DHBW Mannheim

Fakultät Wirtschaft
Studiengang, -richtung: BWL, Industrie

Statistik

2. Semester

Allgemeines

Dozent

- Thilo Klein
- Duale Ausbildung (Bayer AG, Leverkusen)
- Diplom in Wirtschaftspädagogik und Mathematik, FSU Jena
- Master in Operations Research, University of Cambridge
- 2014 Promotion, University of Cambridge
- 2014-2017 Analyst, OECD Statistikdirektorat
- Seit 2017 Ökonom, ZEW Mannheim
- Seit 2018 Lehrbeauftragter für Mathematik, DHBW Mannheim
- Email: thilo@klein.uk

Allgemeines

Modul Wirtschaftsmathematik und Statistik

- Mathematik
  - 30h Präsenzzeit und 45h Selbststudium
  - Lehrbuch: Opitz und Klein (2011). Mathematik – Lehrbuch für Ökonomen
- Statistik
  - 30h Präsenzzeit und 45h Selbststudium
  - Lehrbuch: Quatember (2014). Statistik ohne Angst vor Formeln

Allgemeines

Modul Wirtschaftsmathematik und Statistik

Andreas Quatember

Statistik ohne Angst vor Formeln
Das Studienbuch für Wirtschafts- und Sozialwissenschaftler

4., aktualisierte Auflage

ISBN 978-3-86894-218-7

203 Seiten | 2-farbig
Mai 2014
€ 24,95 [D] | € 25,70 [A] | SFR 33,60

www.pearson-studium.de
www.pearson.ch

Klausur Statistik

Hinweise

- Inhalt
  - Vorlesungsstoff (ca. 33%) und Übungsaufgaben (ca. 67%)
    - Vorlesungsstoff: Theorie verstehen, erklären
    - Übungsaufgaben: Theorie anwenden
  - Transferelemente: Anwendung auf verwandte Fragestellungen

Allgemeines

Was ist Statistik?

- Alle Methoden der Analyse von Daten mit dem Ziel der Informationsbündelung
- Statistik ist Alltag!

- Analysen des Finanzmarktes, z.B. Kursschwankungen
- Big Data, Kundendatenanalysen im Web (Amazon, iTunes, Google)
- Analysen im Sport, z.B. Matchstatistiken im Fußball

Allgemeines

Was ist Statistik?

Allgemeines

Was ist Statistik?

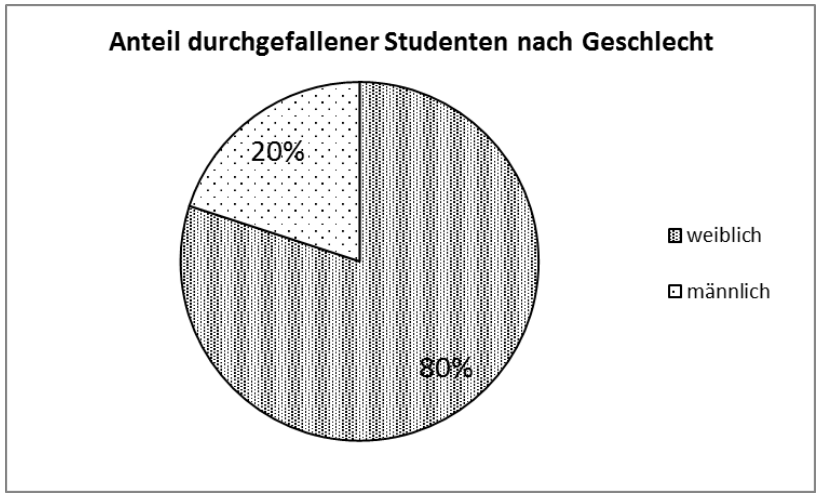
- allerdings schwieriges Image des Faches
  - „und jetzt noch etwas für die Statistiker unter unseren Zusehern“
  - „Mit Statistik lässt sich alles beweisen“
  - „Ich glaube nur den Statistiken die ich selbst gefälscht habe“ (Winston Churchill)
  - „There are three kinds of lies: lies, damned lies and statistics“ (Benjamin Disraeli)

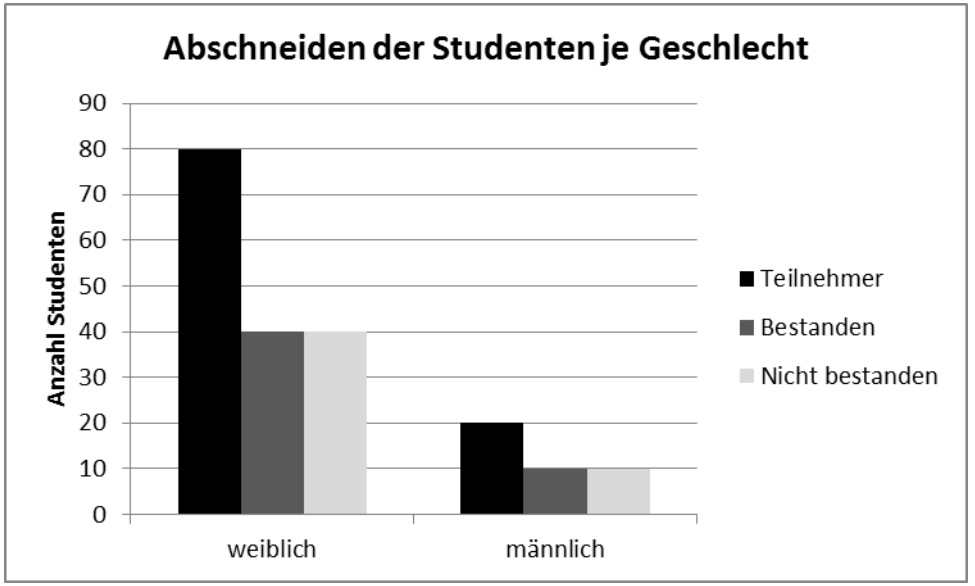
- Verwechselung der Qualität der statistischen Methoden mit der Qualität ihrer Anwendung

Allgemeines

Was ist Statistik?

- Lügen mit Statistik?





Die meisten „Durchfaller“ sind Frauen → Frauen schneiden bei der Klausur schlechter ab

Es gibt mehr weibliche als männliche Studierende und die Durchfallquote ist jeweils gleich → Frauen sind genauso gut wie Männer

Gliederung

1. Beschreibende Statistik

   Es liegen vollständige Daten über eine Grundgesamtheit vor

2. Wahrscheinlichkeitstheorie

   Kombiniert 1 mit 3

3. Schließende Statistik

   Es liegen nur Daten aus einem ausgewählten Teil der Grundgesamtheit vor

Gliederung

- 1. Beschreibende Statistik
- 1.1 Grundbegriffe
- 1.2 Tabellarische und graphische Darstellung von Häufigkeitsverteilungen
- 1.3. Kennzahlen statistischer Verteilungen

Grundbegriffe

Was ist was?

- Erhebungseinheiten: Objekte, über die Daten erhoben werden
- Grundgesamtheit: Gesamtheit aller Erhebungseinheiten
- Merkmal: Eine interessierende Eigenschaft (die analysiert werden soll)
- Merkmalsausprägungen: Die einzelnen möglichen Werte eines Merkmals
- Wertebereich: Alle möglichen Merkmalsausprägungen

Grundbegriffe

Was ist was? – Beispiel 1

Erhebung der Punkteverteilung bei der Statistikklausur

Grundgesamtheit:	alle Prüflinge
Merkmal:	Punkte
Merkmalsausprägungen:	0, 1, 2, ...

Erhebung der Zufriedenheit von Kunden

Grundgesamtheit:	alle Kunden
Merkmal:	Zufriedenheit mit der Beratung
Merkmalsausprägungen:	sehr zufrieden, eher zufrieden, teils-teils, eher unzufrieden, sehr unzufrieden

Grundbegriffe

Was ist was? – Beispiel 1

Erhebung des besten Kinofilms

Grundgesamtheit:	alle teilnahmewilligen Leser und -innen
Merkmal:	bester Film
Merkmalsausprägungen:	Film 1, Film 2, …

Grundbegriffe

Unterscheidung von Merkmalstypen

- Nominal – ordinal – metrisch
  - Nominal: Unterscheidung der Merkmalsausprägungen dem Namen (Bsp: Geschlecht)
  - Ordinal: Merkmalsausprägungen besitzen eine natürliche Reihenfolge (Bsp: Schulnoten)
  - Metrisch: Merkmalsausprägungen lassen sich reihen und haben die gleiche Einheit (Bsp: Körpergröße)
- Diskret – stetig
  - Diskret: Wertebereich umfasst nur bestimmte Merkmalsausprägungen (Bsp: Schulnoten)
  - Stetig: Wertebereich umfasst alle reellen Werte eines Intervalls (Bsp: Körpergröße)

Grundbegriffe

Unterscheidung von Merkmalstypen

Beispiel 2: Merkmalstypen

Merkmal	Merkmalsausprägungen	n / o / m	d / s
Familienstand	ledig (=1), verheiratet (=2), geschieden (=3), verwitwet (=4)	nominal	diskret
100-m-Zeiten	11,21 sec., 11,2435 sec., ...		
Preis eines Sportartikels	29,90 €, 34,90 €, ...		
Platzierungen in einem 100m-Lauf	1., 2., 3., ...		
Weitsprungleistung (in ganzen cm)	516 cm, 492 cm, ...		

Grundbegriffe

Kodierung von Merkmalsausprägungen

Geschlecht: **O** weiblich (=1) **O** männlich (=2)

Alter (in vollendeten Lebensjahren): ..... Jahre

Wie schätzen Sie die didaktisch-methodische Qualität der LVA ein?
O 1 (=sehr gut) **O** 2 **O** 3 **O** 4 **O** 5 (=sehr schlecht)

Waren die angegebenen Lernunterlagen hilfreich?
O 1 **O** 2 **O** 3 **O** 4 **O** 5 (1=sehr hilfreich, … , 5=überhaupt nicht hilfreich)

Dateneingabe für die elektronische Verarbeitung (z.B. in Excel):

Antwort auf 1. Frage

Antwort auf 2. Frage

	A	B	C	D	E
1	ID	Frage 1	Frage 2	Frage 3	Frage 4
2	1	2	21	1	3
3	2	1	38	2	2
4	3	1	42	3	2

1. Erhebungseinheit:
2. Erhebungseinheit:
3. Erhebungseinheit:

Tabellarische und graph. Darstellung von Häufigkeitsverteilungen

Gliederung

- 1.2. Tabellarische und graphische Darstellung von Häufigkeitsverteilungen
- 1.2.1 Häufigkeitsverteilung einzelner Merkmale
- 1.2.2 Häufigkeitsverteilung zweier Merkmale

Häufigkeitsverteilung einzelner Merkmale

Beispiel 3: Tabellarische Darstellung einer Häufigkeitsverteilung

Häufigkeiten (h): Erster Überblick

Punktezahlen (i)	Häufigkeit h
0	1
1	3
2	10
3	16
4	32
5	44
6	20
7	16

N=142

Häufigkeitsverteilung einzelner Merkmale

Tabellarische Darstellung von Häufigkeiten

Häufigkeiten (h): Erster Überblick
Relative Häufigkeiten oder Anteile (p) einer Merkmalsausprägung: $p_i = h_i/N$
Prozentzahlen: $p_i \cdot 100$

Punktezahlen (i)	Häufigkeit h	Relative Häufigkeit p	Prozent
0	1	0,007	0,7
1	3	0,021	2,1
2	10	0,070	7,0
3	16	0,113	11,3
4	32	0,225	22,5
5	44	0,310	31,0
6	20	0,141	14,1
7	16	0,113	11,3
	N=142		

Häufigkeitsverteilung einzelner Merkmale

Tabellarische Darstellung von Häufigkeiten

Relative Summenhäufigkeit (oder empirische Verteilungsfunktion) = Summe der relativen Häufigkeiten einer Merkmalsausprägung und aller kleineren Merkmalsausprägungen

Punktezahlen (i)	Häufigkeit h	Relative Häufigkeit p	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1,000

Nur sinnvoll bei metrischen oder ordinalen Merkmalen!

Häufigkeitsverteilung einzelner Merkmale

Tabellarische Darstellung von Häufigkeiten

Beispiel 4: Zusammenfassung von Merkmalsausprägungen zu Intervallen: Besonders bei stetigen Merkmalen oder Merkmalen mit vielen Ausprägungen

Altersklassen (i)	Häufigkeit h	Relative Häufigkeit p	Prozent	Relative Summenhäufigkeit
0 - 14	10.805.291	0,135	13,5	0,135
15 - 29	13.722.052	0,171	17,1	0,306
30 - 44	15.845.993	0,198	19,8	0,503
45 - 59	18.625.423	0,232	23,2	0,735
60 - 74	13.737.405	0,171	17,1	0,907

N=80.219.659

Anteil der Bevölkerung unter 30 Jahre = 30,6%

Anteil der Bevölkerung zwischen 15 und 59 Jahren als Differenz der relativen Summenhäufigkeiten = 0,735 - 0,135 = 0,600

Häufigkeitsverteilung einzelner Merkmale

Tabellarische Darstellung von Häufigkeiten

Pro und Contra der Zusammenfassung in Intervallen

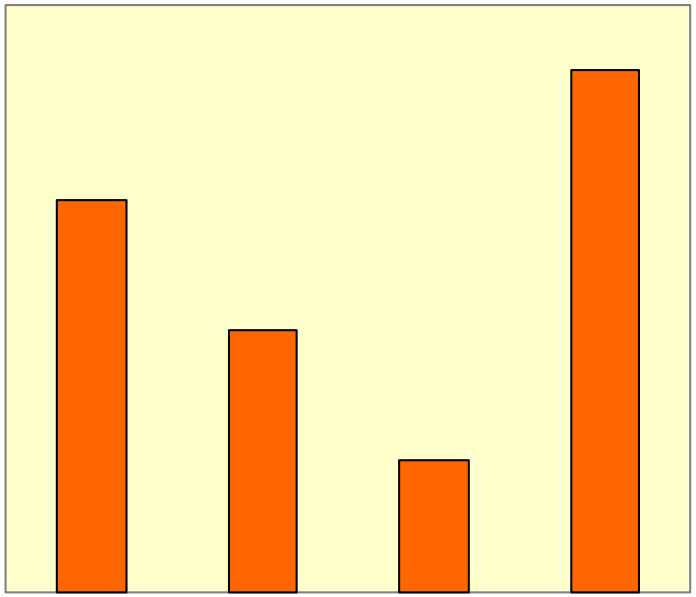
- Vorteil: Bessere Übersicht
- Nachteil: Verlust an Informationen

Häufigkeitsverteilung einzelner Merkmale

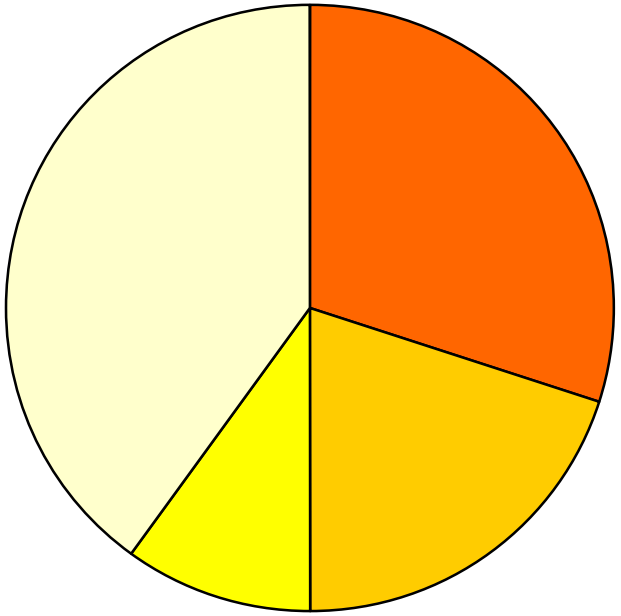
Graphische Darstellung von Häufigkeiten

- Aufgabe: Die wesentlichsten Informationen „auf einen Blick“

Säulendiagramm:
Balken-, Stabdiagramm



Kreisdiagramm:
Kuchen-, Tortendiagramm

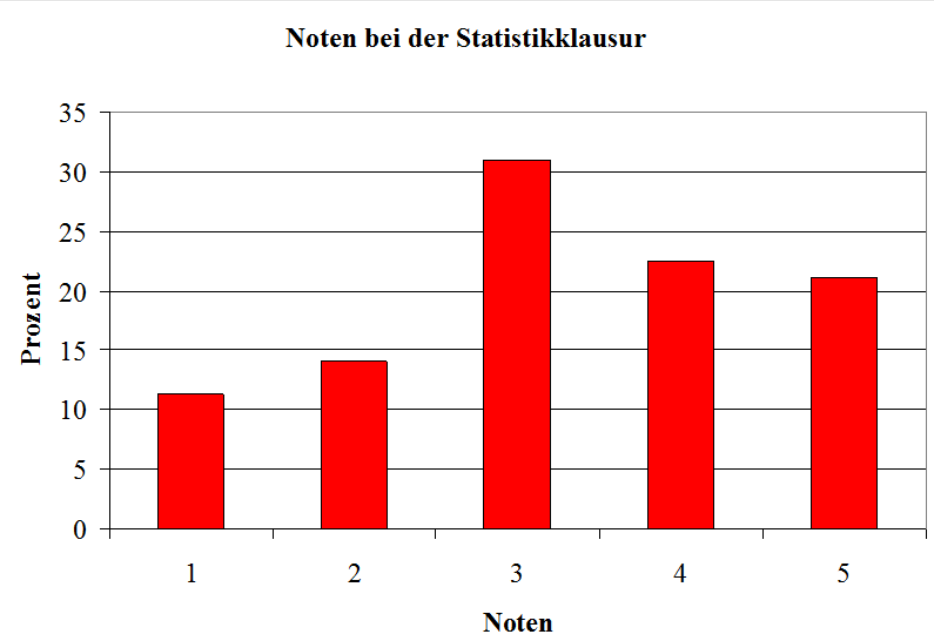


Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung

Note (i)	Häufigkeit h	Relative Häufigkeit p	Prozent
1	16	0,113	11,3
2	20	0,141	14,1
3	44	0,310	31,0
4	32	0,225	22,5
5	30	0,211	21,1

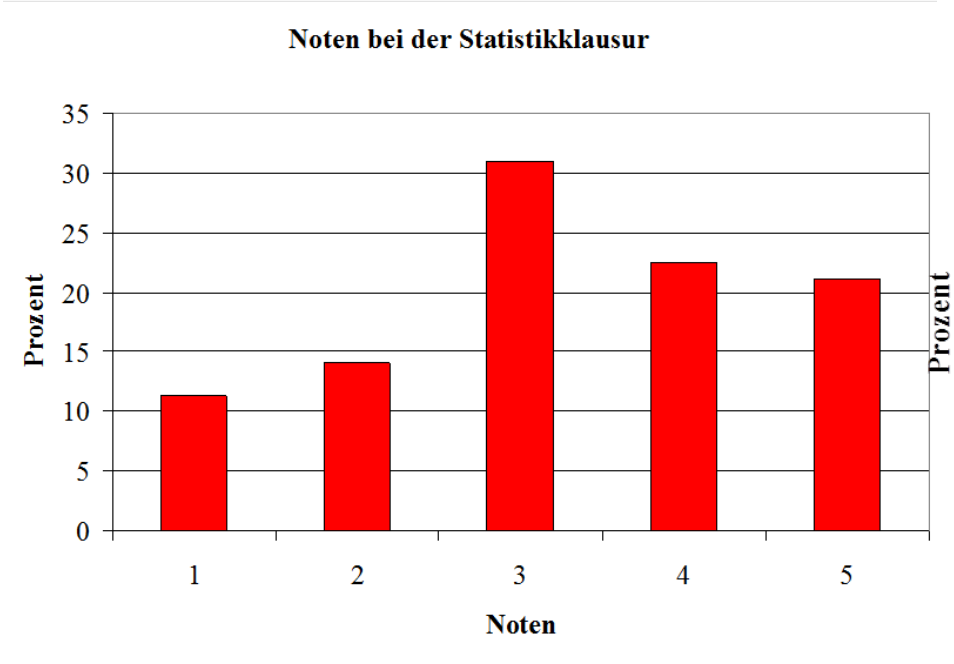


Säulendiagramm

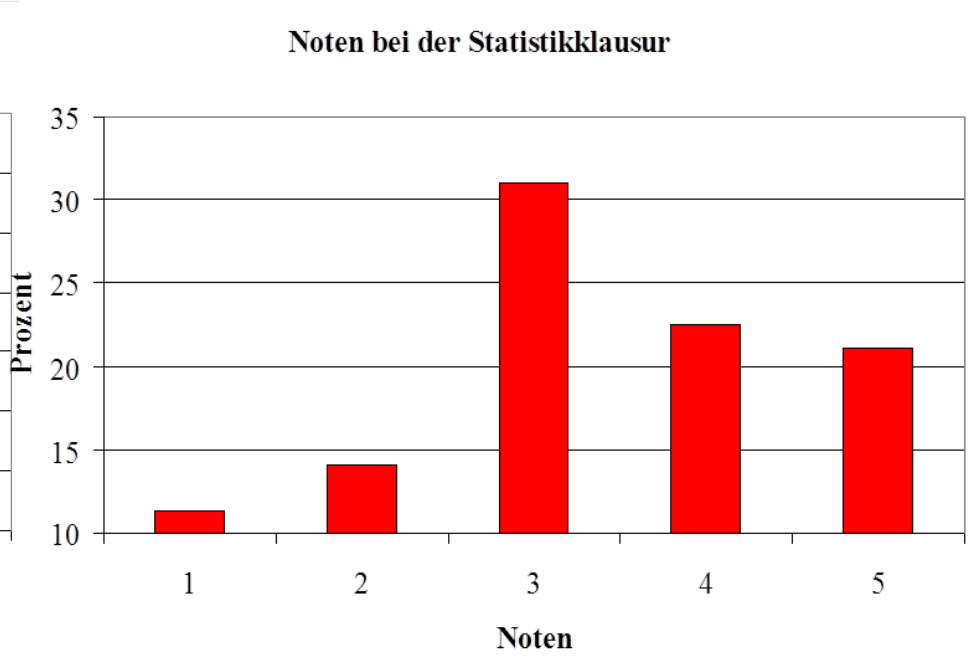
Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



Säulendiagramm



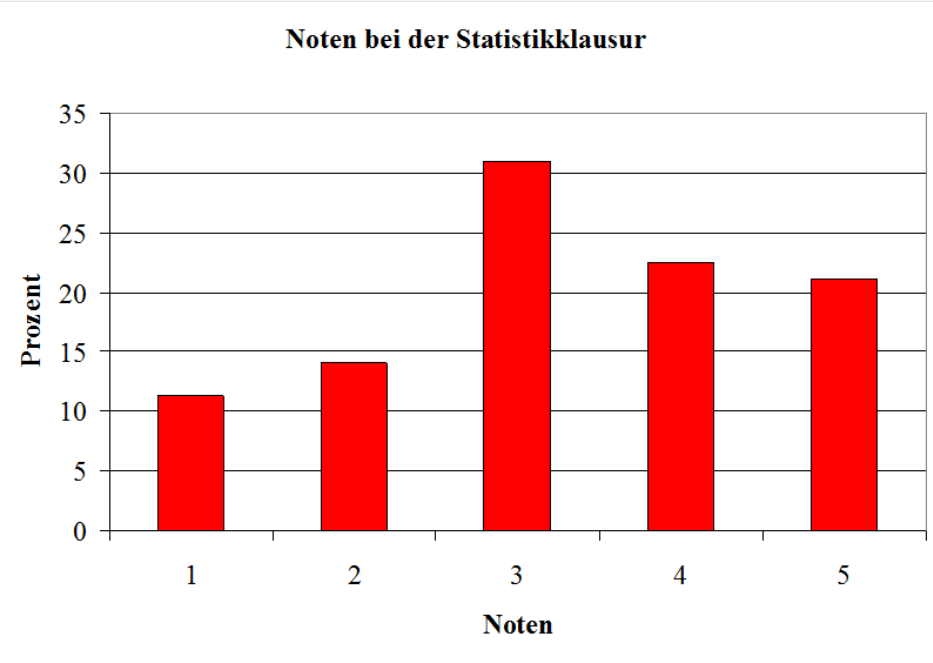
Säulendiagramm mit verschobenem Nullpunkt

→ Falsche Wahrnehmung der Proportionen des Säulendiagramms

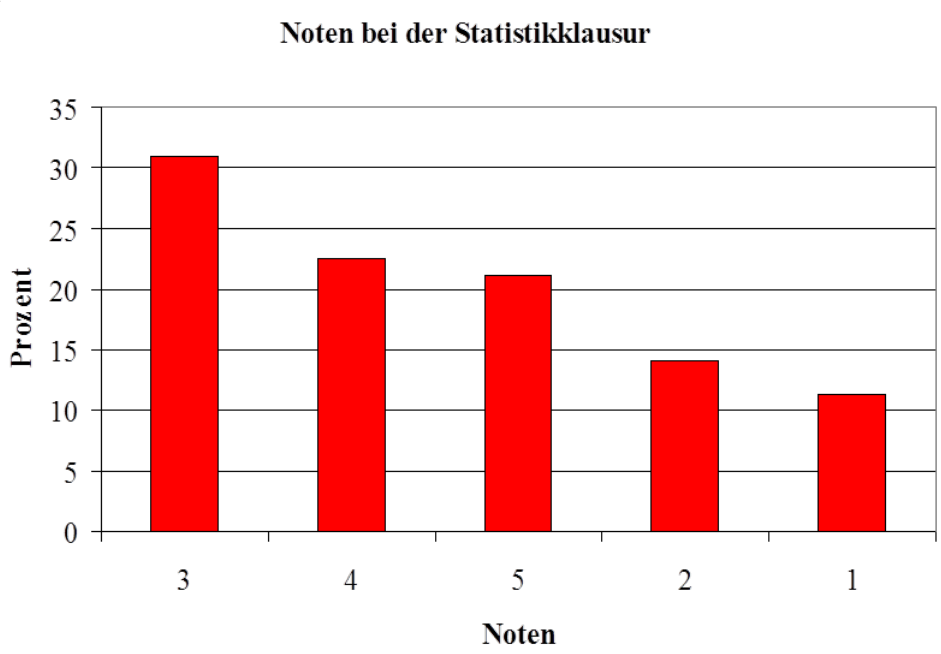
Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



Säulendiagramm



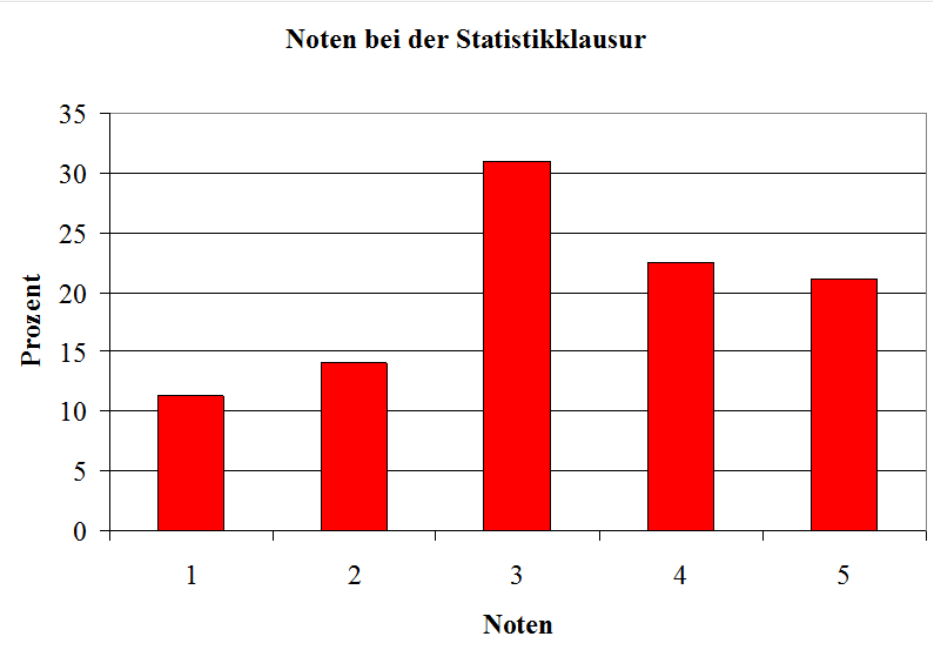
Säulendiagramm mit umgeordneten Merkmalsausprägungen

→ Falsche Wahrnehmung der Verteilung des Säulendiagramms

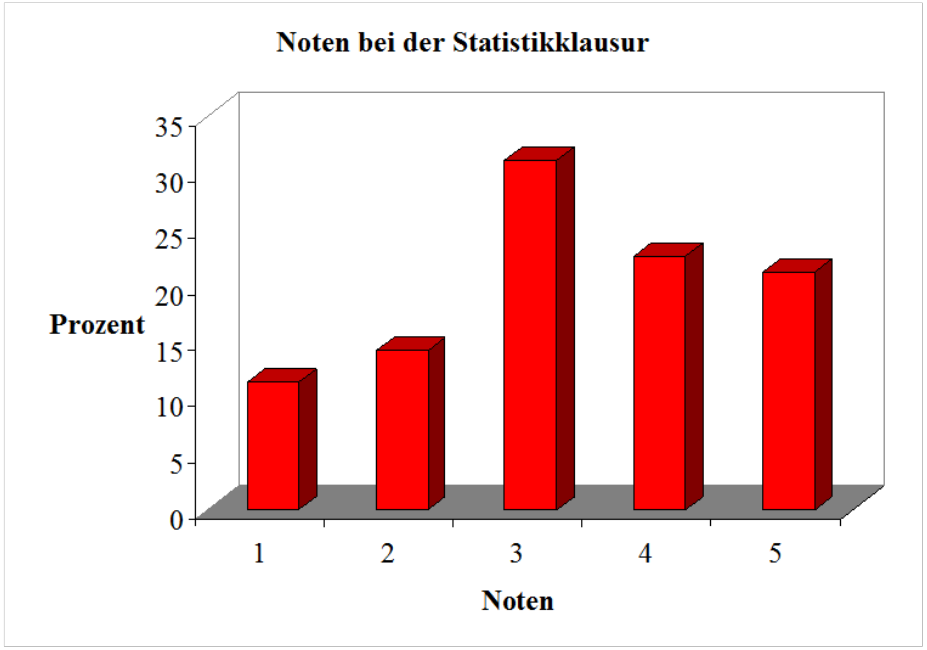
Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



Säulendiagramm



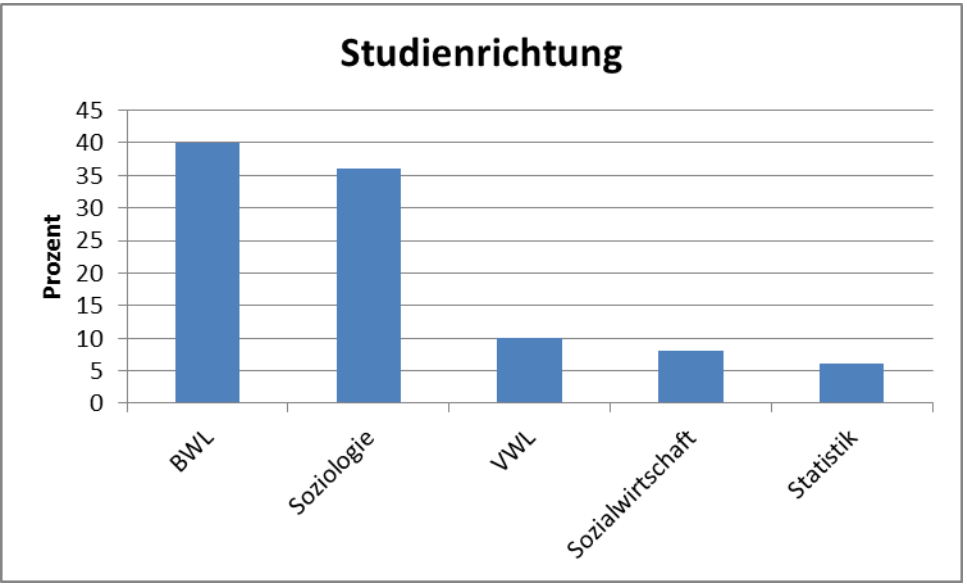
3D Säulendiagramm

→ Verminderte Ablesbarkeit der Säulenhöhen

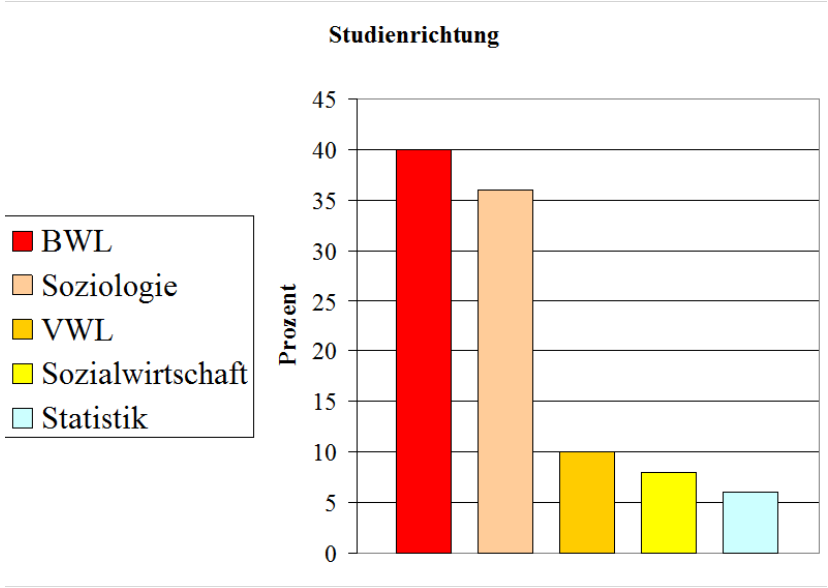
Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



Säulendiagramm mit Bezeichnung der Merkmalsausprägungen auf der x-Achse



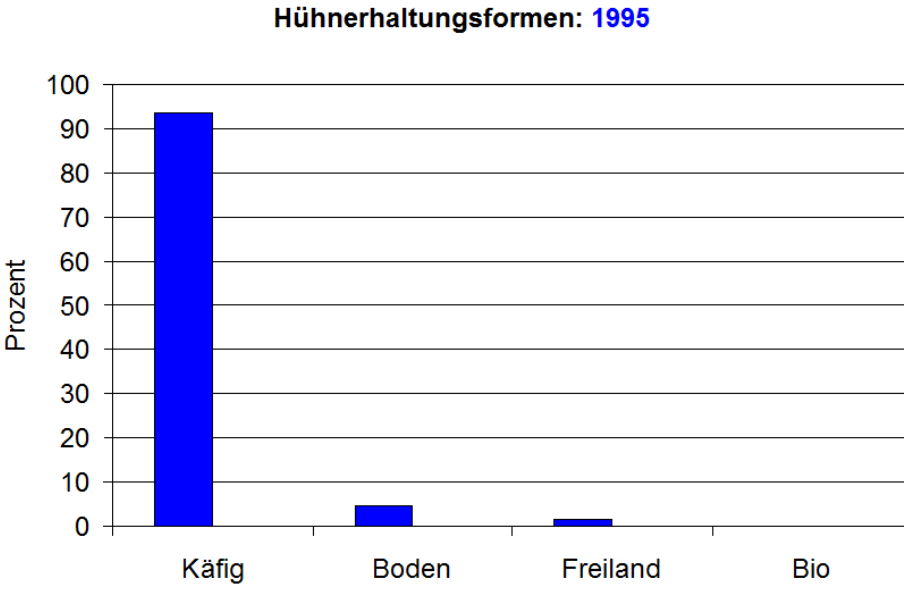
Säulendiagramm mit Legende

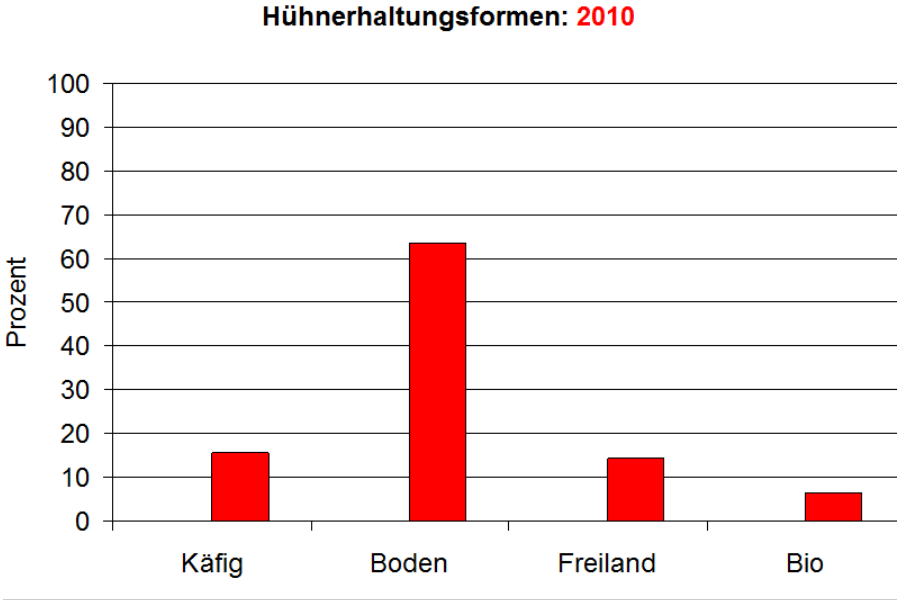
→ Erhöhte Komplexität durch Wechsel zwischen Legende und Diagramm

Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung





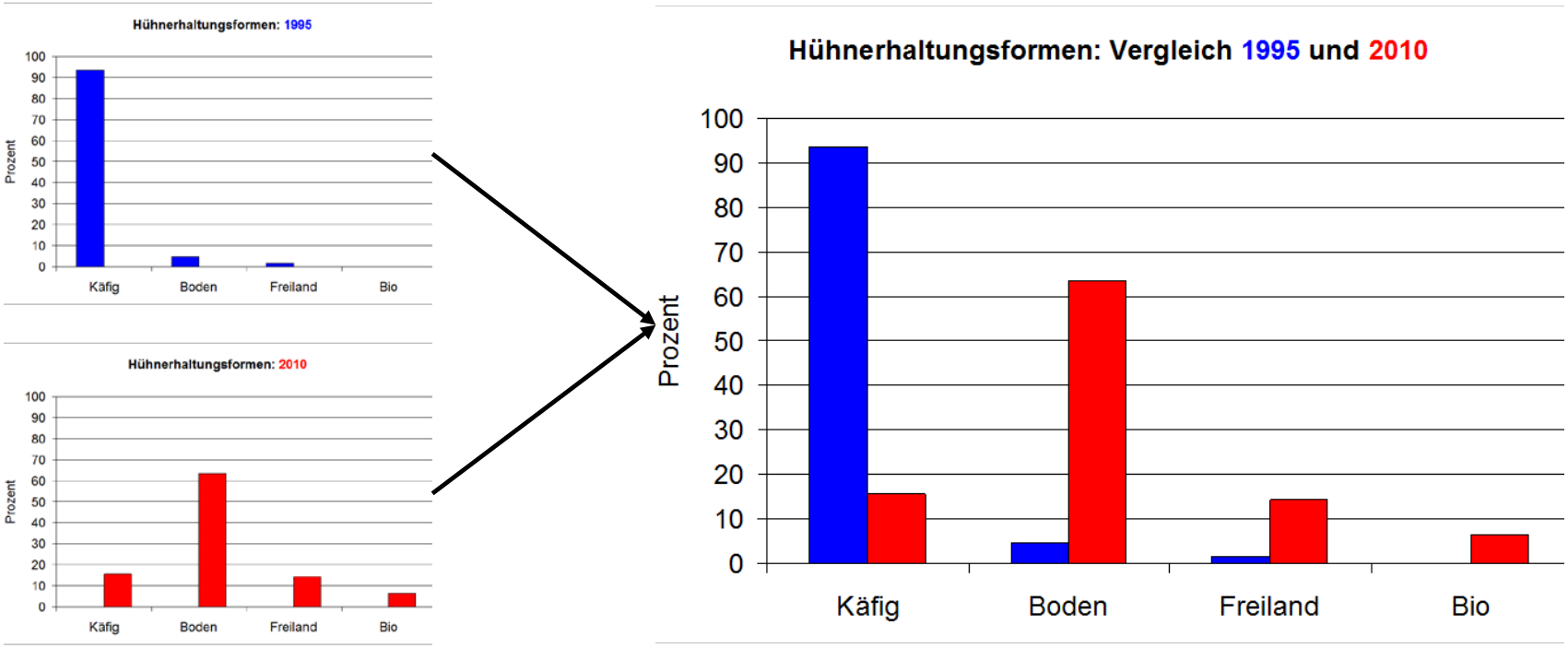
Säulendiagramme über zwei Zeitperioden

→ Erschwert die Vergleichbarkeit

Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



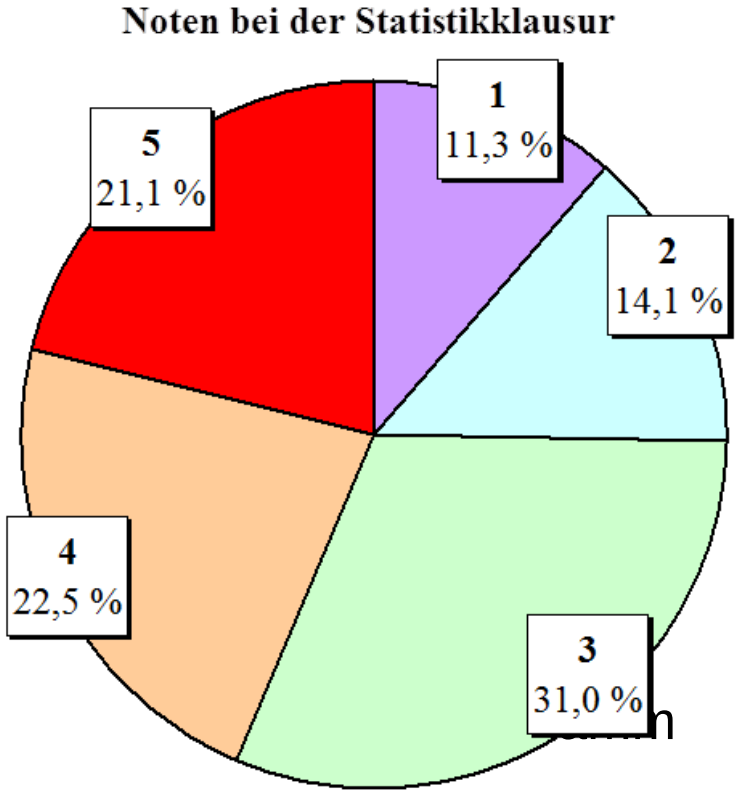
Säulendiagramme über zwei Zeitperioden

→ Kombination in einem Diagramm erleichtert die Vergleichbarkeit

Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Beispiel 5: Graphische Darstellung einer Häufigkeitsverteilung



→ Auch relative Summenhäufigkeiten sind im Kreisdiagramm ablesbar

Häufigkeitsverteilung einzelner Merkmale

Graphische Darstellung von Häufigkeiten

Regeln für die graphische Darstellung

- Säulendiagramme
  - Beschriftungen der x- und y-Achse sind unbedingt anzuführen
  - Nullpunkt der Prozentzahlen auf der y-Achse sollte zum Schnittpunkt der x-Achse liegen
- Säulen- und Kreisdiagramme
  - Titel sind unbedingt anzuführen
  - Ordnung innerhalb der Merkmalsausprägungen beibehalten
  - 3D-Darstellungen vermeiden
  - Direkte Beschriftungen sind Legenden vorzuziehen

→ Die einfachste Grafik ist oft die Beste und spart Zeit!

Gemeinsame Häufigkeitsverteilung 2er Merkmale

Tabellarische Darstellung

- Häufig werden mehrere Merkmale auf einmal erhoben (z.B. Noten der Statistikklausur und Geschlecht der Studierenden)

→ Ermöglicht Vergleiche über Gruppen

Beispiel 6: Tabellarische Darstellung der gemeinsamen Häufigkeitsverteilung zweier Merkmale

		Studienrichtung					
Geschlecht		BWL	Soz	VWL	Sowi	Stat	Summe
	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

Gemeinsame Häufigkeitsverteilung 2er Merkmale

Tabellarische Darstellung

Relative Häufigkeiten (p) = Absolute Häufigkeit / Anzahl Erhebungseinheiten

$p_{ij} = h_{ij}/N$, z.B.: $p_{11} = h_{11}/N$ → 0,22=110/500

		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $N_{i.}$
Geschlecht (i)	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Randv. $N_{.j}$	200	180	50	40	30	500

		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $p_{i.}$
Geschlecht (i)	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Randv. $p_{.j}$	0,40	0,36	0,10	0,08	0,06	1

Randverteilung: Verteilung der einzelnen Merkmale (Geschlecht, Studien-richtung) am Rand der Tabellen

Gemeinsame Häufigkeitsverteilung 2er Merkmale

Tabellarische Darstellung

Vergleich: Häufigkeitsverteilung der Studienrichtung unter Frauen und unter Männern

Bedingte Häufigkeiten = Absolute Häufigkeit / Anzahl Erhebungseinheiten in einer Gruppe: $p_{j|i=k} = h_{kj}/N_{k.}$ z.B. $p_{1|i=1} = h_{11}/N_{1.} = 110/300 = 0{,}37$

		Studienrichtung (j)					
Geschlecht (i)		BWL	Soz	VWL	Sowi	Stat	Summe
	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500

		Studienrichtung (j)					
Geschlecht (i)		BWL	Soz	VWL	Sowi	Stat	Summe
	weiblich	0,37	0,40	0,07	0,10	0,07	1
	männlich	0,45	0,30	0,15	0,05	0,05	1

Gemeinsame Häufigkeitsverteilung 2er Merkmale

Tabellarische Darstellung

Vergleich: Häufigkeitsverteilung der Studienrichtung unter Frauen und unter Männern

Beispiel 7: Tabellarische Darstellung einer bedingten Häufigkeitsverteilung

		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Summe
Geschlecht (i)	weiblich	0,37	0,40	0,07	0,10	0,07	1
	männlich	0,45	0,30	0,15	0,05	0,05	1

Korrekte Aussage durch Berücksichtigung der Grundgesamtheit auf die sich Prozentzahlen beziehen:
Unter den Frauen studieren 37% BWL, 40% Soziologie, …
Unter den Männern studieren 45% BWL, 30% Soziologie, …

Tipp: Verwandeln Sie Tabellen nicht in Zahlengräber.
z.B. Verzicht auf die dritte Nachkommastelle

Kennzahlen statistischer Verteilungen

Allgemeines

- Tabellarische und graphische Darstellung geben einen guten Überblick über die Daten → allerdings ist das nur der Anfang aller Statistik
- Weitere Beschreibung anhand von einzelnen Kennzahlen
- Dabei Bündelung der Informationen auf einen einzigen Repräsentanten der Verteilung

Kennzahlen statistischer Verteilungen

Gliederung

- 1.3 Kennzahlen statistischer Verteilungen
- 1.3.1 Kennzahlen der Lage (Mittelwert, Median, Quartile, Modus)
- 1.3.2 Kennzahlen der Streuung (Varianz, Standardabweichung, Variationskoeffizient)
- 1.3.3 Kennzahlen der Konzentration (Lorenzkurve, Ginikoeffizient)
- 1.3.4 Kennzahlen des statistischen Zusammenhangs (Chi Quadrat χ^2, Cramers V, Kovarianz, Korrelationskoeffizient, Spearmannscher Rangkorrelationskoeffizient, Regressionsrechnung)

Kennzahlen der Lage

Arithmetisches Mittel (1. Variante)

- Idee: Stellvertreter für alle Daten ist jener Wert, der sich bei gleichmäßiger Aufteilung der Summe aller auftretenden Daten (=Merkmalssumme) auf die Erhebungseinheiten ergeben würde
- Beispiel Einkommen von fünf Personen in €:
  - Merkmalsausprägungen: 1.000, 3.000, 4.000, 1.000, 1.000
  - Summe der Merkmalsausprägungen: 1.000 + 3.000 + 4.000 + 1.000 + 1.000 = 10.000
  - Gleichmäßige Aufteilung: 10.000 : 5 = 2.000

Kennzahlen der Lage

Arithmetisches Mittel (1. Variante)

- Formale Umsetzung der Idee des Mittelwerts:
  - Zeichen für den Mittelwert: $\bar{x}$ (sprich „x quer“)
  - N = Anzahl der Erhebungseinheiten
  - x_1= Merkmalsausprägung der 1. Erhebungseinheit
  - x_2= Merkmalsausprägung der 2. Erhebungseinheit …
  - x_i= Merkmalsausprägung der i-ten Erhebungseinheit

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Kennzahlen der Lage

Arithmetisches Mittel (2. Variante)

- Idee: Multiplikation der Merkmalsausprägungen und Häufigkeiten
- Beispiel Einkommen von fünf Personen in €:
  - Merkmalsausprägungen: 1.000, 3.000, 4.000, 1.000, 1.000
  - Merkmalssumme: 1.000 · 3 + 3.000 · 1 + 4.000 · 1

    Merkmalsausprägungen · Häufigkeiten
  - Ergebnis: 10.000 : 5 = 2.000

Kennzahlen der Lage

Arithmetisches Mittel (2. Variante)

- Formale Umsetzung der Idee des Mittelwerts 2. Variante:
  - k = Anzahl der verschiedenen Merkmalsausprägungen
  - h_1= Häufigkeit der 1. Merkmalsausprägung
  - h_2= Häufigkeit der 2. Merkmalsausprägung…
  - h_i= Häufigkeit der i-ten Merkmalsausprägung

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot h_i}{N}$$

- Auch mit der relativen Häufigkeit p

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot h_i}{N} = \sum_{i=1}^{k} x_i \cdot \frac{h_i}{N} = \sum_{i=1}^{k} x_i \cdot p_i$$

Kennzahlen der Lage

Arithmetisches Mittel (2. Variante)

- Beispiel 8: Berechnung des Mittelwerts der Statistikklausur

Punktezahlen (i)	Häufigkeit h	Relative Häufigkeit p
0	1	0,007
1	3	0,021
2	10	0,070
3	16	0,113
4	32	0,225
5	44	0,310
6	20	0,141
7	16	0,113

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot h_i}{N} = \frac{0 \cdot 1 + 1 \cdot 3 + \cdots + 7 \cdot 16}{142} = \frac{651}{142} = 4{,}58$$

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot p_i = 0 \cdot 0{,}007 + 1 \cdot 0{,}021 + \cdots + 7 \cdot 0{,}113 = 4{,}58$$

Kennzahlen der Lage

Geometrisches Mittel

- Achtung: Mittelwert eignet sich nur für metrische Merkmale (und auch da nicht immer)
- Beispiel 9: Der Mittelwert von Wachstumsraten
  - Vor drei Jahren: Umsatz von 20 Mio. €. In den drei Jahren seither jährliche Umsatzzuwächse von 10, 90, 50%. Um wie viel Prozent ist der Umsatz pro Jahr durchschnittlich gestiegen?
  - Mittelwert: (10+90+50) : 3 = 50% ?

Jahr	Umsatzverlauf mit proz. Anstieg	Umsatzverlauf mit Mittelwert
1	$20 \cdot (1+0{,}10) = 20 \cdot 1{,}10 = 22$	$20 \cdot 1{,}5 = 30$
2	$22 \cdot 1{,}90 = 41{,}8$	$30 \cdot 1{,}5 = 45$
3	$41{,}8 \cdot 1{,}50 = 62{,}7$	$45 \cdot 1{,}5 = 67{,}5$

Wachstumsfaktor

Kennzahlen der Lage

Geometrisches Mittel

- Wdhl: $20 \cdot 1{,}5 \cdot 1{,}5 \cdot 1{,}5 = 20 \cdot 1{,}5^3 \neq 62{,}7$
- Welcher konstante Wachstumsfaktor würde also 62,7 ergeben?

$$20 \cdot g^3 = 62{,}7 \rightarrow g = \sqrt[3]{\frac{62{,}7}{20}} = \sqrt[3]{3{,}135} = 1{,}464$$

- Auch aus Wachstumsfaktoren:

$$g = \sqrt[3]{1{,}1 \cdot 1{,}9 \cdot 1{,}5} = \sqrt[3]{3{,}135} = 1{,}464$$

→ Die durchschnittliche jährliche Wachstumsrate liegt bei 46,4% (1,464-1)

- Allgemeine Formel $\boxed{g = \sqrt[n]{x_1 \cdot x_2 \cdot, \ldots, \cdot x_n}}$

Kennzahlen der Lage

Geometrisches Mittel

- Häufiges Anwendungsgebiet des geometrische Mittelwerts: prozentuelles Wachstum von Indizes (z.B. Preisindex für die Lebenshaltung, Aktienindizes, …)
- Bsp: Preisliche Entwicklung eines Warenkorbs:
  Jahr 0 = 100, Jahr 1 = 105; Jahr 2 = 108,15

→Inflationsrate: Quotient des aktuellen Werts des Preisindexes für die Lebenshaltung und des Werts vor genau einem Jahr

→Jahr 1: 105 / 100 =1,05 → Inflationsrate 5%

→Jahr 2: 108,15 / 105 = 1,03 → Inflationsrate 3%

→Durchschnittliche Inflationsrate = $\sqrt[2]{1{,}05 \cdot 1{,}03} = 1{,}03995$ → Knapp unter 4%

Kennzahlen der Lage

Median (Zentralwert)

- Idee des Median ($\tilde{x}$, sprich „x Welle“): Als Stellvertreter für alle Daten gilt jener Wert, der bei – Sortierung der Daten aller N Erhebungseinheiten nach der Größe – in der Mitte steht.
- Bsp: Körpergröße von 5 Erhebungseinheiten (*ungerade* Anzahl):

  148, 158, 148, 160, 155

→Sortierung: 148, 148, 155, 158, 160 → $\tilde{x} = 155$

- Bsp: Körpergröße von 6 Erhebungseinheiten (*gerade* Anzahl):

  148, 158, 148, 160, 155, 157

→Sortierung: 148, 148, 155, 157, 158, 160 → $\tilde{x} = \frac{155+157}{2} = 156$

→Interpretation: (mindestens) die Hälfte der Erhebungseinheiten hat Werte die kleiner gleich dem Median sind, andere Hälfte größer gleich

Kennzahlen der Lage

Median (Zentralwert)

- Allgemeine Formel

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

Kennzahlen der Lage

Median (Zentralwert)

Beispiel 10: Median eines diskreten Merkmals

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1,000

142 Erhebungseinheiten → 71. und 72. stehen in der Mitte
Wann überschreitet die relative Summenhäufigkeit *erstmals* 0,5?
→ $\tilde{x} = 5$

Voraussetzung für Medianberechnung ist die Sortierbarkeit der Merkmalsausprägungen → nur bei metrischen und ordinalen Merkmalen

Kennzahlen der Lage

Median (Zentralwert)

- Vergleich arithmetisches Mittel und Median
  - Einkommensverteilung: 1.000, 1.000, 1.000, 1.000, 11.000
  - Mittelwert: $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{15.000}{5} = 3.000$
  - Median: $\tilde{x} = x_{\frac{n+1}{2}} = x_{\frac{5+1}{2}} = x_3 = 1.000$

→Mittelwert ist anfällig gegenüber Ausreißern (auch Zahlenfehlern)

→Median ist robust gegenüber Ausreißern

Kennzahlen der Lage

Quartile

- Median ist zwar informativ, aber viele Informationen gehen trotzdem verloren: z.B. welche Punktzahl erreichen mindestens 25% der Studenten
- Idee: Median teilte die Verteilung in zwei Hälften → Unterteilung in Viertel auch möglich (Quartile = Viertelwerte der Verteilung)
  - 1. Quartil (unteres Quartil, $Q_{0,25}$) – 25% der Erhebungseinheiten sind kleiner gleich dem 1. Quartilswert
  - 2. Quartil (mittleres Quartil, $Q_{0,50}$) – 50% der Erhebungseinheiten sind kleiner gleich dem 2. Quartilswert (= Median)
  - 3. Quartil (oberes Quartil, $Q_{0,75}$) – 75% der Erhebungseinheiten sind kleiner gleich dem 3. Quartilswert

Kennzahlen der Lage

Quartile

- Allgemeine Formel

$$Q_p = \begin{cases} x_{\lceil n \cdot p \rceil} & \text{für n·p nicht ganzzahlig} \\ \frac{1}{2} \cdot \left(x_{n \cdot p} + x_{n \cdot p + 1} \right) & \text{für n·p ganzzahlig} \end{cases}$$

Kennzahlen der Lage

Quartile

- Bsp: Körpergröße von 7 Erhebungseinheiten

148, 158, 148, 160, 155, 178, 162

Sortierung: 148, 148, 155, 158, 160, 162, 178

1. Quartil: p=0,25 → n·p=7·0,25=1,75 (*nicht ganzzahlig*)
→ $Q_{0,25} = x_{[n \cdot p]} = x_{[1,75]} = x_2 = 148$

- Interpretation: (mindestens) 25% der Erhebungseinheiten haben Werte die kleiner gleich dem 1. Quartilswert sind (sind höchstens 148 cm groß)

Kennzahlen der Lage

Quartile

- Bsp: Körpergröße von 8 Erhebungseinheiten:

148, 158, 148, 160, 155, 178, 162, 165

Sortierung: 148, 148, 155, 158, 160, 162, 165, 178

3. Quartil: p=0,75 → n·p=8·0,75=6 (*ganzzahlig*) →

$$Q_{0,75} = \frac{1}{2} \cdot \left(x_{8 \cdot 0,75} + x_{8 \cdot 0,75+1}\right) = \frac{1}{2} \cdot (x_6 + x_7) = \frac{162 + 165}{2} = 163,5$$

- Interpretation: (mindestens) 75% der Erhebungseinheiten haben Werte die kleiner gleich dem 3. Quartilswert sind (sind höchstens 163,5 cm groß)

Kennzahlen der Lage

Quartile

Beispiel 10: Quartile eines diskreten Merkmals

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1,000

1. Quartil
2. Quartil
3. Quartil

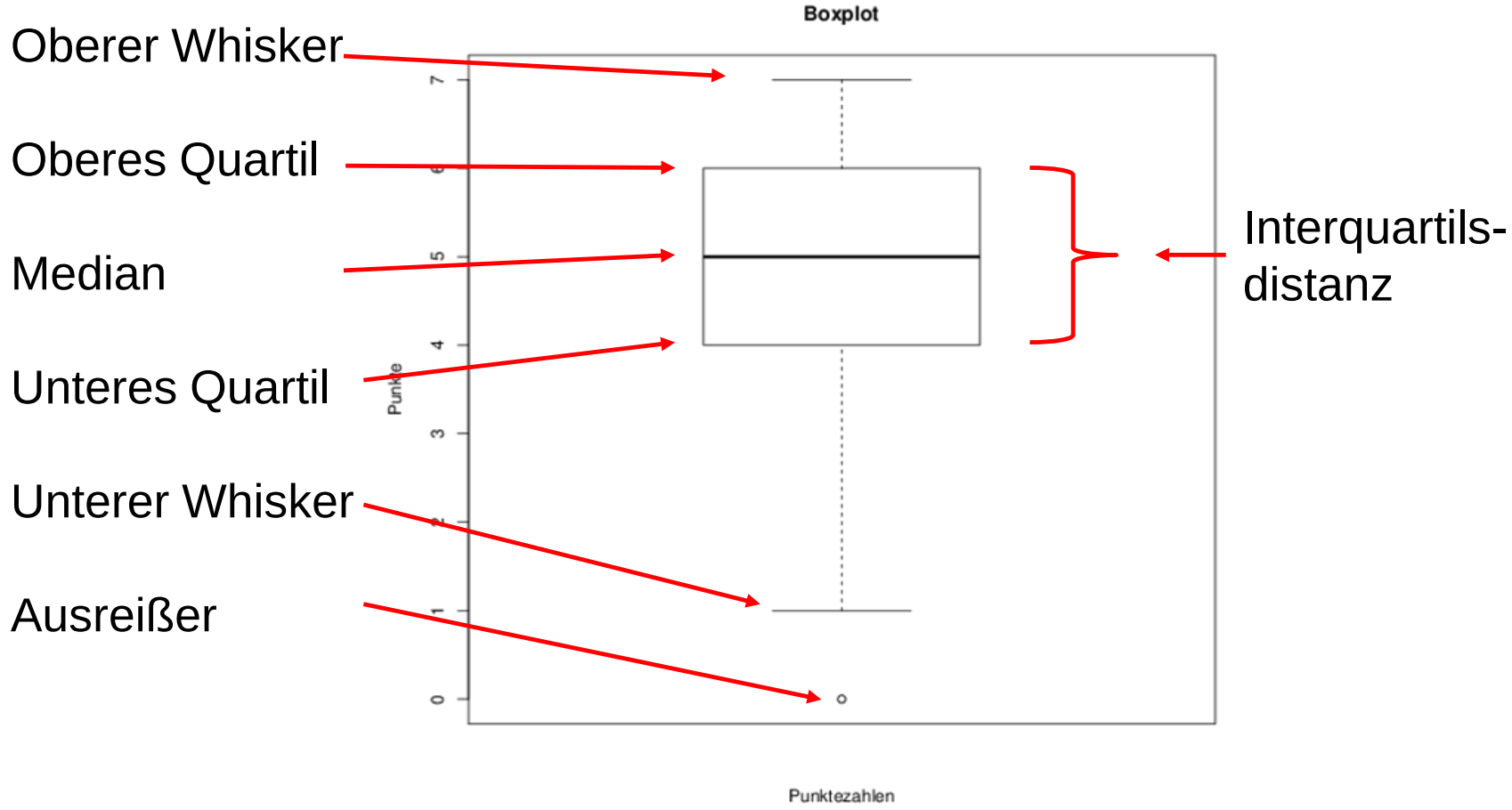
Wann überschreitet die relative Summenhäufigkeit das erste Mal 0,25; 0,5; 0,75

1. Quartil $Q_{0,25}=4$ → (mindestens) 25% der Studenten erreichen höchstens 4 Punkte
2. Quartil $Q_{0,5}=5$ → (mindestens) 50% der Studenten erreichen höchstens 5 Punkte
3. Quartil $Q_{0,75}=6$ → (mindestens) 75% der Studenten erreichen höchstens 6 Punkte

Kennzahlen der Lage

Boxplots

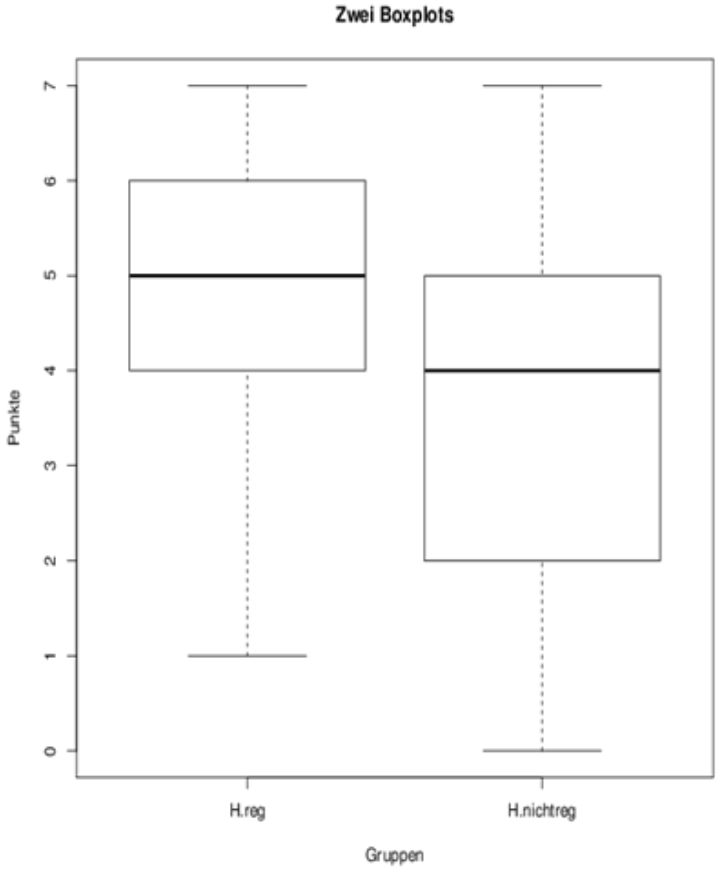
- Zusammenfassung der Lage-Kennzahlen in Box-Plots



Kennzahlen der Lage

Boxplots

- Vergleich von Häufigkeitsverteilungen mit Boxplots



Kennzahlen der Lage

Modus

- Kennzahl der Lage die auch bei nominalen Merkmalen verwendet werden kann
- Idee des Modus: Stellvertreter ist die Merkmalsausprägung mit der größten (relativen) Häufigkeit

Punktezahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenhäufigkeit
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1,000

$x_{mod} = 5$

Kennzahlen der Lage

Modus

Kennzahlen der Streuung

Varianz

- Generelle Überlegung: Lagekennzahl beschreibt die Verteilung mit einer Merkmalsausprägung (stellvertretend für alle) → beschreibt den Charakter der Verteilung nur unzureichend
- Bsp:
  - Einkommen 1: 1.000, 3.000, 4.000, 1.000, 1000
  - Einkommen 2: 1.800, 2.200, 2.400, 1.800, 1.800
  - In beiden Gruppen $\bar{x} = 2000$, aber die Einkommen in der zweiten Verteilung „liegen näher beieinander“ als in der ersten Verteilung
- Idee: Kennzahl für die Streuung als Abstand der Merkmalsausprägungen voneinander oder vom einer fixen Größe

Kennzahlen der Streuung

Varianz

- Idee Varianz: Quadrierte Abweichungen der Merkmalsausprägungen aller Erhebungseinheiten vom Mittelwert bestimmen und davon den Mittelwert berechnen
- Bsp. Einkommen 1: 1.000, 3.000, 4.000, 1.000, 1000 → $\bar{x} = 2000$

→Quadrierte Abweichungen: (1.000-2.000)², (3.000-2.000)², (4.000-2.000)², (1.000-2.000)², (1.000-2.000)²

→Mittelwert berechnen: 8 Mio : 5 = 1,6 Mio. =: Varianz

Kennzahlen der Streuung

Varianz

- Formale Umsetzung der Idee der Varianz:
  - Zeichen für die Varianz: s^2 (sprich „s Quadrat“)

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

- Bsp. Einkommen 2: 1.800, 2.200, 2.400, 1.800, 1800 → $\bar{x} = 2000$

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

$$= \frac{(1.800-2.000)^2 + (2.200-2.000)^2 + (2.400-2.000)^2 + (1.800-2.000)^2 + (1.800-2.000)^2}{5}$$

$$= 64.000$$

Kennzahlen der Streuung

Varianz

- Zusammenfassung der Formeln:

Rohdaten

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

Häufigkeiten

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

Relative Häufigkeiten

$$s^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot p_i$$

Kennzahlen der Streuung

Varianz

- Beispiel 11: Berechnung der Varianz mit Häufigkeiten

Punktezahlen	$(x_i - \bar{x})^2$	Häufigkeit (hi)
0	(0-4,58)² = 20,98	1
1	(1-4,58)² = 12,82	3
2	(2-4,58)² = 6,66	10
3	(3-4,58)² = 2,50	16
4	(4-4,58)² = 0,34	32
5	(5-4,58)² = 0,18	44
6	(6-4,58)² = 2,02	20
7	(7-4,58)² = 5,86	16

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot h_i}{N} = \frac{20{,}98 \cdot 1 + 12{,}82 \cdot 3 + \cdots + 5{,}86 \cdot 16}{142} = 2{,}24$$

Kennzahlen der Streuung

Standardabweichung

- Probleme der Varianz: Quadrierte Abweichungen sind nicht anschaulich im Vergleich zum Mittelwert ($\bar{x} = 2000$ und $s^2 = 1{,}6$ Mio.)
- Idee Standardabweichung: Wurzel aus Varianz bringt Streuungskennzahl auf die selbe Maßeinheit wie die Merkmalsausprägungen und den Mittelwert
- Standardabweichung: $s = \sqrt{s^2}$
- Beispiel 11: $s = \sqrt{2{,}24} = 1{,}5$
- Beispiel Einkommen 1 mit $\bar{x} = 2000$: $s = \sqrt{1.600.000} = 1.264{,}91$ €

Kennzahlen der Streuung

Variationskoeffizient

- Probleme der Varianz und Standardabweichung:
  - sind in bestimmten Einheiten definiert (cm^2 und cm oder $€^2$ und €) → erschwert den Vergleich zwischen Verteilungen mit unterschiedlichen Maßeinheiten
  - Ebenso schwieriger Vergleich von Verteilungen mit unterschiedlichen Mittelwerten
- Bsp. Weitsprungweiten 1: 9m, 10m, 11m → $\bar{x} = 10$, $s^2 = 0{,}67$ und $s = 0{,}82$
- Bsp. Weitsprungweiten 2: 900cm, 1.000cm, 1.100cm → $\bar{x} = 1.000$, $s^2 = 6.666{,}67$ und $s = 81{,}65$
- Idee Variationskoeffizient („sprich v“): Streuung der Merkmale in Relation zum Mittelwert

Kennzahlen der Streuung

Variationskoeffizient

- Formale Umsetzung der Idee des Variationskoeffizienten:
  - Zeichen für Variationskoeffizient: v (sprich „v“)

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

- Bsp. Sprungweiten 1

$\rightarrow v = \frac{0{,}82}{10} = 0{,}082$

- Bsp. Sprungweiten 2

$\rightarrow v = \frac{81{,}65}{1000} = 0{,}082$

Kennzahlen der Konzentration

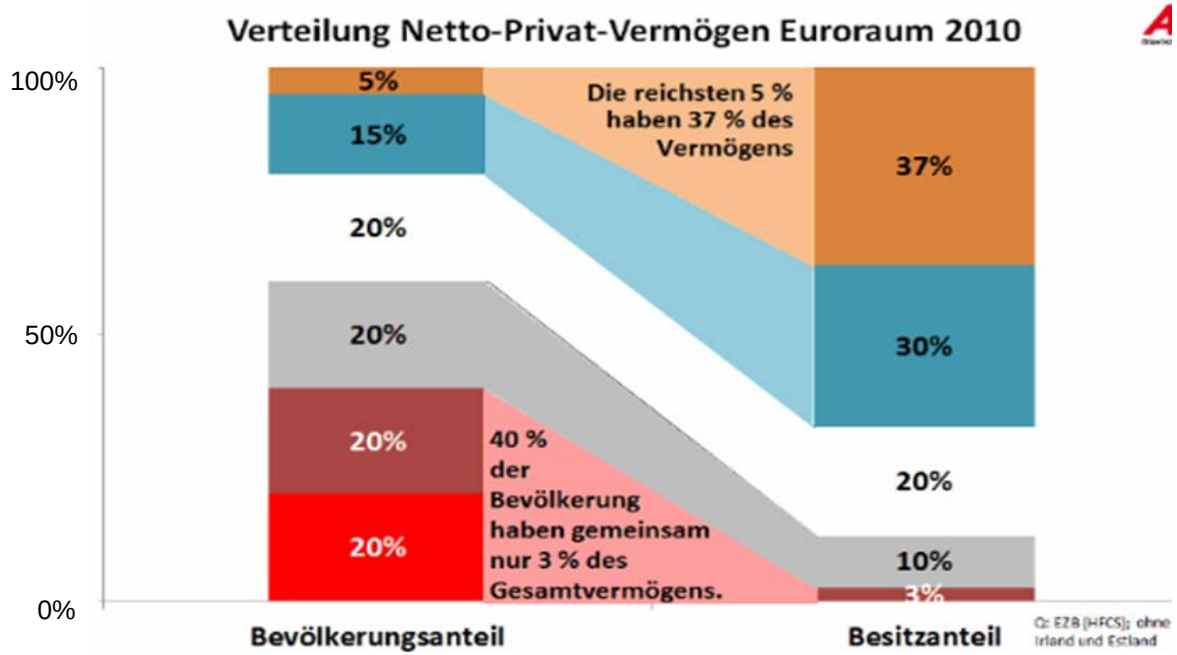
Lorenzkurve

- Generelle Überlegung: Mittelwert und Varianz liefern Aussagen über die gesamte Verteilung, aber wenige Information wie gleichmäßig die Merkmalssumme auf die einzelnen Erhebungseinheiten konzentriert ist.
- Bsp:
  - Einkommen 1: 1.000, 3.000, 4.000, 1.000, 1.000
    mit $\bar{x}$ = 2.000 und s^2 = 1,6 Mio
  - Jede Person erhält jetzt 10.000 € zusätzlich
  - Einkommen 2: 11.000, 13.000, 14.000, 11.000, 11.000
    mit $\bar{x}$ = 12.000 und s^2 = 1,6 Mio
- Frage: Wie gleichmäßig konzentriert sich die Merkmalssumme auf die einzelnen Erhebungseinheiten?

Kennzahlen der Konzentration

Lorenzkurve

- Häufige Anwendung im Bereich der BWL (Marktkonzentration), VWL (Vermögen, Einkommen)



Idee: Gegenüberstellung des Anteils an der Grundgesamtheit und des Anteils an der Merkmalssumme

Kennzahlen der Konzentration

Lorenzkurve

- Beispiel 12: Messung der Konzentration einer Merkmalssumme auf die Erhebungseinheiten

Person	Anteile an Grund-gesamtheit	Kumulierter Anteil an Grund-gesamtheit	Einkom-men	Anteile am Gesamt-einkommen	Kumulierter Anteile am Gesamt-einkommen
A	0,2	0,2	1.000	0,1	0,1
D	0,2	0,4	1.000	0,1	0,2
E	0,2	0,6	1.000	0,1	0,3
B	0,2	0,8	3.000	0,3	0,6
C	0,2	1	4.000	0,4	1
			10.000		

Aussagen: Die ärmsten 40% der Bevölkerung verdienen 20% des Einkommens

Aussagen: Die reichsten 20% der Bevölkerung verdienen 40% des Einkommens

→ Kumulierter Anteil GG (1-0,8)=0,2 vs. Kumulierter Anteil E. (1-0,6)=0,4

Kennzahlen der Konzentration

Lorenzkurve

- Beispiel 12: Messung der Konzentration einer Merkmalssumme auf die Erhebungseinheiten

Person	Anteile an Grund-gesamtheit	Kumulierter Anteil an Grund-gesamtheit	Einkom-men	Anteile am Gesamt-einkommen	Kumulierter Anteile am Gesamt-einkommen
A	0,2		1.000	0,1	
D	0,2		1.000	0,1	
E	0,2		1.000	0,1	
B	0,2		3.000	0,3	
C	0,2		4.000	0,4	
			10.000		

Aussagen: Die ärmsten 40% der Bevölkerung verdienen 20% des Einkommens

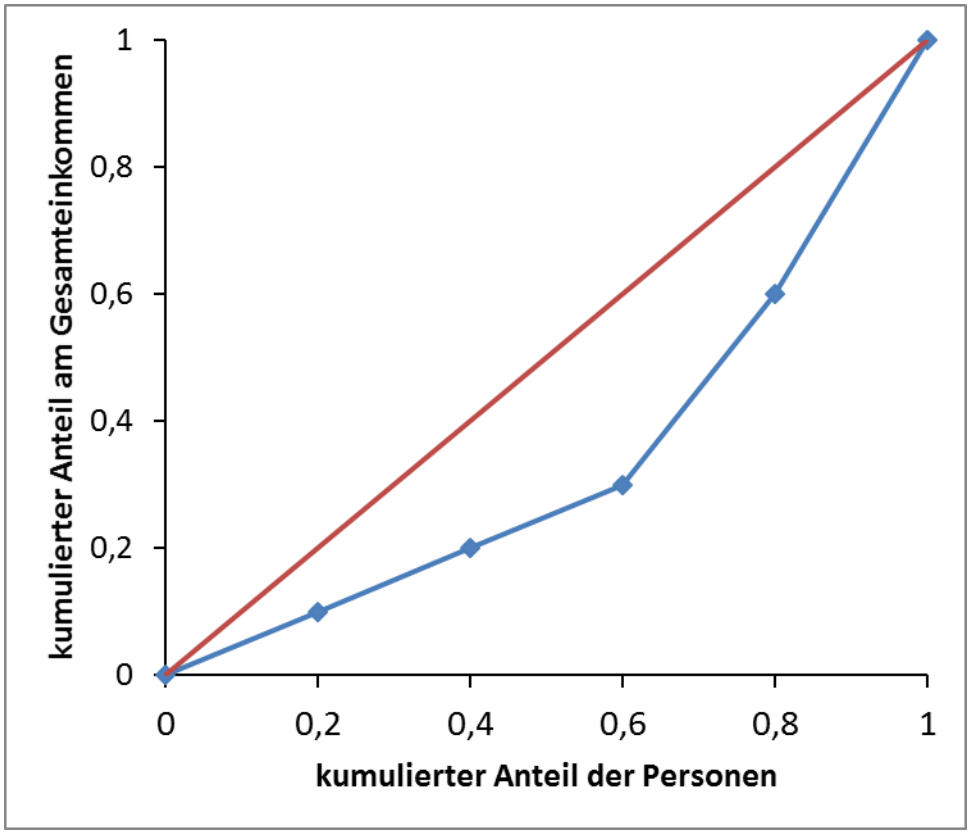
Aussagen: Die reichsten 20% der Bevölkerung verdienen 40% des Einkommens

→ Kumulierter Anteil GG (1-0,8)=0,2 vs. Kumulierter Anteil E. (1-0,6)=0,4

Kennzahlen der Konzentration

Lorenzkurve

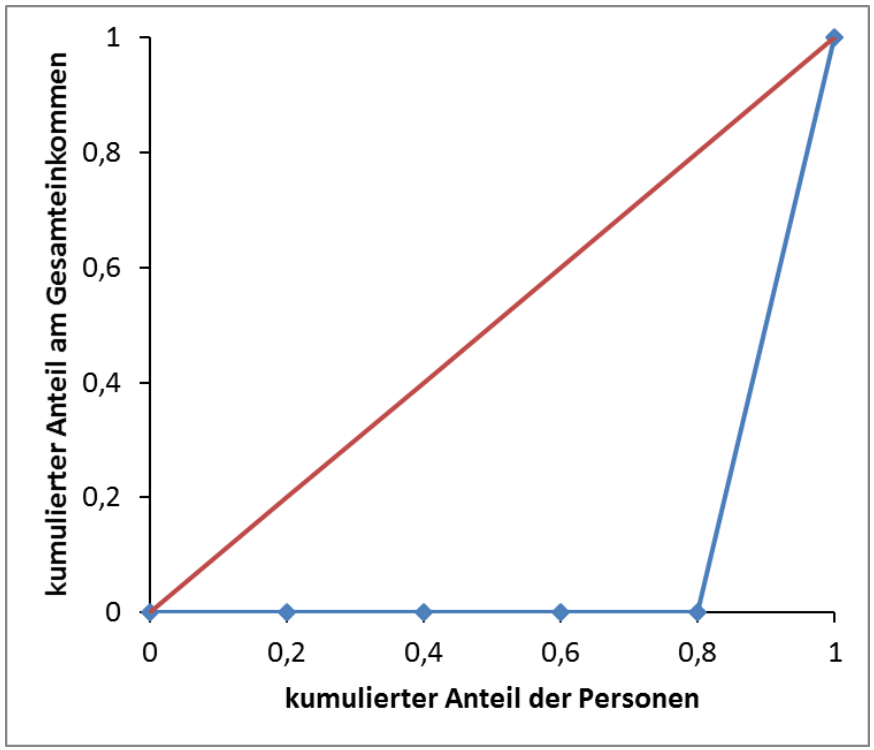
- Graphische Veranschaulichung durch Lorenzkurve



Kennzahlen der Konzentration

Lorenzkurve

- Nullkonzentration vs. Maximalkonzentration



Nullkonzentration: Gleichverteilung der Einkommen über die Erhebungseinheiten

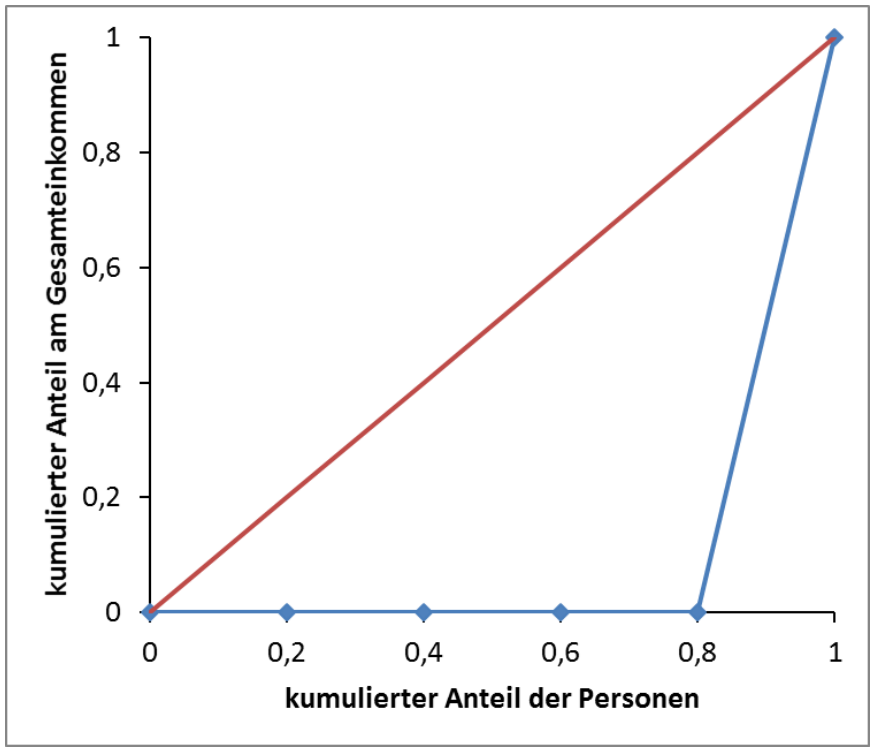
Maximalkonzentration: Eine Erhebungseinheit verdient gesamtes Einkommen, die anderen nichts

Fläche zwischen Diagonale und Lorenzkurve als Maß für die Konzentration der Einkommen

Kennzahlen der Konzentration

Lorenzkurve

- Nullkonzentration vs. Maximalkonzentration



Nullkonzentration: Fläche zwischen Diagonale und Lorenzkurve = 0

Maximalkonzentration: Fläche zwischen Diagonale und Lorenzkurve ist:

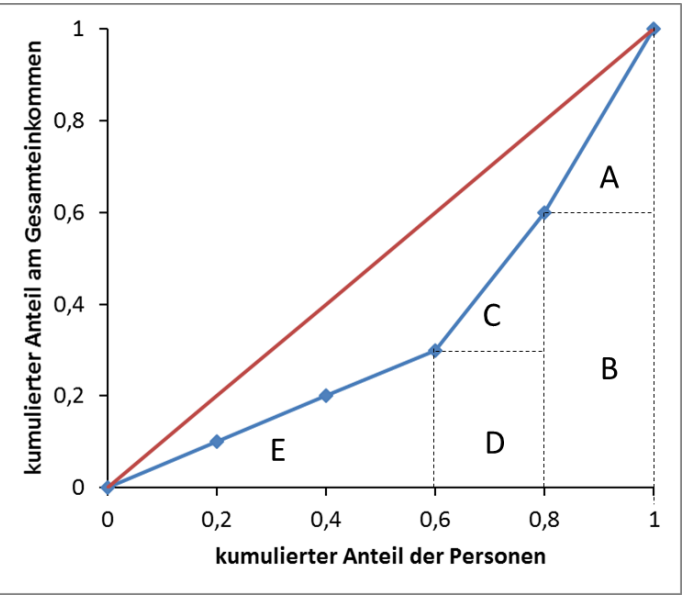
$$\frac{1}{2} - \frac{1 \cdot \frac{1}{N}}{2} = \frac{1}{2} - \frac{1}{2 \cdot N} = \frac{1}{2} \cdot \left(1 - \frac{1}{N}\right)$$

In diesem Beispiel = 0,4

Kennzahlen der Konzentration

Ginikoeffizient

- Bisherige Maßzahl abhängig von N (Anzahl der Erhebungseinheiten)
- Idee normierter Ginikoeffizient: Fläche zwischen Lorenzkurve und der Diagonale dividiert durch maximale Fläche zwischen Lorenzkurve und Diagonale → 0 bei Nullkonzentration, 1 bei Maximalkonzentration



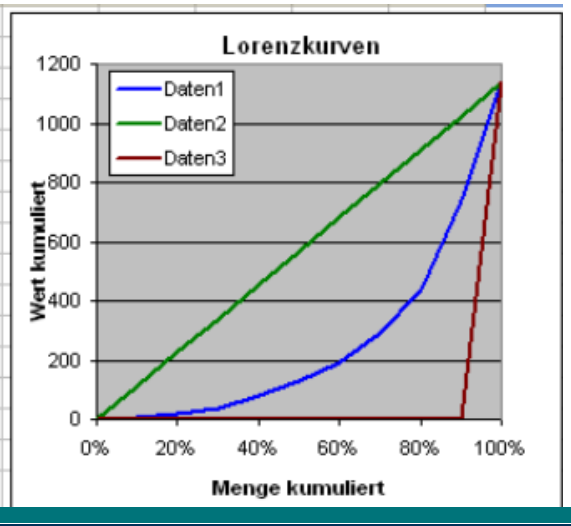
Fläche Lorenzkurve als Summe der Abschnitte auf der x-Achse
Fläche A+B (0,8-1) =(1-0,8)·(1-0,6)/2+(1-0,8)·0,6 = 0,16
Fläche C+D (0,6-0,8) = (0,8-0,6)+
(0,6-0,3)/2(0,8-0,6)·0,3 = 0,09
Fläche E (0-0,6) = (0,6-0)·(0,3-0)/2 = 0,09
Summe Flächen A-E = 0,34
Fläche Diagonale = 1 · 1 / 2 = 0,5
Differenz = 0,5-0,34 = 0,16
Fläche Maximalkonzentration = 0,4
Normierter Ginikoeffizient = 0,16 : 0,4 = 0,4

Kennzahlen der Konzentration

Ginikoeffizient

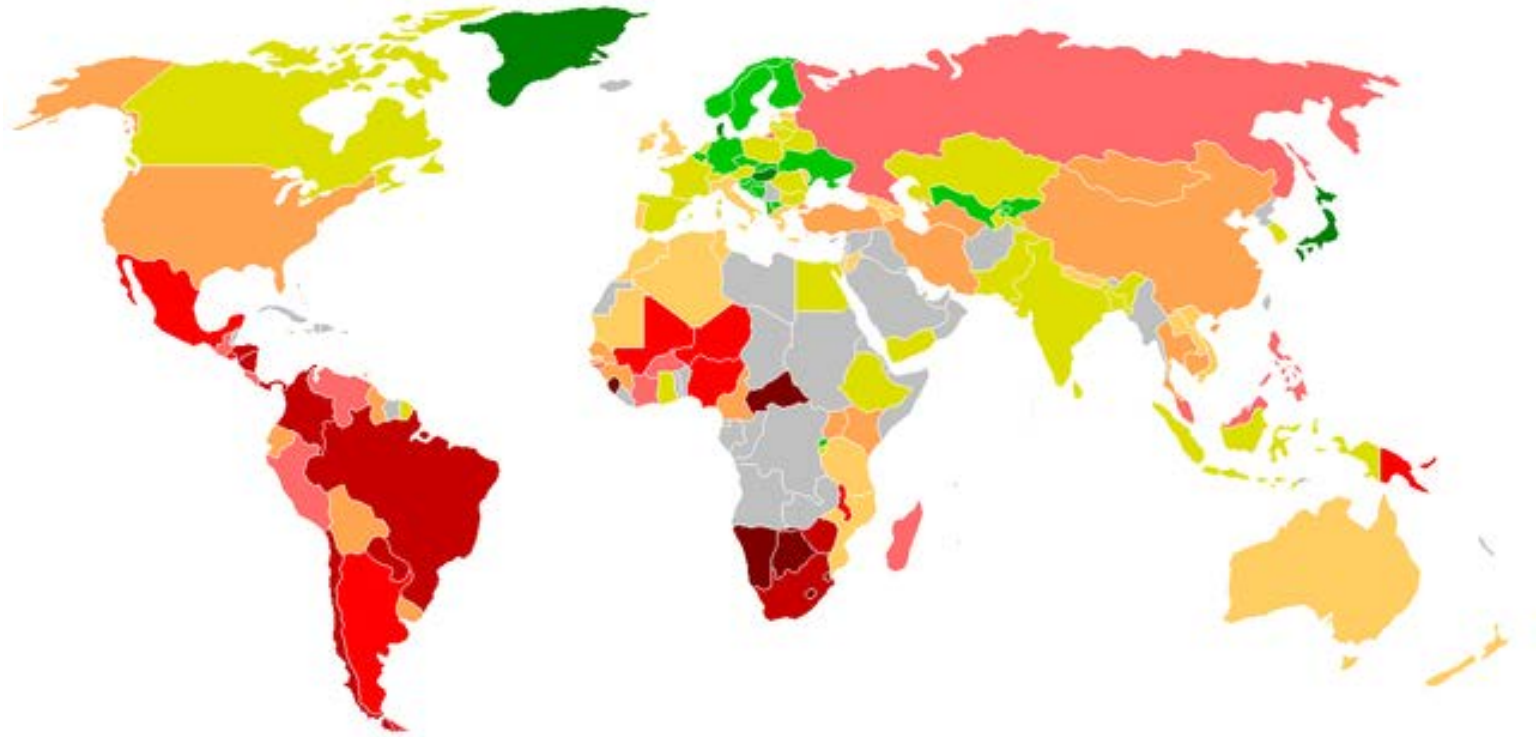
- Zusammenfassung:
  - Lorenzkurve und Ginikoeffizient sind eng verbunden
  - Komplizierte Formel (daher weggelassen)
  - Übungsaufgabe an einfachen Beispielen per Taschenrechner lösbar, bei umfangreicheren Datensätzen Nutzung von Softwareprogrammen empfohlen (auch Excel kann das)

		Daten1	Daten2	Daten3
1				
2		8	113,8	0
3		10	113,8	0
4		20	113,8	0
5		40	113,8	0
6		50	113,8	0
7		60	113,8	0
8		100	113,8	0
9		150	113,8	0
10		300	113,8	0
11		400	113,8	1138
12	G	**0,6247**	**0,0000**	**1,0000**
13				
14				
15				



Kennzahlen der Konzentration

Ginikoeffizient

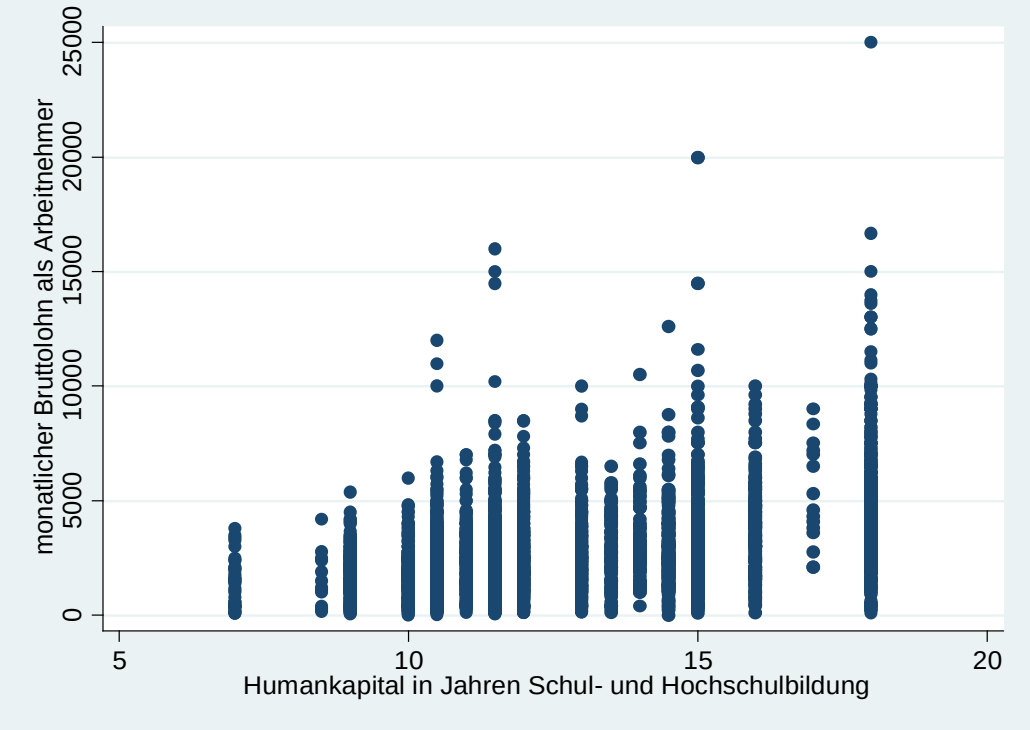


Color	Gini coefficient		0,35 - 0,39		0,55 - 0,59
	< 0,25		0,40 - 0,44		> 0,60
	0,25 - 0,29		0,45 - 0,49		NA
	0,30 - 0,34		0,50 - 0,54		

Kennzahlen des statistischen Zusammenhanges

Was bedeutet Zusammenhang?

- Statistischer Zusammenhang: Die Verteilung eines Merkmals hängt mit der Verteilung eines anderen Merkmals zusammen
- Beispiel: Bildung und Einkommen



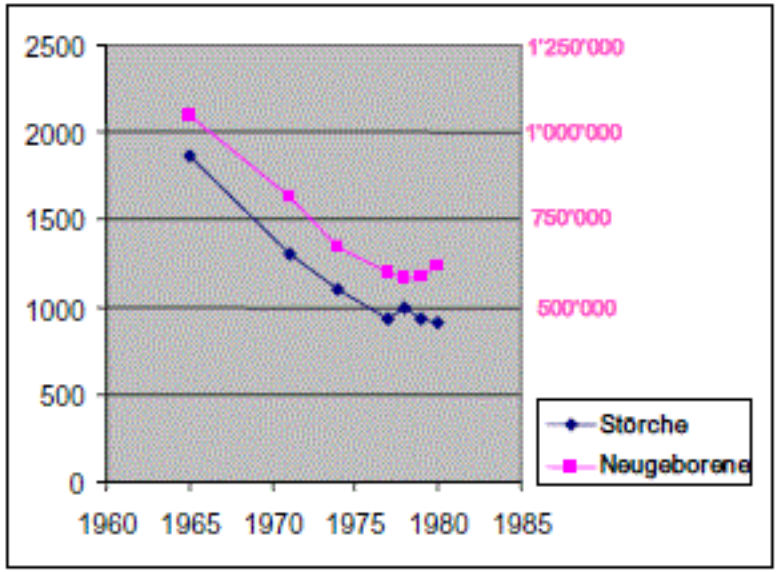
Quelle: SOEP

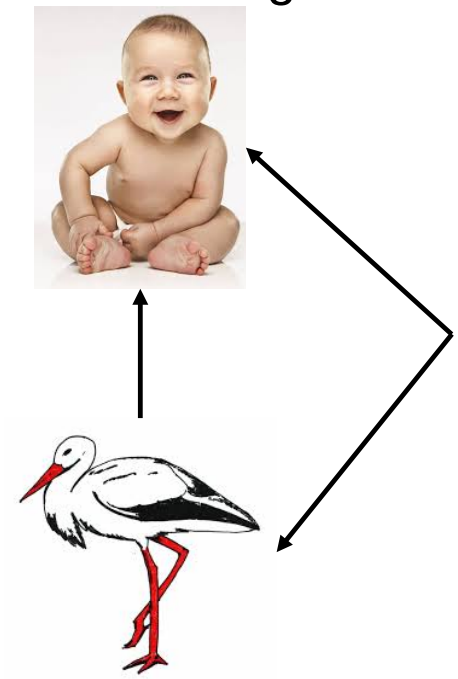
Jahre Bildung	Mittelwert monatlicher Bruttolohn in €
0-10	1.722
11-15	2.522
Größer 15	4.100

Kennzahlen des statistischen Zusammenhanges

Was bedeutet Zusammenhang?

- Kausaler Statistischer Zusammenhang: Die Verteilung eines Merkmals bestimmt ursächlich die Verteilung eines anderen Merkmals (→ Ursache und Wirkung)
- Statistischer vs. kausaler Zusammenhang





Kennzahlen des statistischen Zusammenhanges

Chi-Quadrat

- Beispiel 13: Messung eines Zusammenhangs von nominalen Merkmalen

Absolute Häufigkeiten		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Summe
Geschlecht (i)	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Summe	200	180	50	40	30	500
Relative Häufigkeiten		0,40	0,36	0,10	0,08	0,06	1

Bedingte Häufigkeitsverteilung		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Summe
Geschlecht (i)	weiblich	0,37	0,40	0,07	0,10	0,07	1
	männlich	0,45	0,30	0,15	0,05	0,05	1

Wenn *kein* statistischer Zusammenhang zwischen Geschlecht und Studienrichtung vorliegt → gleiche bedingte Häufigkeitsverteilung unter Frauen und Männern

Kennzahlen des statistischen Zusammenhanges

Chi-Quadrat

- Idee: Wenn *kein* statistischer Zusammenhang → erwartete relative Häufigkeiten

Beobachtete absolute Häufigkeiten h_{ij}^b		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $N_{i.}$
Geschlecht (i)	weiblich	110	120	20	30	20	300
	männlich	90	60	30	10	10	200
	Randv. $N_{.j}$	200	180	50	40	30	500

Beobachtete relative Häufigkeiten p_{ij}^b		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $p_{i.}$
Geschlecht (i)	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Randv. $p_{.j}$	0,40	0,36	0,10	0,08	0,06	1

22% weibliche BWL Studenten. Wenn das Geschlecht nicht die Studienwahl beeinflussen würde, würden sowohl 40% der Männer als auch Frauen BWL studieren. Welchen Anteil an weiblichen BWL Studenten würde man erwarten? → 60% weibliche Studenten · 40% BWL Studium = 24%

$$p_{ij}^e = p_{i.}^b \cdot p_{.j}^b$$

Erwartete relative Häufigkeiten p_{ij}^e		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $p_{i.}$
Geschlecht (i)	weiblich	0,24	0,216	0,06	0,048	0,036	0,60
	männlich	0,16	0,144	0,04	0,032	0,024	0,40
	Randv. $p_{.j}$	0,40	0,36	0,10	0,08	0,06	1

Kennzahlen des statistischen Zusammenhanges

Chi-Quadrat

- Idee Chi-Quadrat: Verwendung der Differenzen der beobachteten und der bei Fehlen eines Zusammenhangs erwarteten (relativen) Häufigkeiten

Beobachtetet relative Häufigkeiten p_{ij}^b		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Randv. $p_{i.}$
Geschlecht (i)	weiblich	0,22	0,24	0,04	0,06	0,04	0,60
	männlich	0,18	0,12	0,06	0,02	0,02	0,40
	Randv. $p_{.j}$	0,40	0,36	0,10	0,08	0,06	1

Erwartete relative Häufigkeiten p_{ij}^e		Studienrichtung (j)					
		BWL	Soz	VWL	Sowi	Stat	Summe
Geschlecht (i)	weiblich	0,24	0,216	0,06	0,048	0,036	0,60
	männlich	0,16	0,144	0,04	0,032	0,024	0,40
	Summe	0,40	0,36	0,10	0,08	0,06	1

Je größer die Differenz zwischen beobachteter und erwarteter Häufigkeit desto stärker ist der Zusammenhang zwischen Geschlecht und Studienwahl → Zusammenfassung der Differenzen in einer Kennzahl!

Kennzahlen des statistischen Zusammenhanges

Chi-Quadrat

- Formale Umsetzung Chi-Quadrat:

$$\chi^2 = N \cdot \sum \frac{\left(p_{ij}^b - p_{ij}^e\right)^2}{p_{ij}^e}$$

Wenn die Merkmale nicht statistisch zusammenhängen χ^2=0
Beispiel 13:

$$\chi^2 = 500 \cdot \left[\frac{(0{,}22 - 0{,}24)^2}{0{,}24} + \frac{(0{,}24 - 0{,}216)^2}{0{,}216} + \cdots\right] = 18{,}06$$

→ Normierung notwendig

Kennzahlen des statistischen Zusammenhanges

Cramers V

- Idee und formelles Umsetzung Cramers V: Chi-Quadrat normieren so dass Kennzahl zwischen 0 (kein Zusammenhang) und 1 (vollständiger Zusammenhang) liegt

$$V = \sqrt{\frac{\chi^2}{N \cdot (\min(k, l) - 1)}}$$

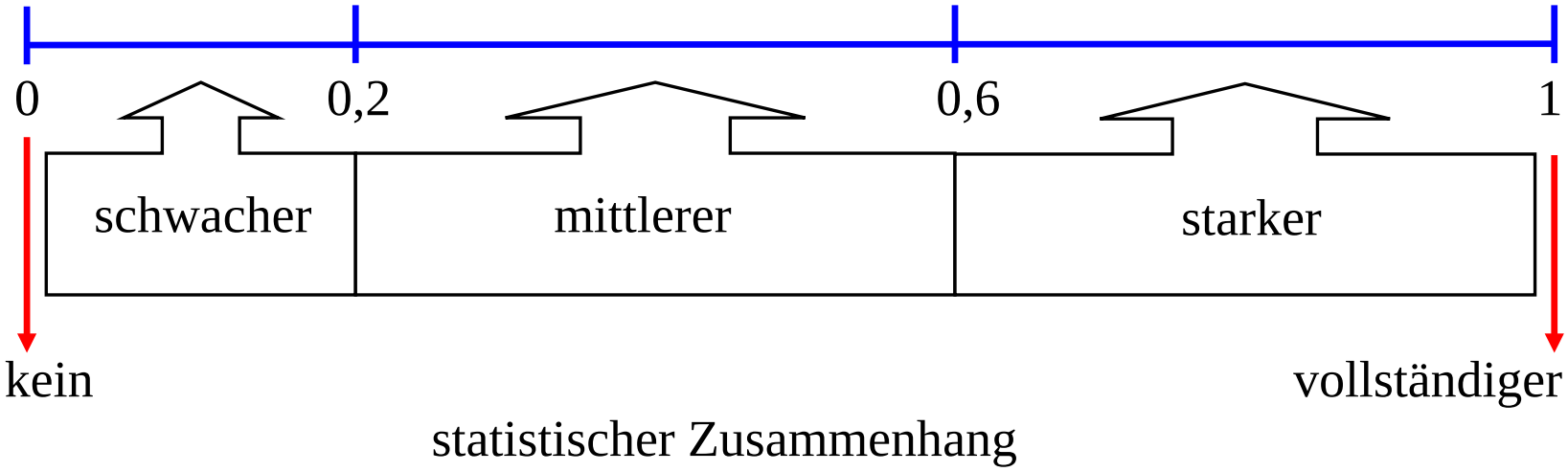
k, l = die Anzahl der Merkmalsausprägungen der beiden Merkmale;
min(k,l) = die kleinere der beiden Anzahlen
Beispiel 13:

$$V = \sqrt{\frac{18{,}06}{500 \cdot (2 - 1)}} = 0{,}19$$

Kennzahlen des statistischen Zusammenhanges

Cramers V

- Interpretation Cramers V (Faustregeln)

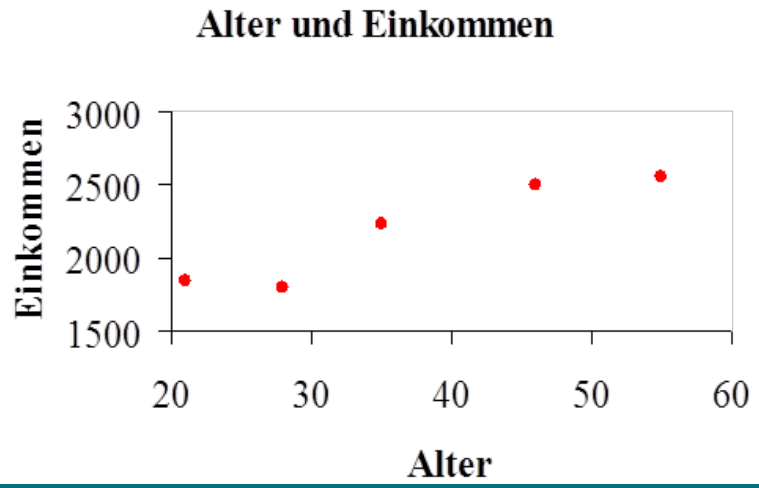


Kennzahlen des statistischen Zusammenhanges

Kovarianz

- Generelle Überlegung: Metrische Merkmale erlauben genauere Aussagen zum Zusammenhang
- Beispiel 14: Erhebung von zwei metrischen Merkmalen

Person	A	B	C	D	E
Alter	21	46	55	35	28
Einkommen	1.850	2.500	2.560	2.230	1.800

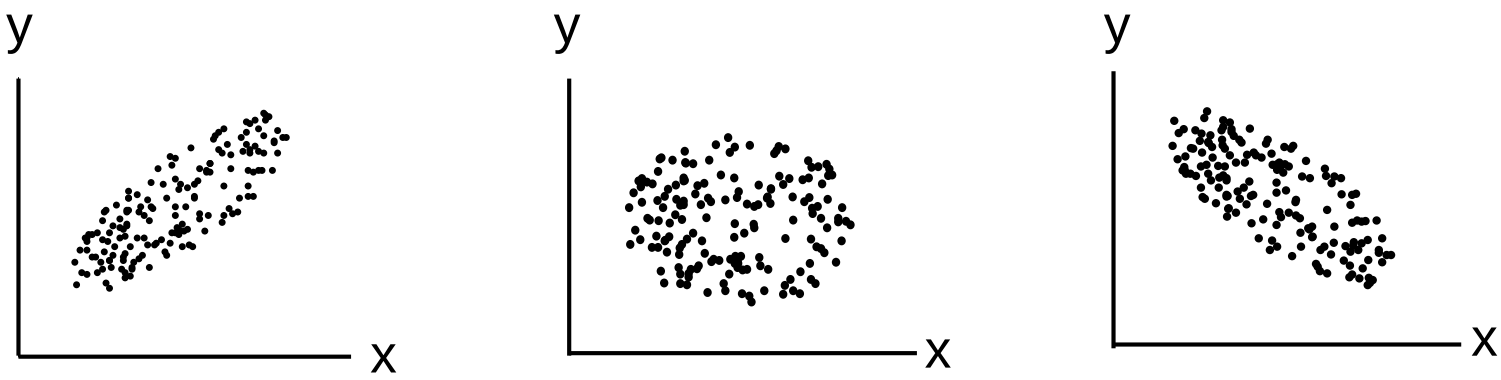


Streudiagramm

Kennzahlen des statistischen Zusammenhanges

Kovarianz

- Idee: Richtung des Zusammenhangs sollte sich im Vorzeichen der Kennzahl widerspiegeln (positiv mit >0, kein Zusammenhang =0, negativ <0)
- Beispiel: Drei Streudiagramme für beliebige Merkmale x und y

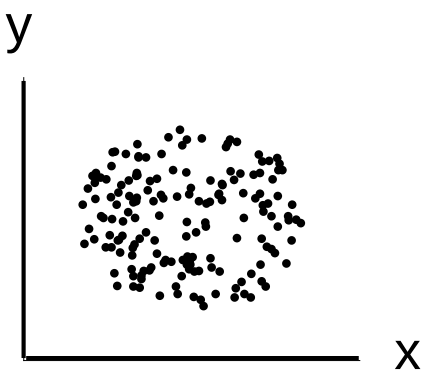


positiv | keine Richtung | negativ

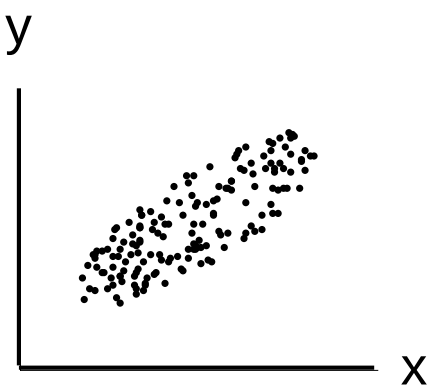
Kennzahlen des statistischen Zusammenhanges

Kovarianz

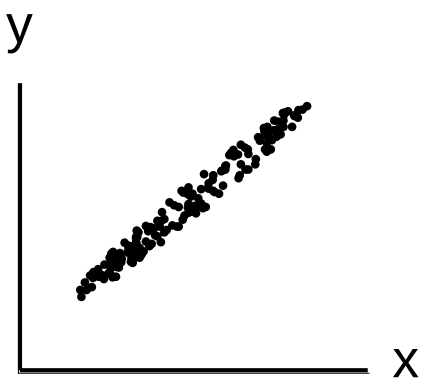
- Idee: Stärke des Zusammenhangs sollte sich in der Größe der Kennzahl widerspiegeln (starker Zusammenhang mit großen Werten, kleiner Zusammenhang mit kleineren Werten)
- Beispiel: Drei Streudiagramme für beliebige Merkmale x und y



kein Zusammenh.



mittlerer Zusammenh.



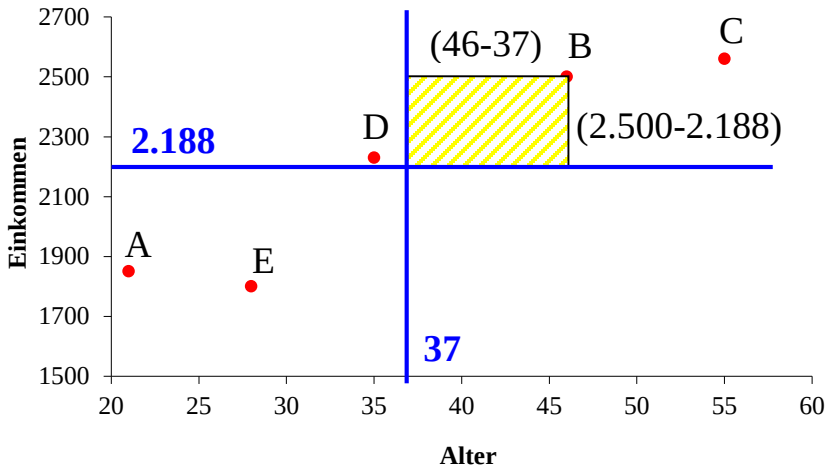
starker Zusammenh.

Kennzahlen des statistischen Zusammenhanges

Kovarianz

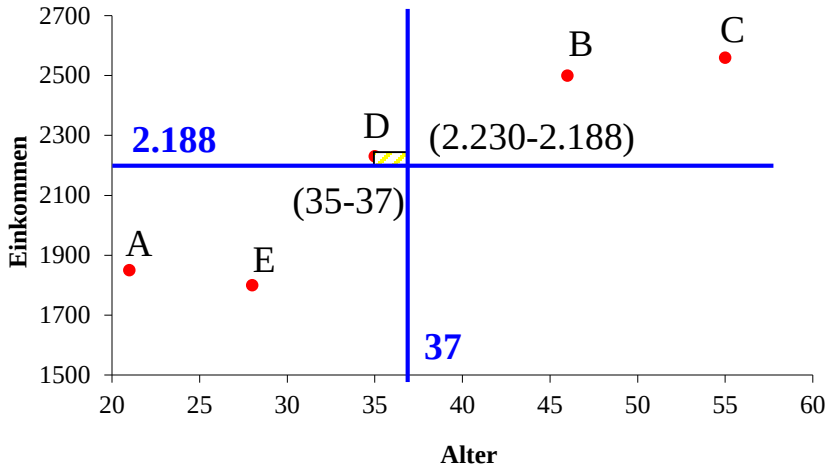
- Idee: Berechnung des Produkts der Abweichungen vom Mittelwert für jede Erhebungseinheit $(x_i - \bar{x}) \cdot (y_i - \bar{y})$

Punkt B



Punkte B, C, A und E mit positiven gerichteten Rechtecksflächen

Punkt D



Punkte D mit negativer gerichteter Rechtecksfläche

Kennzahlen des statistischen Zusammenhanges

Kovarianz

- Formale Umsetzung der Kovarianz (s_{xy}): Mittelwert der gerichteten Rechtsecksflächen $(x_i - \bar{x}) \cdot (y_i - \bar{y})$

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

Im Beispiel 14

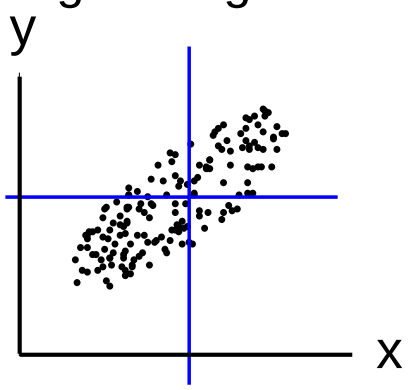
$$s_{xy} = \frac{(21 - 37) \cdot (1.850 - 2188) + \cdots + (28 - 37) \cdot (1.800 - 2188)}{5} = 3.664$$

Kennzahlen des statistischen Zusammenhanges

Kovarianz

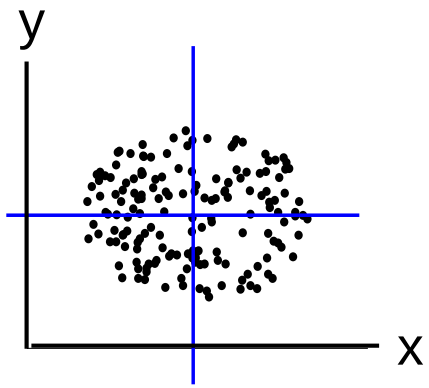
$$s_{xy} = \frac{\sum_{i=i}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

Idee: Richtung der Kovarianz sollte sich im Vorzeichen widerspiegeln
Check: *Vorzeichen* der Kovarianz abhängig vom Ausmaß positiver und negativer gerichteter Rechtecksflächen



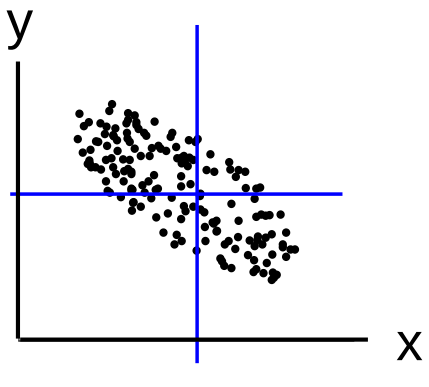
positiv

$s_{xy} > 0$



keine Richtung

$s_{xy} = 0$



negativ

$s_{xy} < 0$

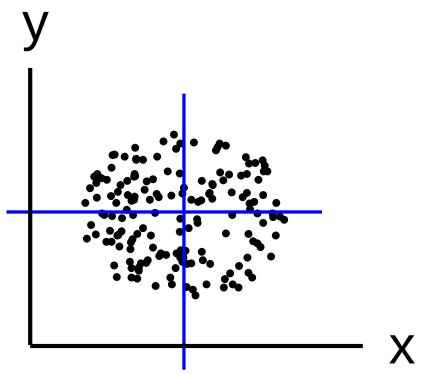
Kennzahlen des statistischen Zusammenhanges

Kovarianz

$$s_{xy} = \frac{\sum_{i=i}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

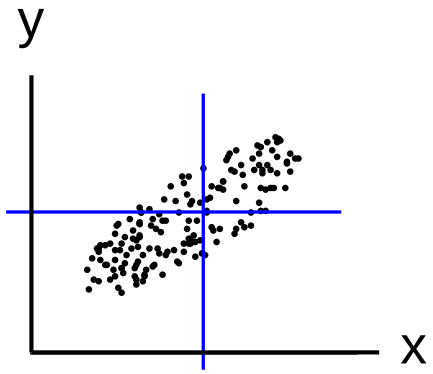
Idee: Stärke des Zusammenhangs sollte sich in der Größe der Kennzahl widerspiegeln

Check: *Größe* der Kovarianz abhängig von der Größe der Rechtecksflächen



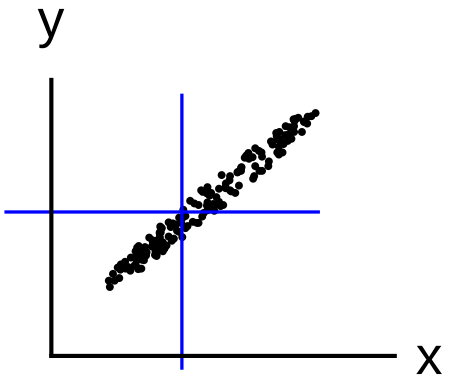
kein Zusammenh.

$s_{xy} = 0$



mittlerer Zusammenh.

$s_{xy} > 0$



starker Zusammenh.

$s_{xy} \gg 0$

Kennzahlen des statistischen Zusammenhanges

Korrelationskoeffizient

- Problem der Kovarianz (s_{xy}): nicht beschränkter Wertebereich → erschwert den Vergleich von verschiedenen Kovarianzen → Normierung notwendig
- Idee Korrelationskoeffizient: Division der Kovarianz durch das Produkt der Standardabweichungen der Merkmale → Messung des linearen statistischen Zusammenhanges

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

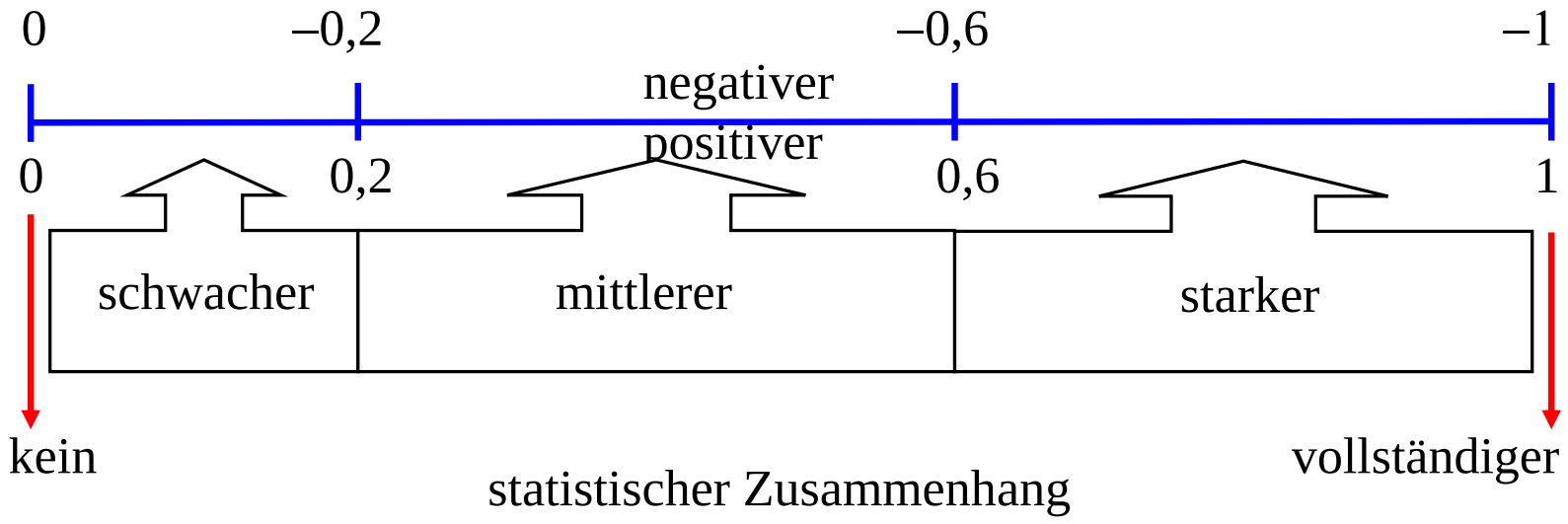
Im Beispiel 14

$$r = \frac{3.664}{12{,}21 \cdot 316{,}95} = 0{,}947$$

Kennzahlen des statistischen Zusammenhanges

Korrelationskoeffizient

- r schwankt zwischen -1 und +1
- Interpretation r (Faustregeln)



Kennzahlen des statistischen Zusammenhanges

Spearmanscher Rangkorrelationskoeffizient

- Problem des Korrelationskoeffizienten: nur bei metrischen Merkmalen, aber nicht bei ordinalen Merkmalen (z.B. Noten, Rangstufen)

Beispiel 16: Erhebung von zwei ordinalen Merkmalen

Studierende	A	B	C	D	E	F
Mathe-Note x	1	1	5	5	4	2
Statistik-Note y	2	2	5	4	4	3

r=0,96

Andere Kodierung der Noten: 1, 10, 100, 1.000, 10.000

Studierende	A	B	C	D	E	F
Mathe-Note x	1	1	10000	10000	1000	10
Statistik-Note y	10	10	10000	1000	1000	100

r=0,68

Idee Rangkorrelationskoeffizient: Korrelation der Ränge der Erhebungseinheiten (anstatt der Merkmalsausprägungen)

Kennzahlen des statistischen Zusammenhanges

Spearmanscher Rangkorrelationskoeffizient

Studierende	A	B	C	D	E	F
Mathe-Note x	1	1	5	5	4	2
Statistik-Note y	2	2	5	4	4	3

Studierende	A	B	C	D	E	F
Mathe-Rang u	1,5	1,5	5,5	5,5	4	3
Statistik-Rang v	1,5	1,5	6	4,5	4,5	3

Korrelationskoeffizient der Rangzahlen ist unabhängig von der gewählten Kodierung → Interpretation wie normaler Korrelationskoeffizient

$$r = \frac{s_{uv}}{s_u \cdot s_v}$$

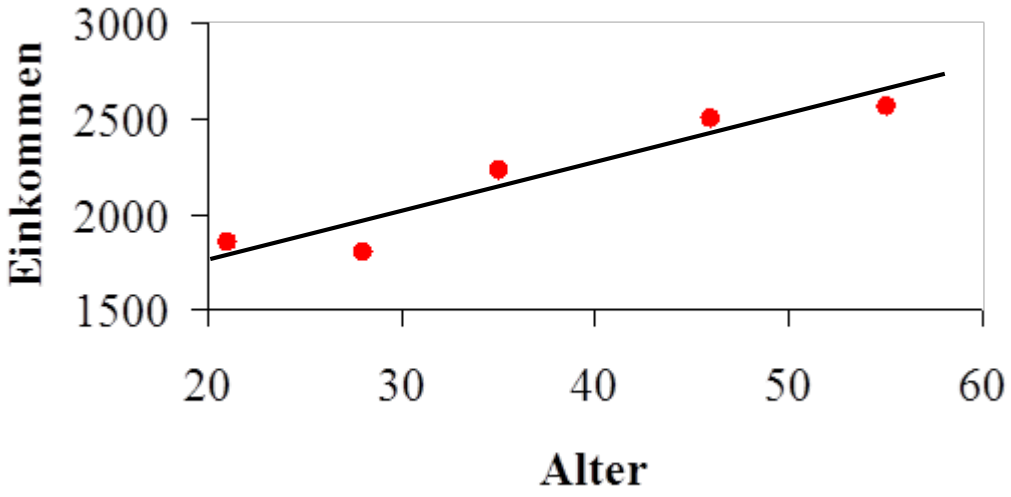
Im Beispiel: $r = \frac{2{,}625}{1{,}658 \cdot 1{,}658} = \frac{2{,}625}{2{,}75} = 0{,}955$

Kennzahlen des statistischen Zusammenhanges

Regressionsrechnung

- Überlegung: Kann man mit der Kenntnis über die Richtung und Größe des statistischen Zusammenhangs weitergehende Aussagen (z.B. kausale Aussagen und Prognosen) treffen
- Bestimmung einer Funktion die den Zusammenhang zweier Merkmale in einer Funktion (zumeist lineare Gleichung) erfasst

→Regressionsgerade als Gerade die „am Nächsten zu den Punkten“ liegt



Kennzahlen des statistischen Zusammenhanges

Regressionsrechnung

- Formale Umsetzung: Methode der kleinsten Quadrate (Extremwertaufgabe) ergibt die Gleichung der Regressionsgerade

$$y = b_1 \cdot x + b_2$$

Mit dem Regressionskoeffizienten (Steigung): $b_1 = \frac{s_{xy}}{s_x^2}$

und der Konstante (Achsenabschnitt): $b_2 = \bar{y} - b_1 \cdot \bar{x}$

Beispiel 15: Berechnung der Gleichung der Regressionsgeraden:

$b_1 = \frac{3.664}{149{,}2} = 24{,}6$ und $b_2 = 2.188 - 24{,}6 \cdot 37 = 1.279{,}4$

Regressionsgerade: $y = 24{,}6 \cdot x + 1.279{,}4$

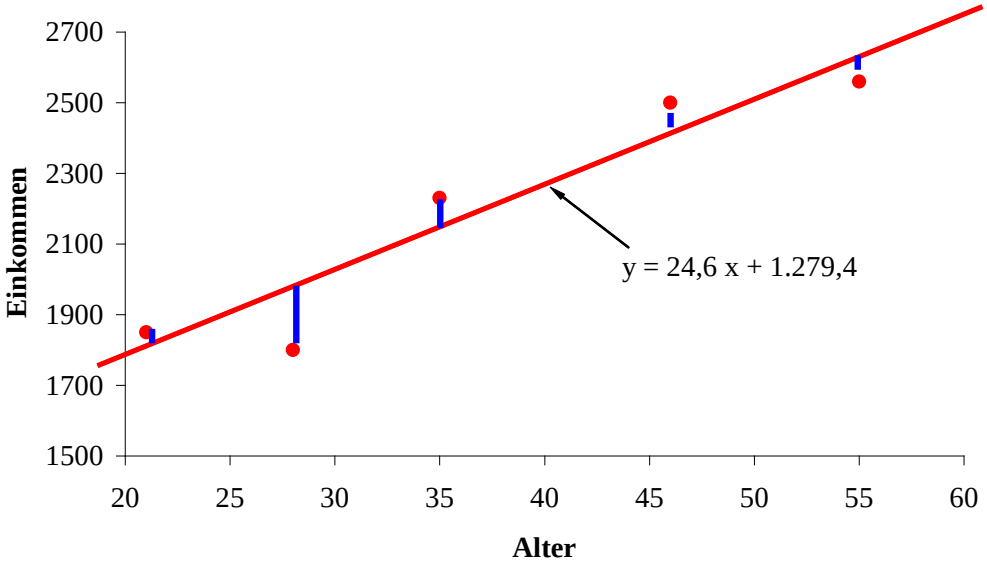
Achtung weitere Bezeichnungen: y = Regresssand oder abhängiges Merkmal oder Variable; x = Regression oder unabhängiges Merkmal oder Variable

Kennzahlen des statistischen Zusammenhanges

Regressionsrechnung

Für jede Erhebungseinheit gibt es
2 Werte: z.B. für Person A
1) Beobachtetes Einkommen:
Person A = 1.850

2) Geschätztes Einkommen
Person A = 24,6·21+1.279,4 = 1.796



Differenz zwischen beobachtetem Einkommen und geschätztem Einkommen ist Residuum → ein Teil des beobachteten Einkommens kann nicht durch die Regressionsfunktion erklärt werden → man macht einen Fehler bei der Schätzung!!!
z.B. Person A: 1.850 - 1.796 = 54

→ Summe der quadrierten Residuen ist ein Maß für die Güte der Regression

Kennzahlen des statistischen Zusammenhanges

Regressionsrechnung

- Verwendung der Regressionsgeraden zur Schätzung fehlender Werte bzw. zur Prognose

  Regressionsgerade: $y = 24{,}6 \cdot x + 1.279{,}4$

  Bsp: Alter $x = 40$ → Einkommen $y = 24{,}6 \cdot 40 + 1.279{,}4 = 2.263{,}4\ €$

- Vertrauen in die Schätzung

  Bestimmtheitsmaß B: $\boxed{B = r^2}$

  Im Beispiel 15: $B = 0{,}947^2 = 0{,}897$

  B gibt den Anteil der durch die Regression erklärten Varianz der abhängigen Variable an („Erklärungsanteil“)