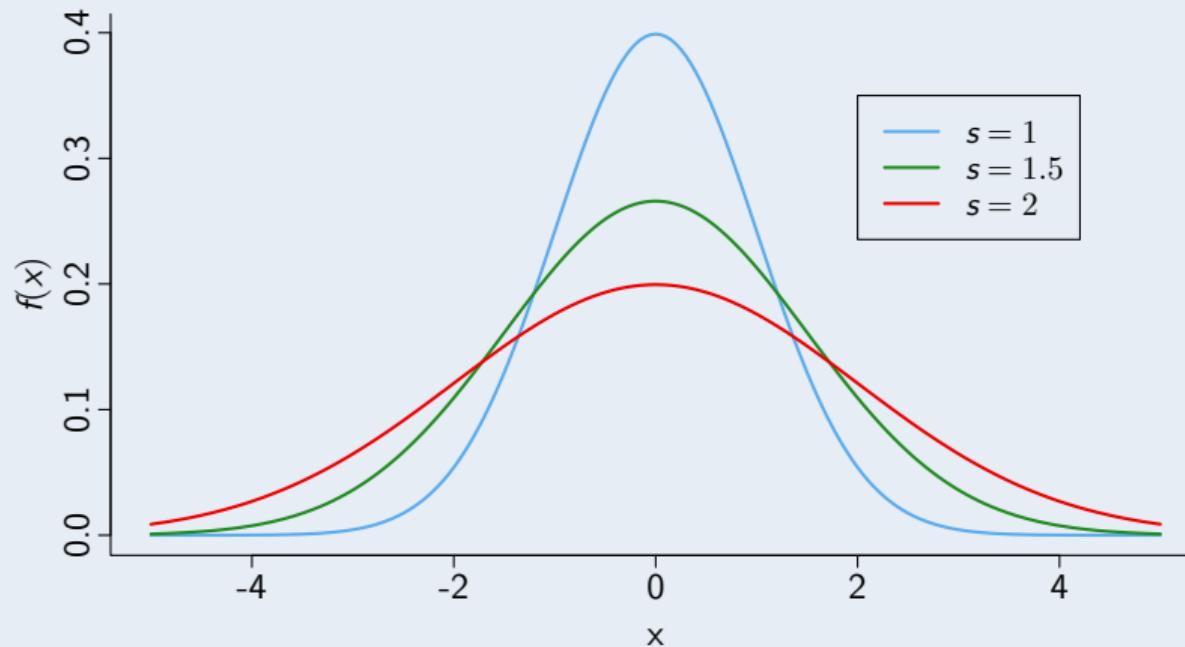
- Wir haben Lageparameter eindimensionaler Datensätze besprochen. Diese Parameter beschreiben einen Datensatz kompakt und machen verschiedene Datensätze vergleichbar.
- Das Verständnis des hier behandelten Stoffes ist für den weiteren Verlauf der Vorlesung elementar! Arbeiten Sie daher nicht nur die 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

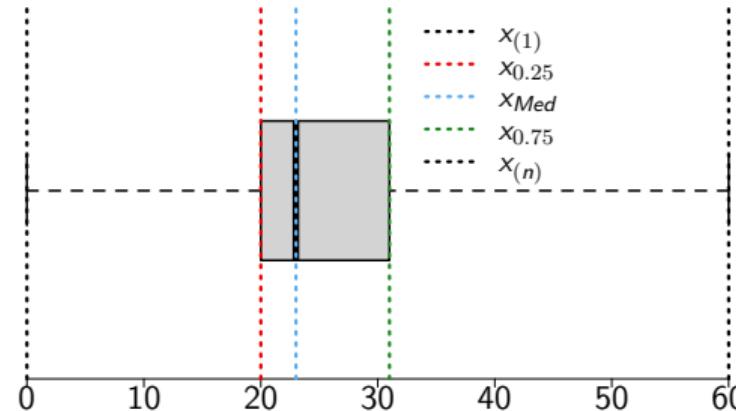
$$x_{0,75} = x_{(15)}, \quad \text{da} \quad np = 20 \cdot 0,75 = 15.$$

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](#)

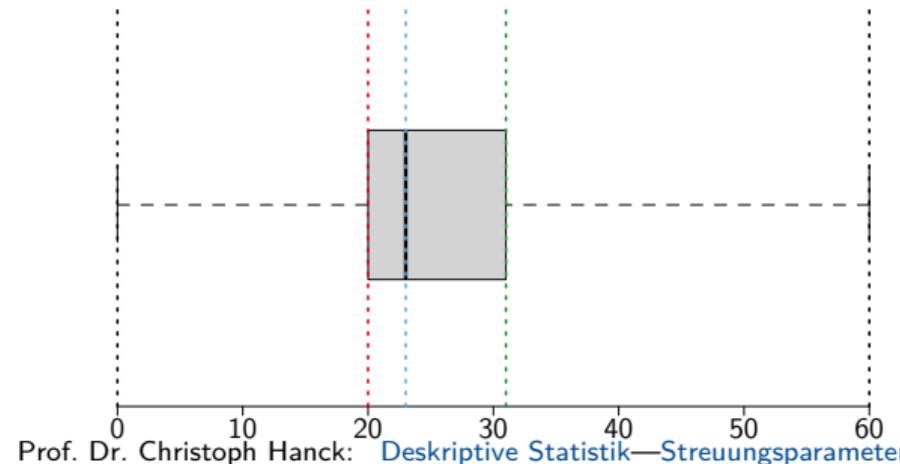
Absolute Streuungsparameter

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)
```

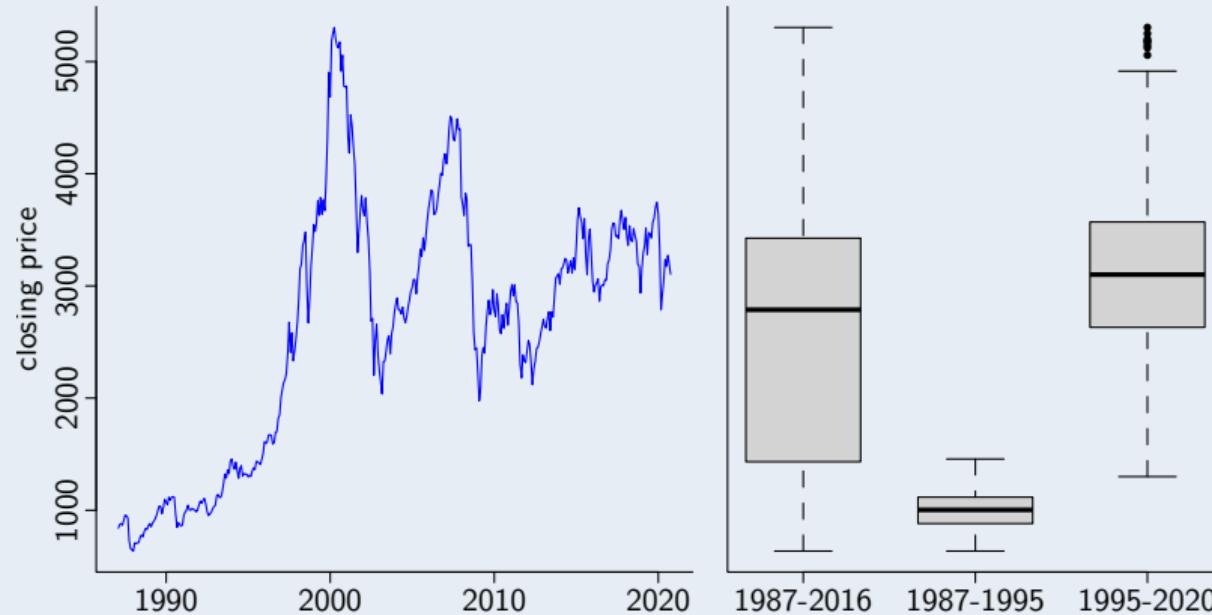
```
bxpl_data <- boxplot(x, horizontal = T, range = 0, cex.axis = 1.2, frame = F)  
abline(v = bxpl_data$stats, lty = 3, lwd = 2,  
       col = c("black", "red2", "steelblue2", "forestgreen"))
```



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).

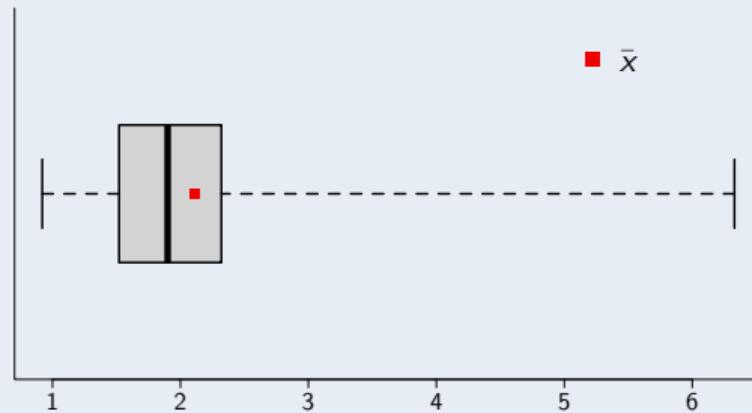
$$\bar{x} = 2.11$$

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

$$R = 6.33 - 0.92 = 5.41$$

$$Q = 2.32 - 1.52 = 0.8$$

$$s = 0.97$$



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

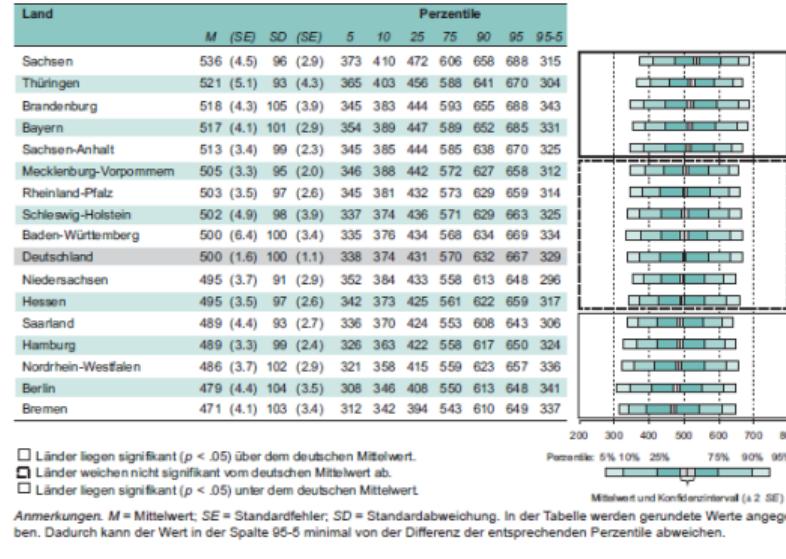
Kahoot!

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

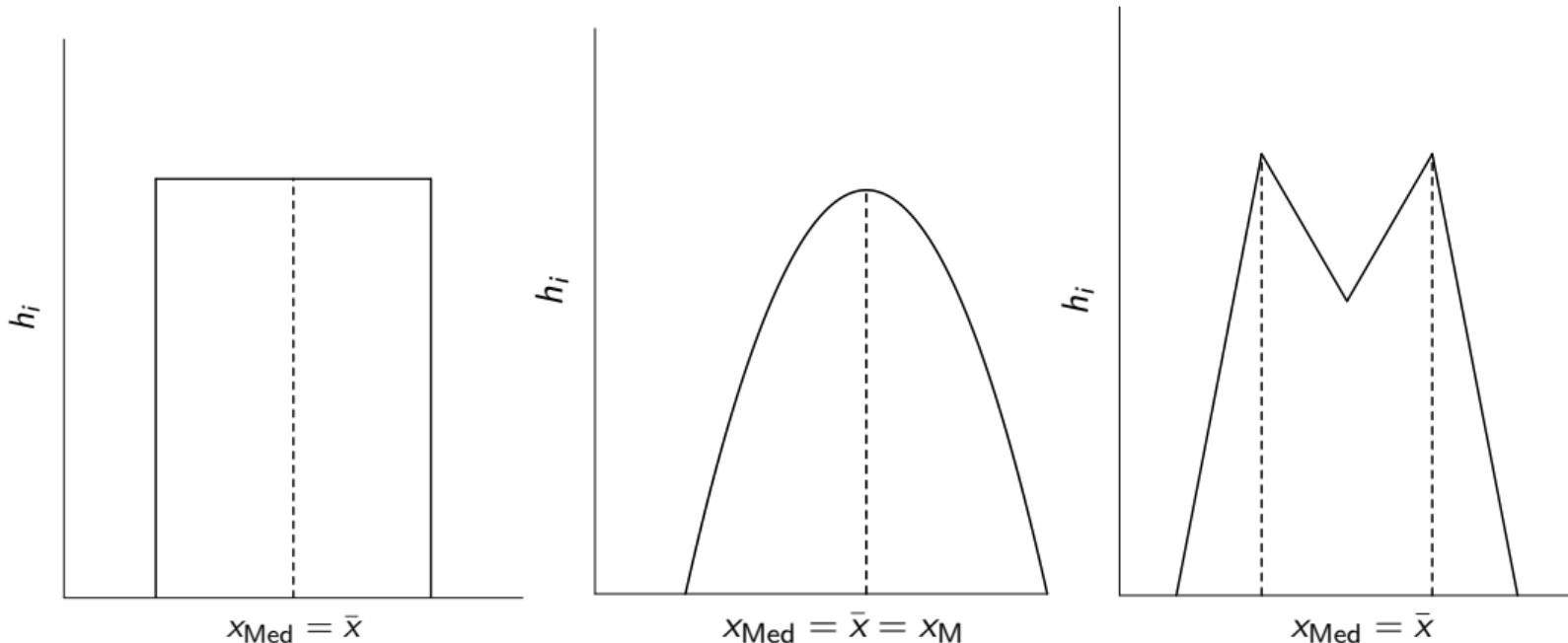
Das Konzept der Kurtosis

- 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

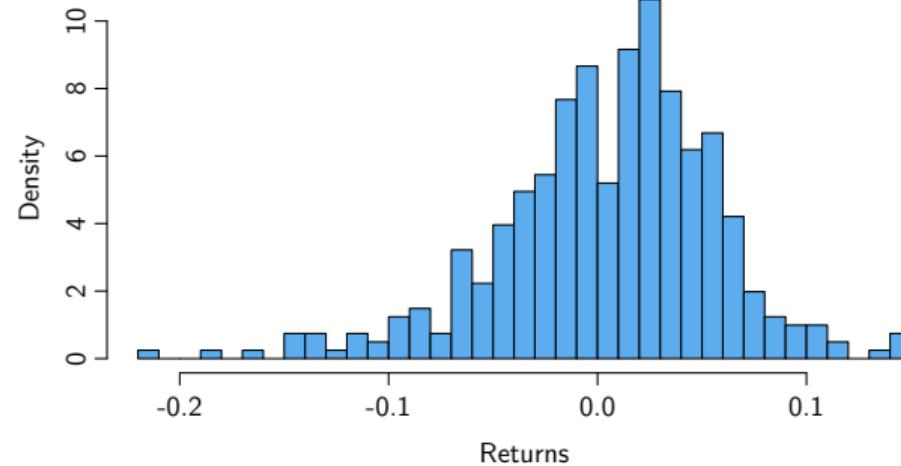
Hier sind drei symmetrische Häufigkeitsverteilungen wiedergegeben. Nur die zweite besitzt einen eindeutigen Modus.



- 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

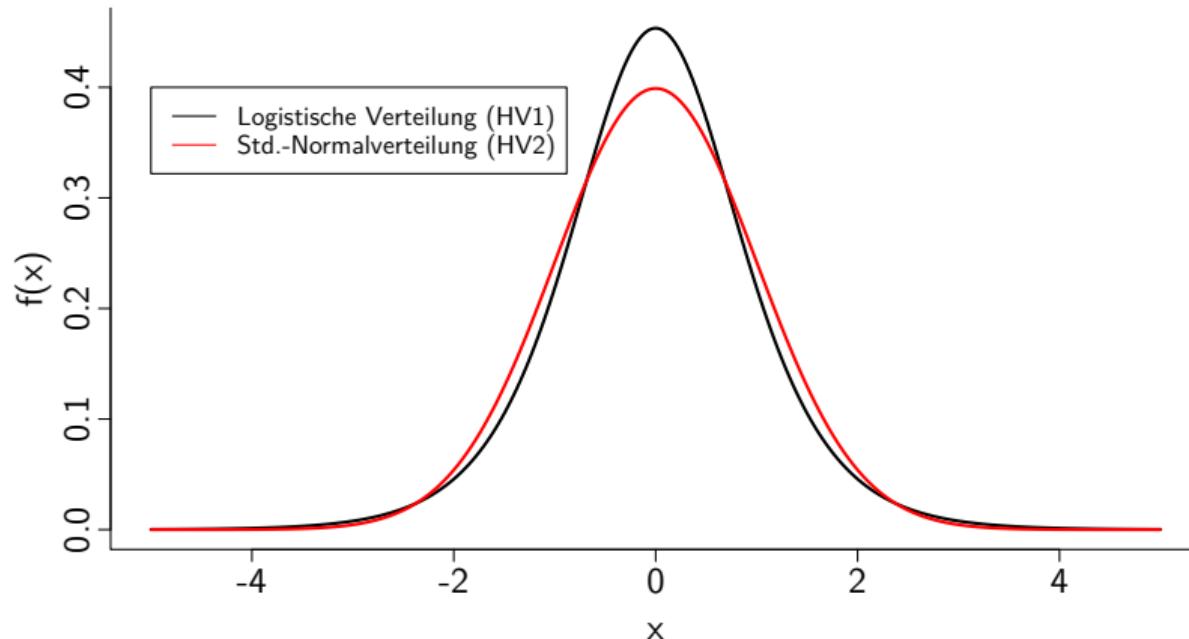


Renditen des Euro Stoxx 50 (01/1987 - 10/2020, Daten: Yahoo Finance)

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

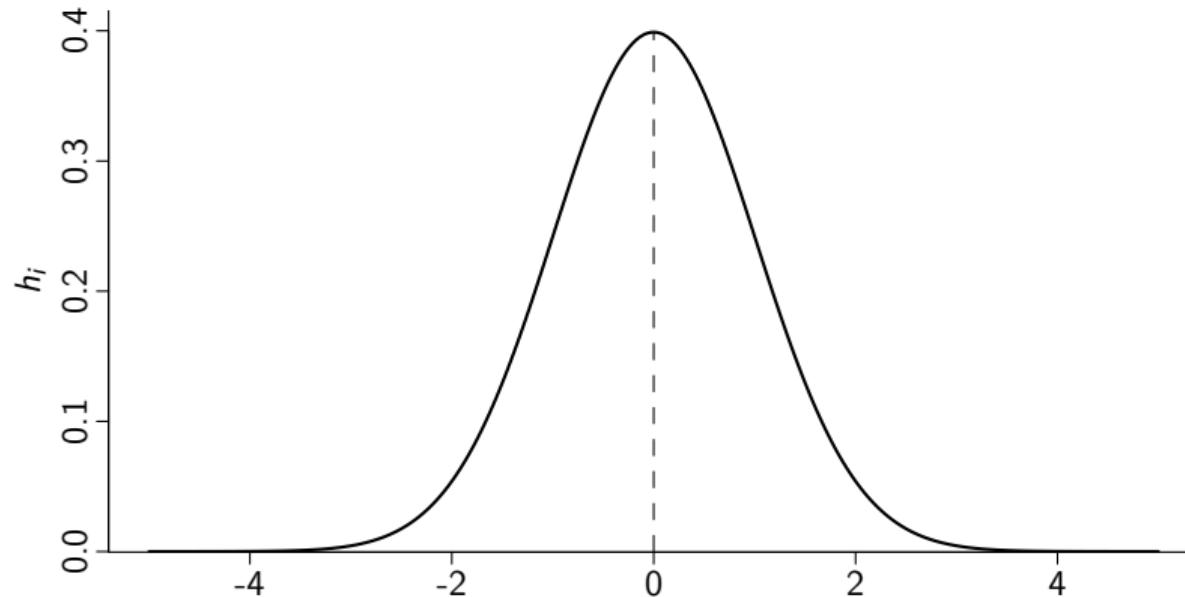
Kurtosisparameter

- 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

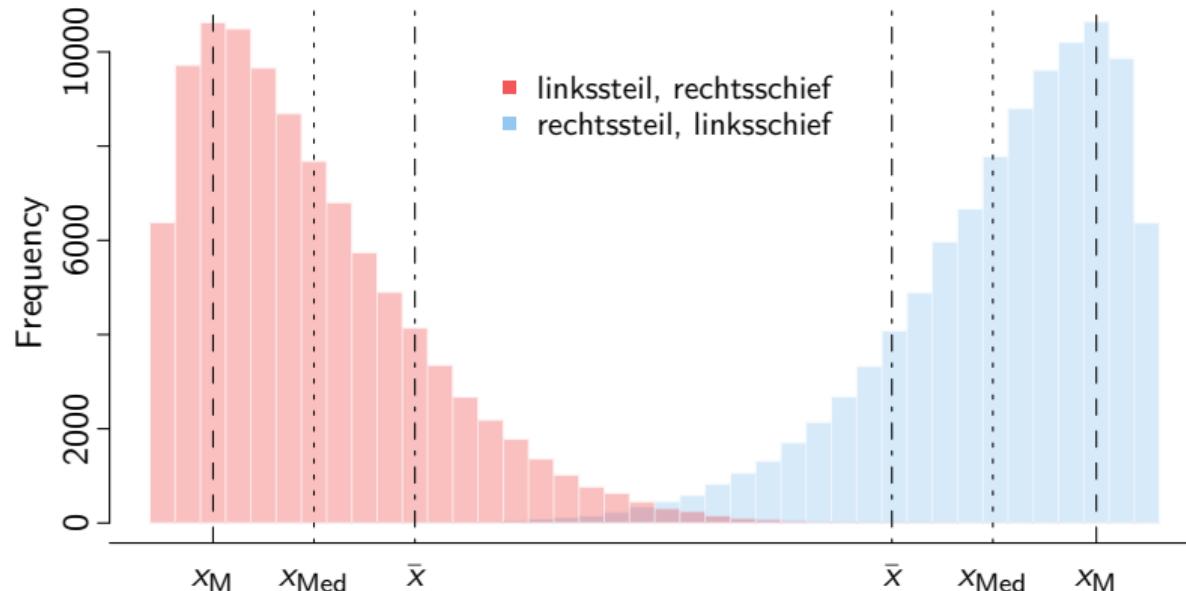
(thetaK <- mean((x - mean(x))^4))
## [1] 24.40873

(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).

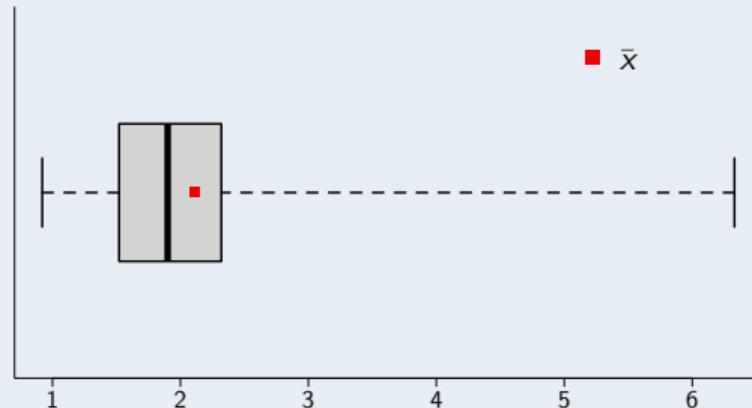
$$\bar{x} = 2.11$$

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

$$R = 6.33 - 0.92 = 5.41$$

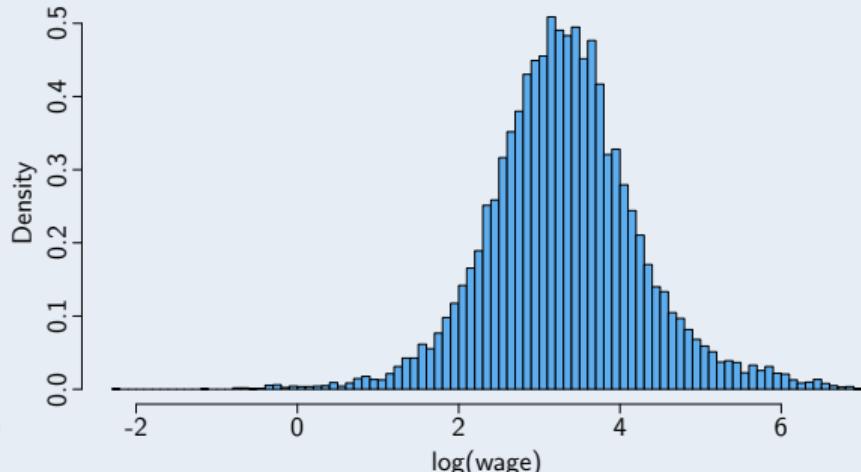
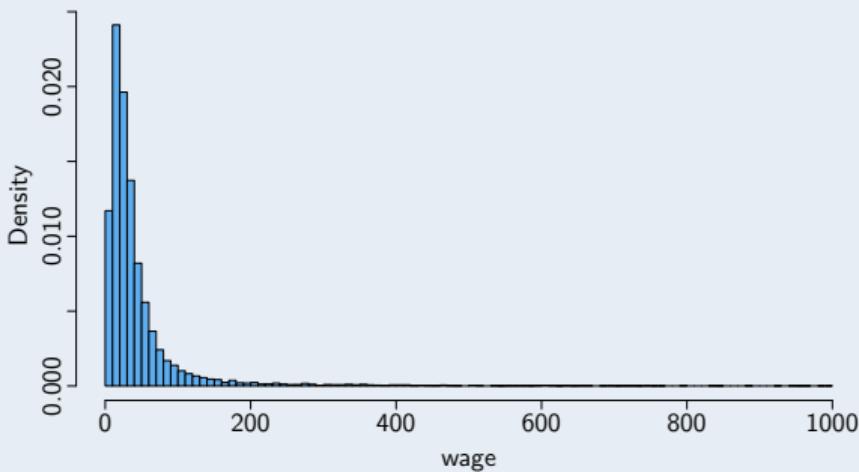
$$Q = 2.34 - 1.505 = 0.835$$

$$s = 0.97$$



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

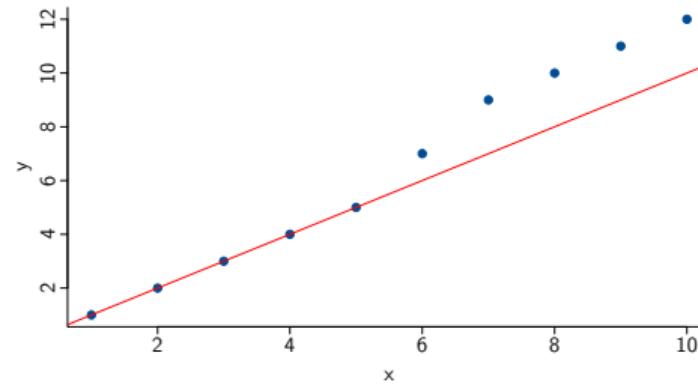
- Stimmen die relativen Häufigkeitsverteilungen zweier Datensätze überein, gilt das auch für ihre p -Quantile $x_p = y_p$.
- In einem Quantil-Quantil-Diagramm (kurz **Q-Q-Plot**) werden die Quantilspaare (x_p, y_p) für verschiedene p als Punkte in ein Koordinatensystem eingetragen.
- Liegen alle Punkte 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")
```

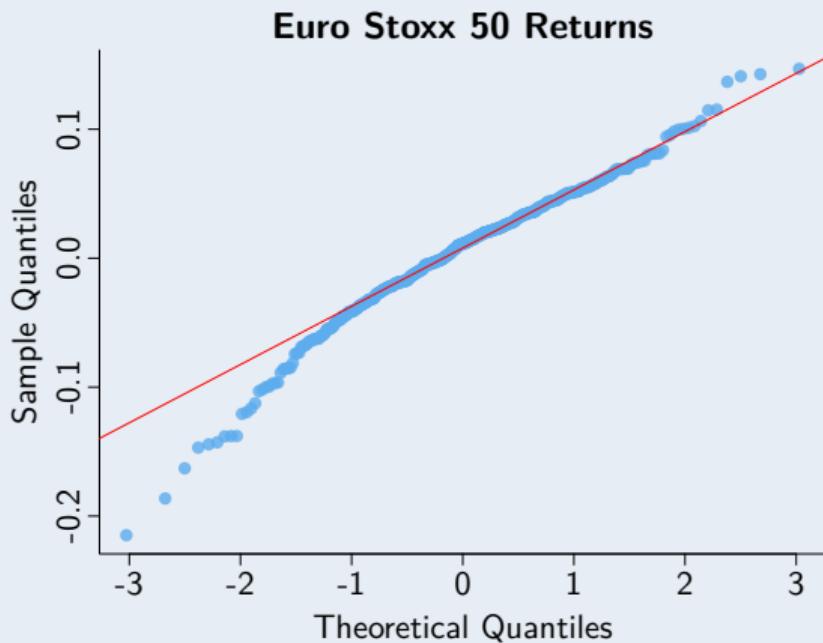
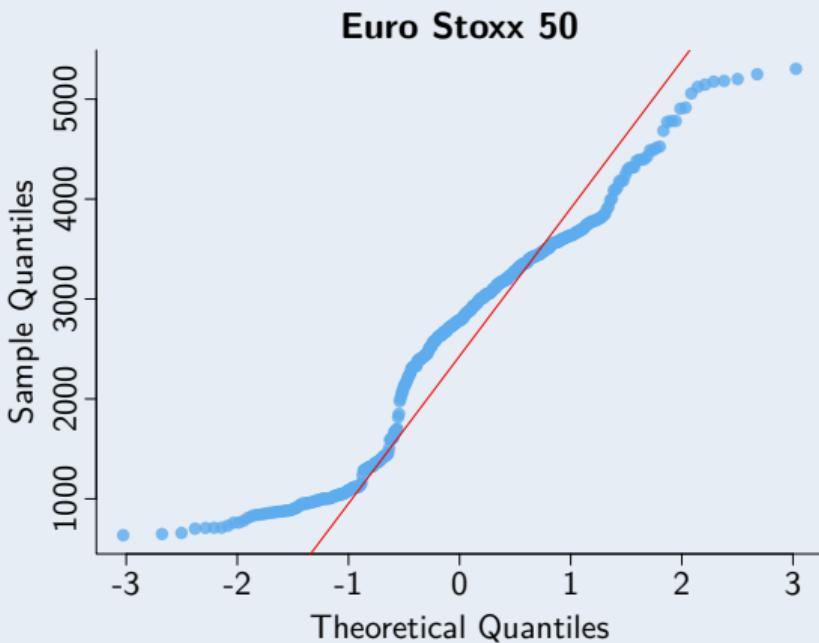


Im Q-Q-Plot ist zu erkennen, dass die unteren Dezile der Verteilungen übereinstimmen. Dies sieht man daran, dass die Daten mit der 45° -Geraden zusammenfallen. Für die Dezile oberhalb des Medians hingegen unterscheiden sich die Verteilungen. Da $\bar{y} > \bar{x}$, liegen die Punkte oberhalb der 45°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

Schiefe

Quantil-Quantil-Diagramm

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



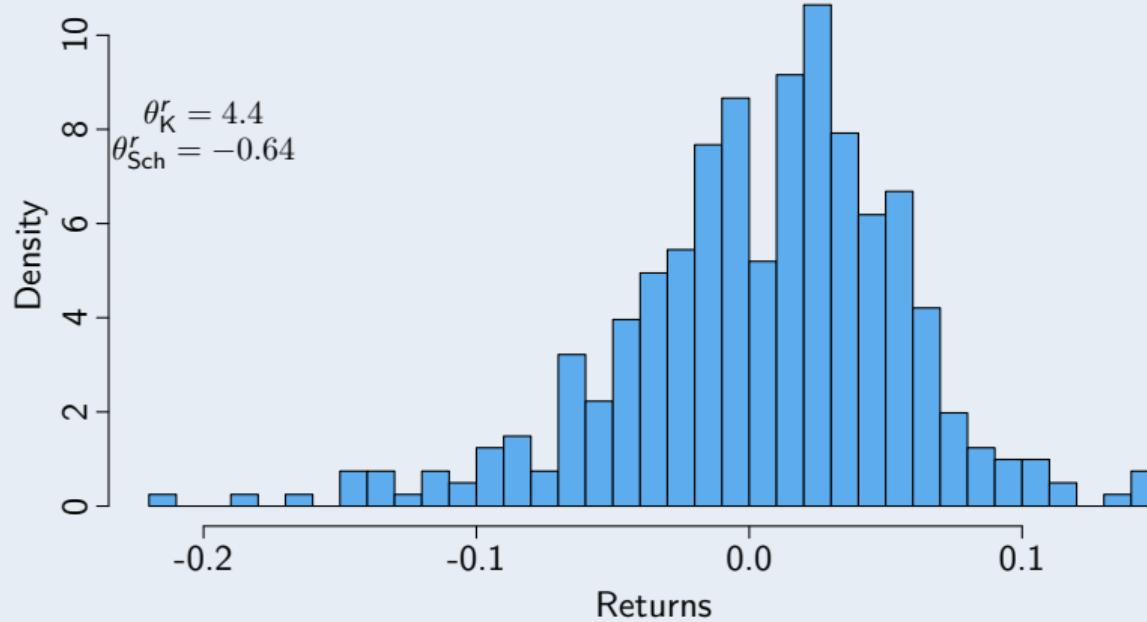
Daten: Yahoo Finance

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.

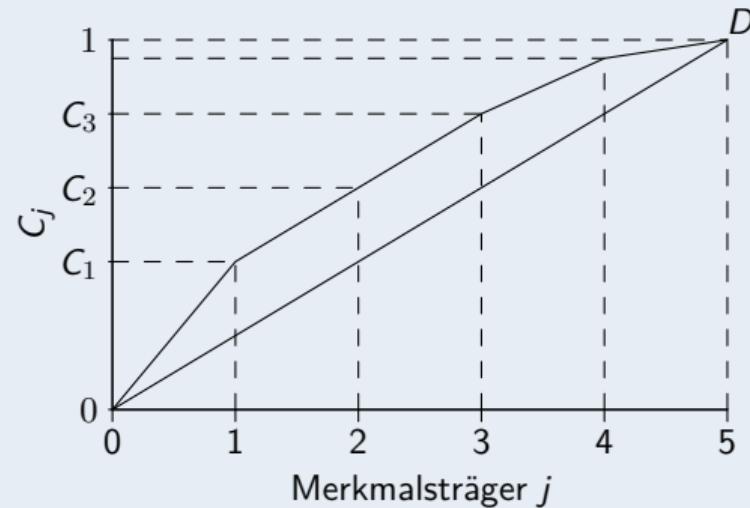
j	x_j	c_j	C_j
1	40	0,40	0,40
2	20	0,20	0,60
3	20	0,20	0,80
4	15	0,15	0,95
5	5	0,05	1,00

$n\bar{x} = 100$

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

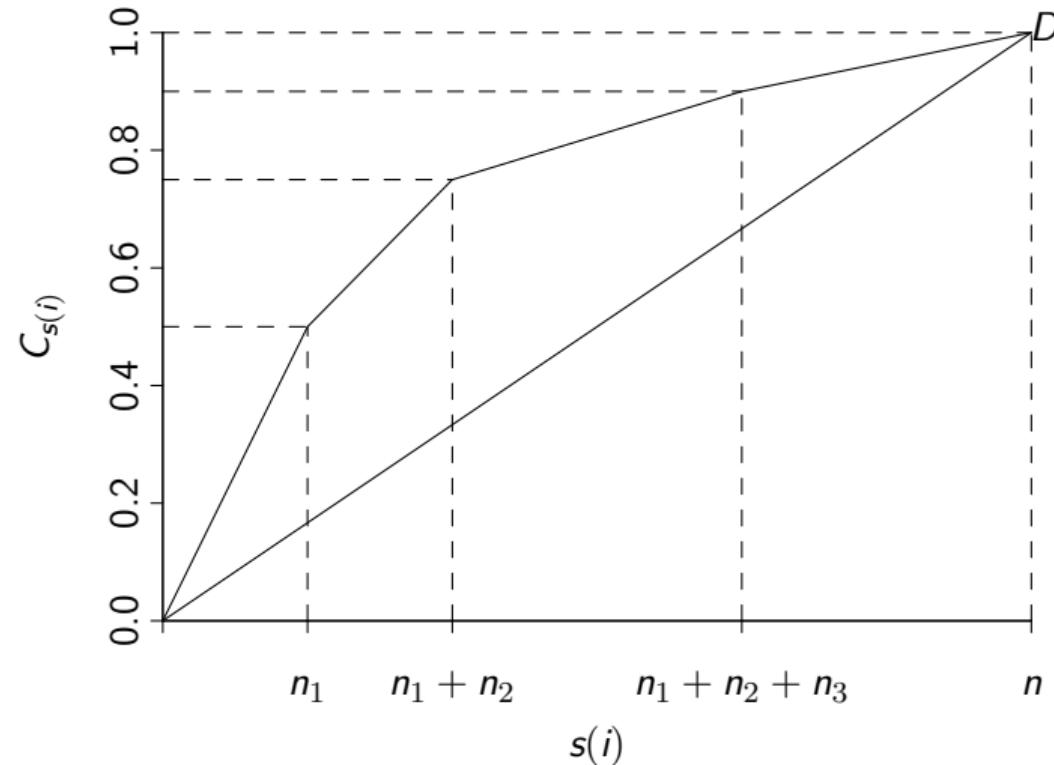
$$C_{s(i)} = \frac{\sum_{r=1}^i n_r x_r}{n \bar{x}}$$

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:

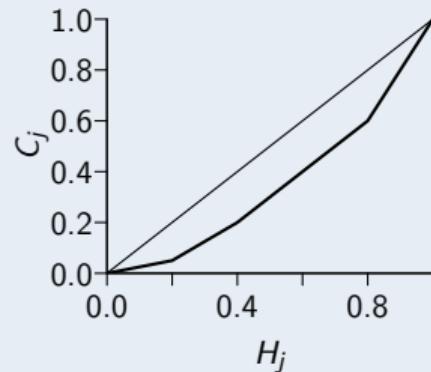
j	x_j	c_j	C_j	$H_j = \frac{j}{n}$
1	5	0,05	0,05	0,2
2	15	0,15	0,20	0,4
3	20	0,20	0,40	0,6
4	20	0,20	0,60	0,8
5	40	0,40	1,00	1,0

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

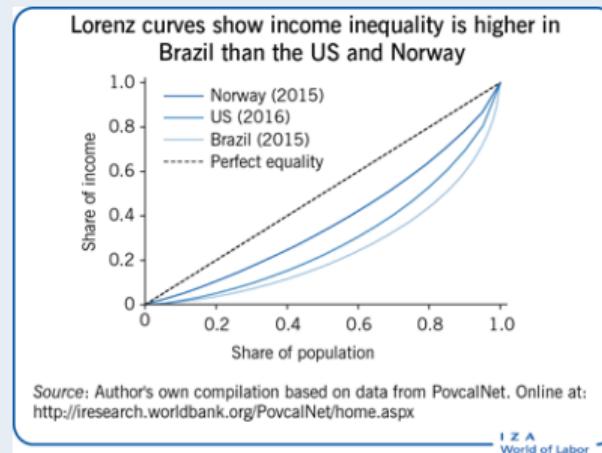
Beispiel 6.7: Fortsetzung.

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.

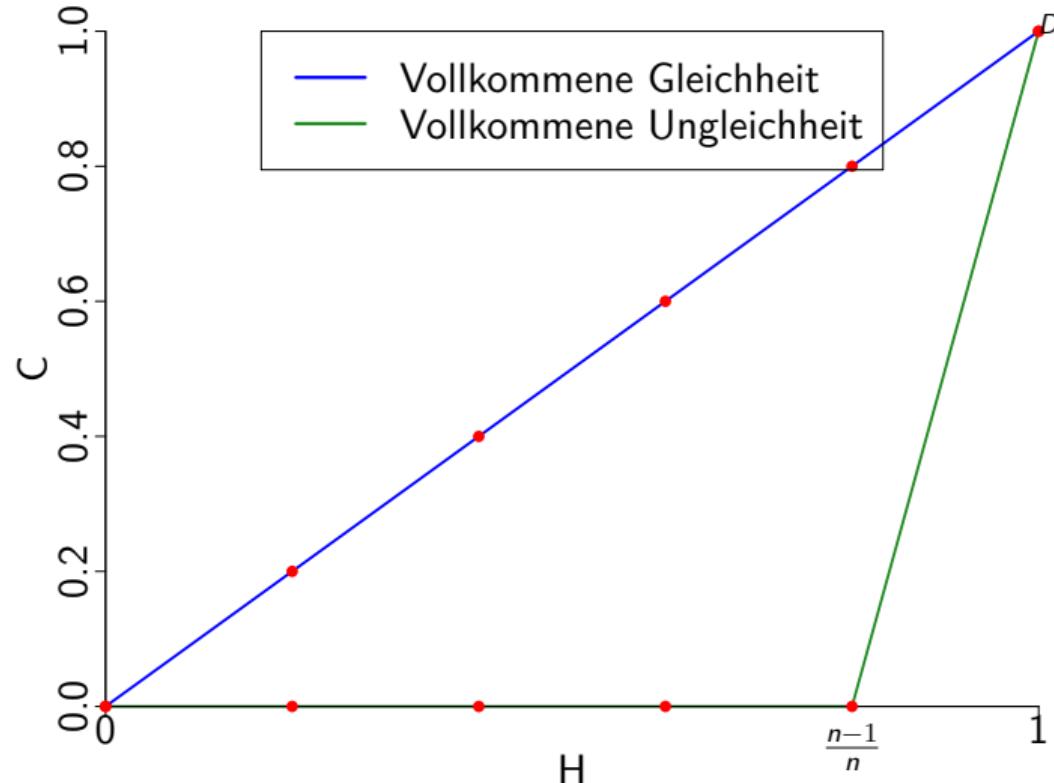
Relative Konzentration

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

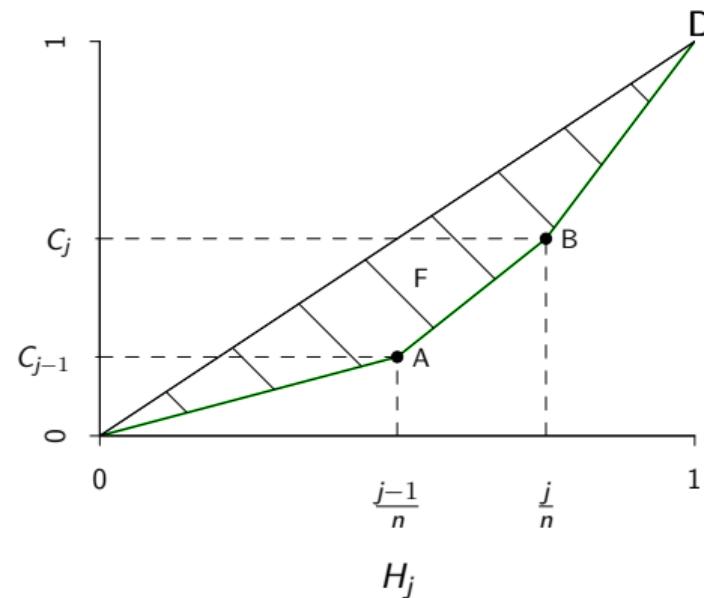
Grenzlagen der Lorenzkurve



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

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

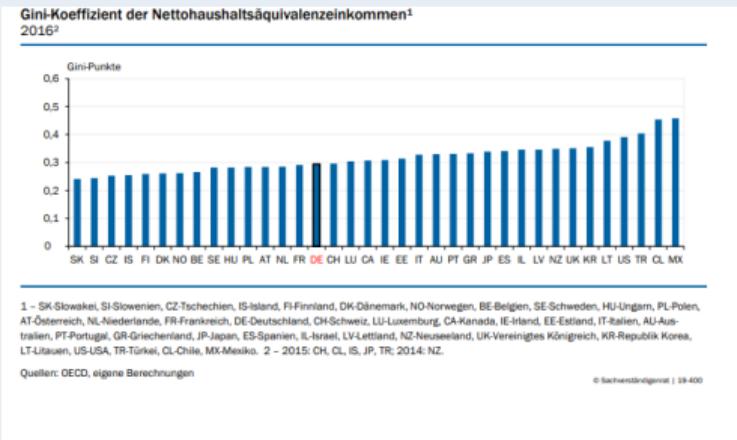
- Für die Konstruktion eines Konzentrationsmaßes mit Werten in $[0, 1]$ gibt es zwei Möglichkeiten:
 - ① Beziehe F auf $\frac{1}{2}$ (Fläche des Dreiecks $(0, 1, D)$) oder
 - ② auf F_{max} .

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

Relative Konzentration

Gini-Koeffizient D_G

Beispiel 6.9: OECD.

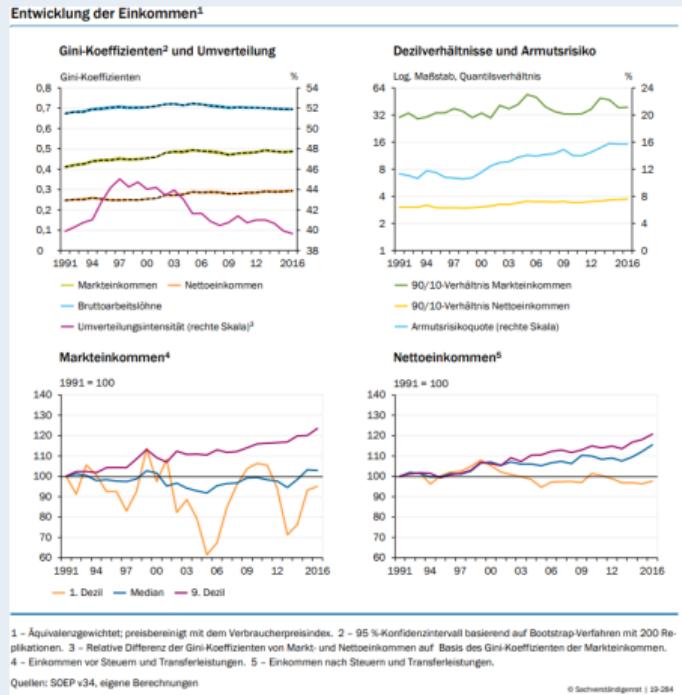


Quelle: OECD (2020), How's Life? 2020:Measuring Well-being, OECD Publishing

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)
```

