Vorlesung: Statistik I für Wirtschaftswissenschaft

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

WiSe 2016/2017

Termine und Informationen

Homepage:

`http://www.stablab.stat.uni-muenchen.de/WiwiStat1_1617`

Vorlesung:

Prof. Helmut Küchenhoff

Di 16:00 - 18:00 Audi max

Übung (wöchentlich):

Ansprechperson: Veronika Deffner

Übung 1:	Mi. 12.15 - 13.45 Uhr	Schellingstr. 3, S 003
Übung 2:	Mi. 14.15 - 15.45 Uhr	Schellingstr. 3, S 001
Übung 3:	Do. 10.15 - 11.45 Uhr	Schellingstr. 3, S 001
Übung 4:	Do. 10.15 - 11.45 Uhr	Schellingstr. 3, S 002
Übung 5:	Do. 12.15 - 13.45 Uhr	Schellingstr. 3, S 001
Übung 6:	Do. 12.15 - 13.45 Uhr	Schellingstr. 3, S 002
Übung 7:	Do. 18.00 - 19.30 Uhr	Schellingstr. 3, S 001

Literatur

L.Fahrmeir, Ch. Heumann, R.Künstler, I.Pigeot, G.Tutz:
Statistik - Der Weg zur Datenanalyse
Springer-Verlag, 8. Auflage, 2016

H.Toutenburg, C.Heumann:
Deskriptive Statistik - Eine Einführung in Methoden und Anwendungen mit R und SPSS
Springer-Verlag, 2009

Dank

an Christian Heumann für Materialien und Folien

8 Komplexe Zusammenhänge

9 Regression und Mittelwertsvergleiche

10 Verhältniszahlen und Indizes

11 Zeitreihen

12 Wahrscheinlichkeit

- **Einführung: Was ist Statistik?**

Statistik

- Businesses are collecting more data than they know what to do with. To turn all this information into competitive gold, they'll need new skills and a new management style.
- Data-driven decisions are better decisions—it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition.

Aus: Andrew McAfee and Erik Brynjolfsson: Big Data: The Management revolution. Harvard Business Review October 2012, 60-68.

Statistik für BWL und VWL ?

- Datenanalyse ermöglicht es Unternehmen, die Wertschöpfungskette an allen Stellen zu optimieren, Im Einkauf, im Marketing, in Verkauf, Preisgestaltung und Management.
- Viele moderne Unternehmen haben als wichtigsten Wert Daten und Informationen (Google, Facebook)
- Statistische Methoden sind ein zentrales Hilfsmittel zur Analyse und Prognose von volkswirtschaftlichen Daten
- Auswertung von Maßnahmen von Regierungen und Institutionen werden mit statistischen Methoden bewertet

HR Analystics

- Nutzung von Befragungsdaten von MitarbeiterInnen zur Reduktion von Kündigungen
- Einstellungsstrategien aus Performance-Daten (talent analytics and big data) zur Effizienzsteigerung
- Xerox used big data to reduce the attrition rate in its call centers by 20%. To do that, it had to understand what was causing the turnover, and determine ways to improve employee engagement.

Marketing

- Flexible Preisgestaltung über Internet- Verkauf ermöglicht viele Strategien
- Projekt mit Fluglinie zur Daten- gesteuerten Preisgestaltung
- Experimente mit verschieden gestalteten Internetseiten
- Erfolge von Mailing-Aktionen abhängig von guter Datenanalyse

Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

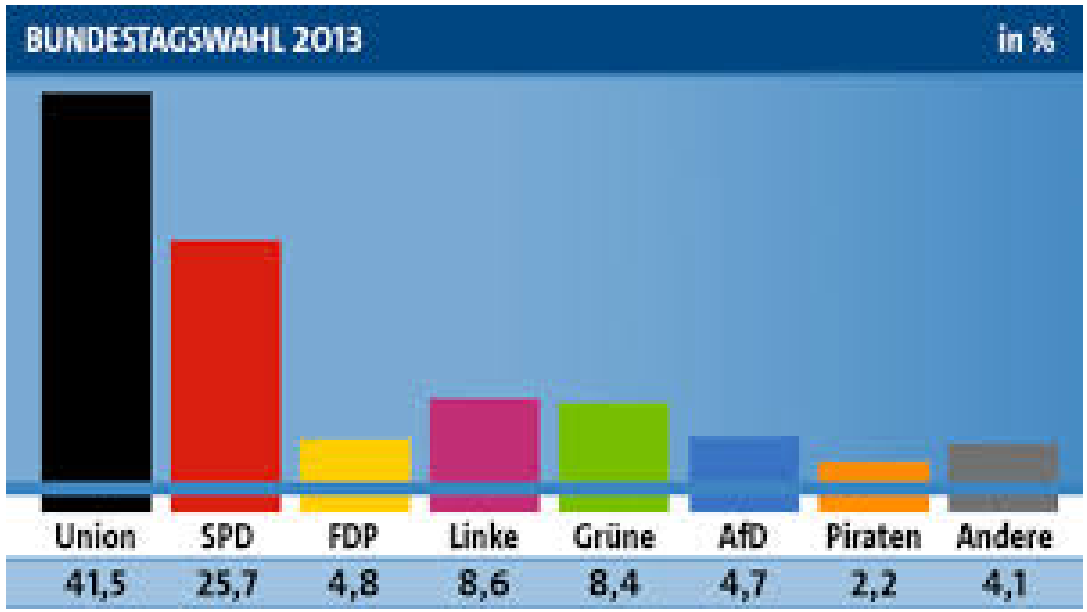
CDU/CSU	SPD	FDP	Linke	Grüne	AFD
42,0	26,0	4,7	8,5	8,0	4,9

Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

CDU/CSU	SPD	FDP	Linke	Grüne	AFD
42,0	26,0	4,7	8,5	8,0	4,9

Ergebnis:

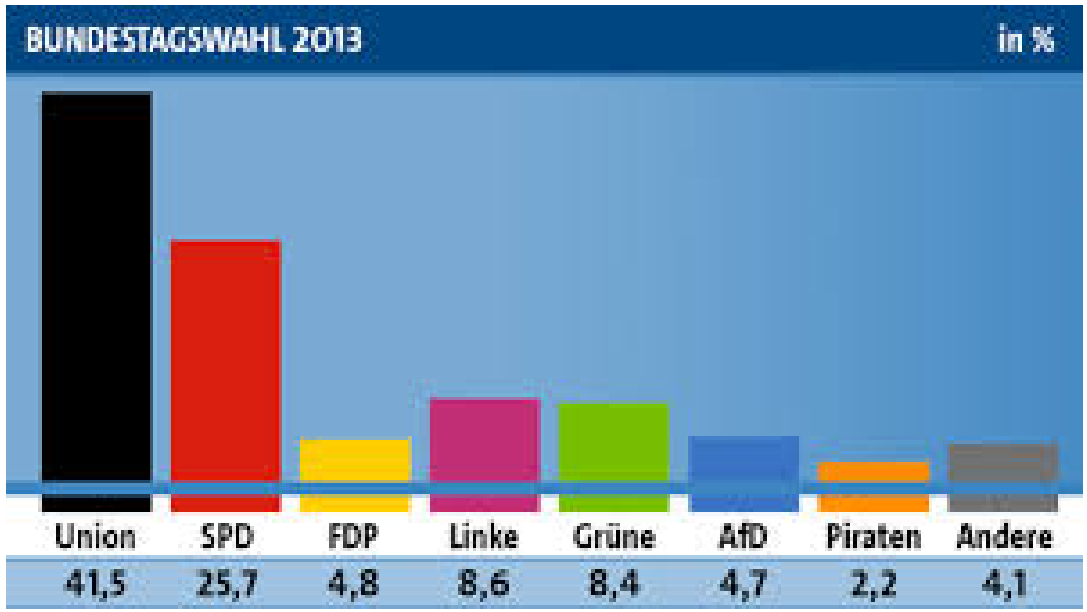


Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

CDU/CSU	SPD	FDP	Linke	Grüne	AFD
42,0	26,0	4,7	8,5	8,0	4,9

Ergebnis:

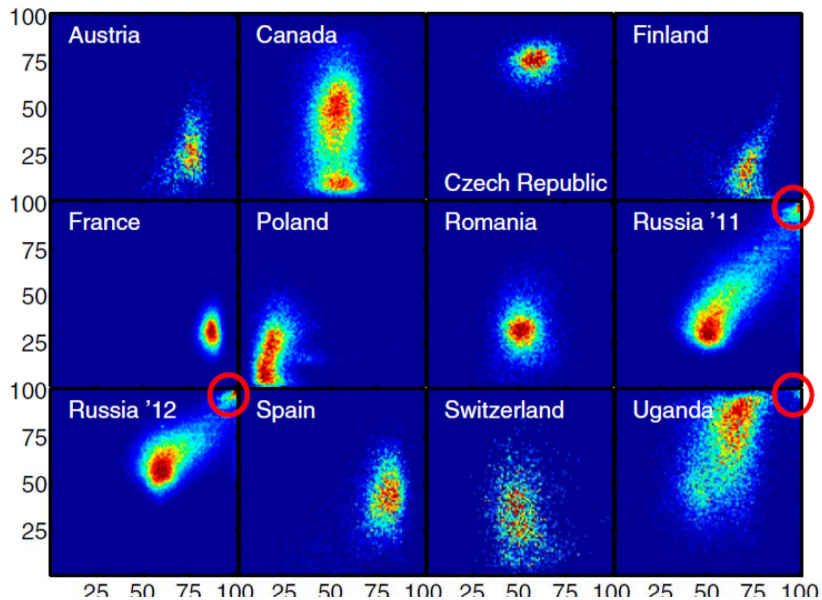


Basis: Nachwahlbefragung 100 000 Wahlberechtigte

`http://wahl.tagesschau.de/wahlen/2013-09-22-BT-DE/index.shtml`

Beispiel 2: Wahlfälschung

Arbeit von Klimek et al.
Einfache Idee: Untersuche Zusammmenhang zwischen Wahlergebnis (Stimmenanteil des Siegers) gegen die Wahlbeteiligung.



Beispiel 3: Lebenszufriedenheit und Alter

Gibt es eine Midlife Crisis?
Analysen von Panel-Daten zur subjektiven Lebenszufriedenheit mit semiparametrischen Regressionsmodellen
In Zusammenarbeit mit Sonja Greven, Andrea Wiencierz, Christoph Wunder

C. Wunder, A. Wiencierz, J. Schwarze, and H. Küchenhoff. Well-being over the Life Span: Semiparametric evidence from British and German Longitudinal Data. *Review of Economics and Statistics* 95(1):154-167, 2013.

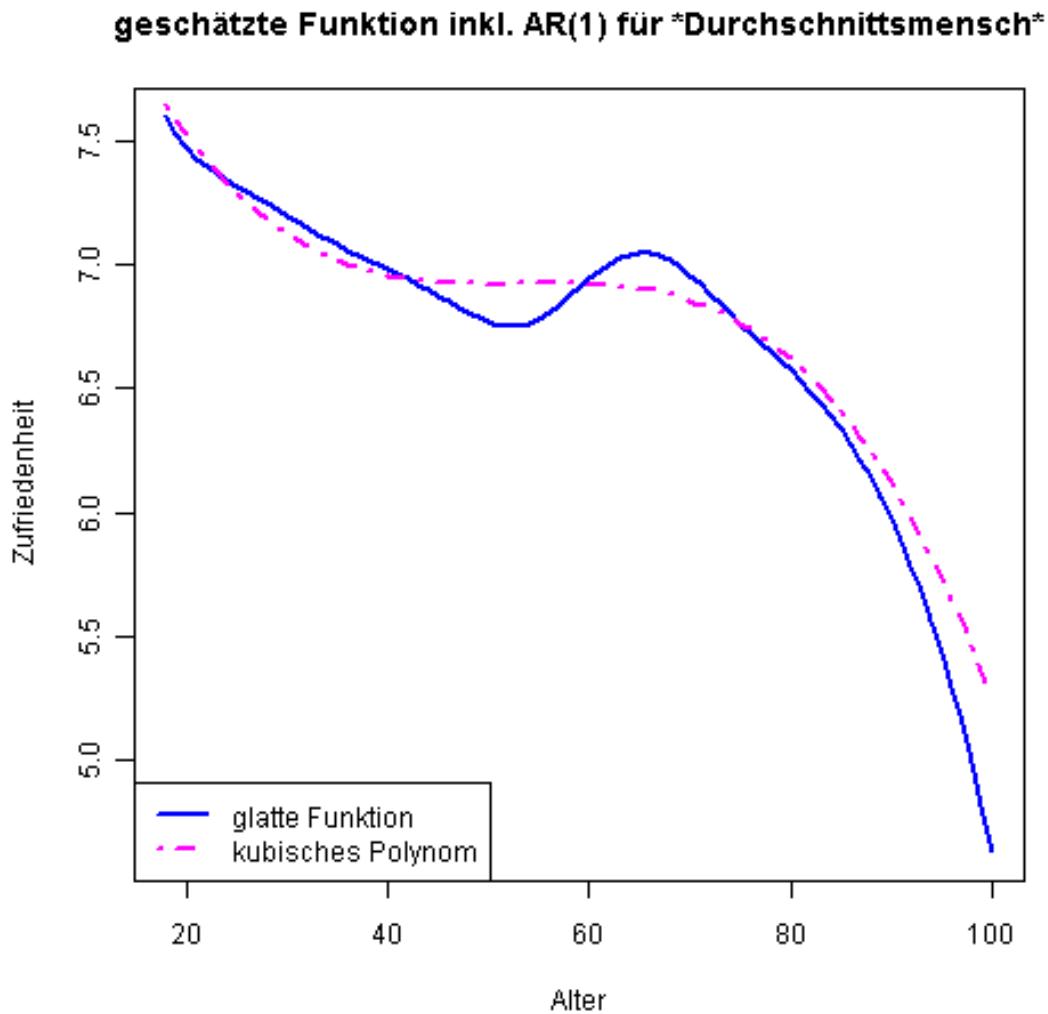
A. Wiencierz, S. Greven, and H. Küchenhoff. Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic Journal of Statistics* 5:1718-1734, 2011.

Datengrundlage

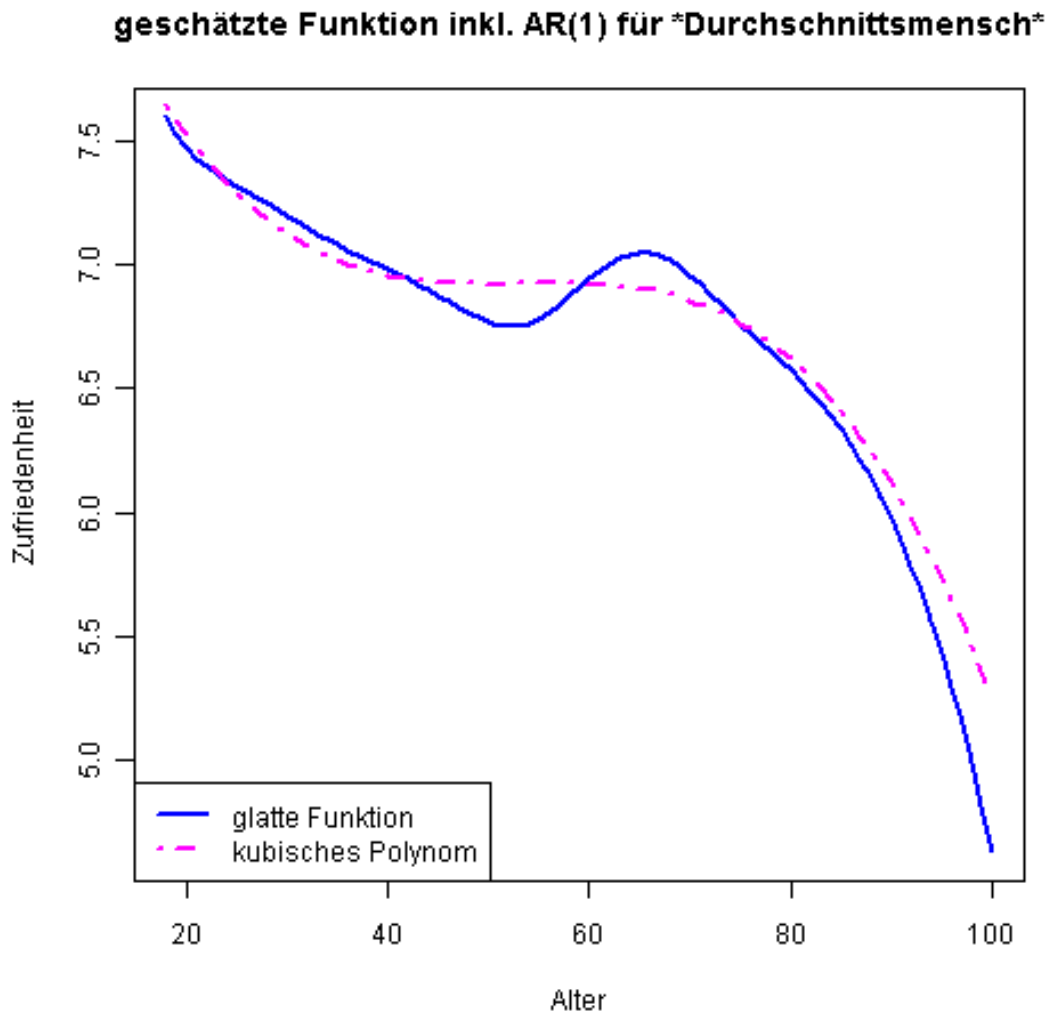
- Daten stammen aus den Haushaltsstichproben A (Westdeutsche) und C (Ostdeutsche) des Sozio-Ökonomischen Panels (SOEP)
- für die ausgewählten Modellvariablen liegen Beobachtungen aus den Jahren 1992, 1994 bis 2006 vor
- durchschnittliche Anzahl von Beobachtungen pro Person: 7.77
- in die Modellberechnungen gingen 102 708 vollständige Beobachtungen von 13 224 Individuen ein
- Anzahl Beobachtungen pro Jahr:

1992	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
8 145	7 720	7 943	7 606	8 052	7 550	7 403
2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
7 628	7 092	7 068	7 000	6 876	6 543	6 082

Ergebnis für Alterseffekt



Ergebnis für Alterseffekt



Midlife-Crisis nur bei glatter Funktion erkennbar.

Beispiel 5: Mineralwasserstudie

Studie in Zusammenarbeit mit Prof. Adam (LMU)
Fragestellung: Schmeckt mit Sauerstoff angereichertes Mineralwasser besser als gewöhnliches Mineralwasser ?

- Doppel–Blindstudie
- Kontroll–Gruppe: zweimal das gleiche Wasser ohne O_2
- Verum–Gruppe: Beim zweiten Mal mit O_2 angereichertes Mineralwasser

Ergebnis (Clausnitzer et al., 2004) :

Beispiel 5: Mineralwasserstudie

Studie in Zusammenarbeit mit Prof. Adam (LMU)
Fragestellung: Schmeckt mit Sauerstoff angereichertes Mineralwasser besser als gewöhnliches Mineralwasser ?

- Doppel–Blindstudie
- Kontroll–Gruppe: zweimal das gleiche Wasser ohne O_2
- Verum–Gruppe: Beim zweiten Mal mit O_2 angereichertes Mineralwasser

Ergebnis (Clausnitzer et al., 2004) :
Placebo: 76% gaben an, dass das zweite Wasser anders schmeckt
Verum : 89 % gaben an, dass das zweite Wasser anders schmeckt
Signifikanter Effekt $\rightarrow$ Zulassung

Beispiel 5: Mineralwasserstudie

Studie in Zusammenarbeit mit Prof. Adam (LMU)
Fragestellung: Schmeckt mit Sauerstoff angereichertes Mineralwasser besser als gewöhnliches Mineralwasser ?

- Doppel–Blindstudie
- Kontroll–Gruppe: zweimal das gleiche Wasser ohne O_2
- Verum–Gruppe: Beim zweiten Mal mit O_2 angereichertes Mineralwasser

Ergebnis (Clausnitzer et al., 2004) :
Placebo: 76% gaben an, dass das zweite Wasser anders schmeckt
Verum : 89 % gaben an, dass das zweite Wasser anders schmeckt
Signifikanter Effekt $\rightarrow$ Zulassung

Ziele und Methoden

- Randomisierte Studie (Doppelblind)
- Entscheidungsfindung durch statistischen Test
- Quantifizierung des Effekts

Was ist Statistik?

Definition Statistik

Statistik als Wissenschaft bezeichnet eine Methodenlehre, die sich mit der Erhebung, der Darstellung, der Analyse und der Bewertung von Daten auseinander setzt. Ein zentraler Aspekt ist dabei die Modellbildung mit zufälligen Komponenten.

Was ist Statistik?

Definition Statistik

Statistik als Wissenschaft bezeichnet eine Methodenlehre, die sich mit der Erhebung, der Darstellung, der Analyse und der Bewertung von Daten auseinander setzt. Ein zentraler Aspekt ist dabei die Modellbildung mit zufälligen Komponenten.

Teilgebiete:

- Deskriptive Statistik: beschreibend
- Explorative Datenanalyse: Suche nach Strukturen
- Induktive Statistik: Schlüsse von Daten auf Grundgesamtheit oder allgemeine Phänomene

Zitate

- „Traue keiner Statistik, die Du nicht selbst gefälscht hast“ (**nicht** von Churchill)

Zitate

- „Traue keiner Statistik, die Du nicht selbst gefälscht hast“
  (**nicht** von Churchill)
- „Statistics is a body of methods for making wise decisions in the face of uncertainty“
  (W.A. Wallis, A.V. Roberts)

Zitate

- „Traue keiner Statistik, die Du nicht selbst gefälscht hast“
  (**nicht** von Churchill)
- „Statistics is a body of methods for making wise decisions in the face of uncertainty“
  (W.A. Wallis, A.V. Roberts)
- „Statistisches Denken wird eines Tages für mündige Staatsbürger ebenso wichtig sein, wie die Fähigkeit zu lesen und zu schreiben“
  (H.G. Wells)

Zitate

- „Traue keiner Statistik, die Du nicht selbst gefälscht hast“
  (**nicht** von Churchill)
- „Statistics is a body of methods for making wise decisions in the face of uncertainty“
  (W.A. Wallis, A.V. Roberts)
- „Statistisches Denken wird eines Tages für mündige Staatsbürger ebenso wichtig sein, wie die Fähigkeit zu lesen und zu schreiben“
  (H.G. Wells)
- You can’t manage what you don’t measure (W.E. Deming)

Zitate

- „Traue keiner Statistik, die Du nicht selbst gefälscht hast“ (**nicht** von Churchill)
- „Statistics is a body of methods for making wise decisions in the face of uncertainty“ (W.A. Wallis, A.V. Roberts)
- „Statistisches Denken wird eines Tages für mündige Staatsbürger ebenso wichtig sein, wie die Fähigkeit zu lesen und zu schreiben“ (H.G. Wells)
- You can’t manage what you don’t measure (W.E. Deming)
- Seit der Finanzkrise gibt es in der Ökonomie einen starken Trend zur empirischen Forschung, die enorm aufgewertet wurde. Kein Wissenschaftler bekommt heute ein makroökonomisches Paper in einem guten Journal veröffentlicht, in dem er nicht saubere... empirische Forschung präsentiert (SZ 14.10.2016).

BIG DATA

- Analyse und Verarbeitung großer Datenmengen
- Drei Vs
  - Volume
  - Velocity
  - Variety

1. **Datenerhebung und Messung**
   - Die Messung
   - Skalenniveaus

Vorlesungseinheiten (vorläufig)

1. Datenerhebung und Messung
2. Häufigkeitsverteilungen
3. Lagemaße
4. Streuungsmaße
5. Konzentrationsmaße
6. Zusammenhangmaße
7. lineare Regression
8. Indizes

Grundbegriffe

- Statistische Einheit, Untersuchungseinheit
- Grundgesamtheit/ Population
- Teilgesamtheit/ Stichprobe
- Merkmal
- Merkmalsausprägung

Untersuchungseinheit und Grundgesamtheit

Definition Untersuchungseinheit

Die Objekte, auf die sich eine statistische Analyse bezieht, heißen Untersuchungseinheiten. Diese werden im folgenden durch das Symbol ω dargestellt.

Definition Grundgesamtheit

Die Zusammenfassung aller *Untersuchungseinheiten* bildet die Grundgesamtheit. Sie wird durch das Symbol Ω dargestellt. Die Beziehung zwischen Untersuchungseinheiten und zugehöriger Grundgesamtheit lässt sich kurz wie folgt umschreiben: $\omega_i \in \Omega$.

Untersuchungseinheit und Grundgesamtheit

1. Beispiel: Personalstruktur einer Firma

Wenn wir uns für die Personalstruktur einer Firma interessieren, so besteht die Grundgesamtheit Ω aus der gesamten Belegschaft; jeder einzelne Mitarbeiter stellt eine Untersuchungseinheit dar.

2. Beispiel: Wirtschaftskraft der chemischen Industrie in Europa

Hier sind die europäischen Chemiefirmen die Untersuchungseinheiten. Zusammengefasst ergibt sich aus ihnen die Grundgesamtheit der europäischen Chemieindustrie.

Merkmal, Merkmalsausprägung und Merkmalsraum

Definition Merkmal bzw. statistische Variable

Bestimmte *Aspekte* oder *Eigenschaften* einer Untersuchungseinheit bezeichnet man als Merkmal oder statistische Variable.

Definition Merkmalsausprägung

Eine Merkmalsausprägung ist der *konkrete Wert* eines Merkmals, die eine bestimmte Untersuchungseinheit aufweist. Es gibt zwei Typen von Ausprägungen:

Qualitativ Sie lassen sich durch die *verschiedenartigen* Ausprägungen charakterisieren.

Quantitativ Diese sind *messbar* und werden durch Zahlen erfasst. Bei diesen gibt es eine weitere Unterscheidung:

diskret Der zugehörige Zustandsraum ist *abzählbar* groß.
stetig Es sind *überabzählbar* viele Elemente im Zustandsraum.

Merkmal, Merkmalsausprägung und Merkmalsraum

Definition Merkmalsraum oder Zustandsraum

Die Menge *aller möglichen* Merkmalsausprägungen bildet den Merkmalsraum oder Zustandsraum.

Merkmal, Merkmalsausprägung und Merkmalsraum

Beispiele Merkmal bzw. statistische Variable

1. Farbe eines Produkts
2. bestellte Produkte pro Auftrag
3. Gewinn/Verlust pro Monat

Beispiele Merkmalsausprägung

1. 6. Gut = blau
2. 19. Kunde = 17 Stück
3. März 2010 = +23.500 €

Weitere Klassen

- Quasi-stetiges Merkmal
  diskret, sehr kleine Einheiten, „praktisch“ stetig.
  Beispiel: Monetäre Größen in Cent, usw.
- Gruppierte Daten, Häufigkeitsdaten: stetiges oder quasi-stetiges Merkmal X
  Wertebereich wird in Gruppen (Klassen, Kategorien) eingeteilt.
  Beispiele: Gehalt in Gehaltsklassen, Alter in Altersklassen
  Bemerkung: Gruppierung dient auch dem Datenschutz!

Datengewinnung und Erhebungsarten

- Vollerhebung:
  Alle statistischen Einheiten der Grundgesamtheit werden untersucht („erhoben“).
- Stichprobe = Teilerhebung
- Zufallsstichprobe:
  statistische Einheiten der Stichprobe werden zufällig nach einem bestimmten Mechanismus gezogen
  Mehr dazu in Statistik II (induktive Statistik) und in der Vorlesung Stichprobenverfahren
- Bewusste Auswahlverfahren „Expertenauswahl“
- Quotenauswahl

Induktive Statistik in der Regel nur mit zufälliger Stichprobe möglich!

Erhebungsarten

- Querschnittsdaten:
  Ein oder mehrere verschiedene Merkmale werden an einer Reihe von Objekten einmal erhoben (zu einem bestimmten Zeitpunkt oder in einem bestimmten Zeitraum)

Erhebungsarten

- Querschnittsdaten:
  Ein oder mehrere verschiedene Merkmale werden an einer Reihe von Objekten einmal erhoben (zu einem bestimmten Zeitpunkt oder in einem bestimmten Zeitraum)
- Zeitreihe
  Beispiele: Aktienkurse, Wirtschaftsentwicklung

Erhebungsarten

- Querschnittsdaten:
  Ein oder mehrere verschiedene Merkmale werden an einer Reihe von Objekten einmal erhoben (zu einem bestimmten Zeitpunkt oder in einem bestimmten Zeitraum)
- Zeitreihe
  Beispiele: Aktienkurse, Wirtschaftsentwicklung
- Longitudinal-, Längsschnitt- oder Paneldaten:
  Ein oder mehrere Merkmale werden mehrmals zu verschiedenen Zeitpunkten an einer Reihe von Objekten erhoben.

  Beispiel: Sozioökonomisches Panel

Experimente

Es werden in der Regel verschiedene „Behandlungen “verglichen
Experimentator greift ein

- Randomisierte Studie: Zuordnung von Einheiten zu Behandlungen erfolgt durch Losverfahren (Randomisierung)
- Randomisierte Experimente in der BWL (Spiele, WWW)

Messen

„Measurement is the contact of reason with nature“ Henry Margenau (1959)

Messen

„Measurement is the contact of reason with nature“ Henry Margenau (1959)

„In its broadest sense, measurement is the assignment of numerals to objects or events according the rules“

Messen

„Measurement is the contact of reason with nature“ Henry Margenau (1959)

„In its broadest sense, measurement is the assignment of numerals to objects or events according the rules“

Messen bedeutet die Zuordnung von Zahlen zu Ausprägungen von Merkmalen an Objekten.

- Physikalische Messung
  Beispiele: Gewicht, Blutdruck, Fettaufnahme
- Psychologie
  Beispiele: Intelligenz, Gewaltbereitschaft, Performance
- Wirtschaftswissenschaften
  Beispiele: Inflation, Bruttosozialprodukt, Umsatz,

Definition

Peter	$\longrightarrow$	1.84
Stefan	$\longrightarrow$	1.91
Laura	$\longrightarrow$	1.72

Merkmal definiert Relation (Struktur) zwischen den Objekten.

Messung: strukturerhaltende Abbildung (Homomorphismus)

Peter ist kleiner als Stefan $\Leftrightarrow 1.84 < 1.91$

Nominalskala

- Beispiele:
  Diagnosen, Geschlecht
- Struktur:
  keine
- Mögliche Aussagen:
  gleich, ungleich

Ordinal- oder Rangskala

- Beispiele:
  Schulbildung, soziale Schicht, Schulnoten
- Struktur:
  lineare Ordnung
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner

Intervallskala

- Beispiele:
  Umsatz, Preisindex, Schulnoten (??)
- Struktur:
  Abstände sinnvoll definiert
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner, Differenzen

Verhältnisskala

Intervallskala mit Nullpunkt

- Beispiele:
  Gewinn, Preis, Beschäftigungsdauer, Alter
- Struktur:
  Abstände definiert, Nullpunkt
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner, Differenzen, Verhältnis

Absolutskala

- Beispiel:
  Häufigkeit
- Struktur:
  Einheit liegt auf natürliche Weise fest

Skalenniveau

Beachte:

- Je höher das Skalenniveau, desto mehr Interpretationen und Rechnungen sind möglich
- Je höher das Skalenniveau, desto mehr (implizite) Annahmen werden gemacht

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
intervall	ja	ja	ja	nein
verhältnis	ja	ja	ja	ja

Indexbildung

Bildung von Einzelindikatoren zu einer neuen Variablen
Häufig: Bildung von (gewichteten) Summen von einzelnen Variablen

Beispiel:

$$\text{Pflege–Qualität} = a_1 \cdot Q(\text{Essen}) + a_2 \cdot Q(\text{Medizinische Versorgung}) + \ldots$$

Indexbildung folgt nur theoretischen Vorgaben und fachspezifischen Überlegungen

Fragen der Statistik:

- Gleichheit sinnvoll ? (Dimensionsreduktion zulässig)
- Ordnung bzw. Abstände sinnvoll ?

2 Datenorganisation und Häufigkeitsverteilungen

- Datenorganisation
- Statistik-Software
- Häufigkeiten

Deskriptive Statistik

„Data is merely the raw material of knowledge.“

Ziel: Beschreibung von Daten mit möglichst geringem Informationsverlust

- Eigenschaften und Strukturen sichtbar machen
- Graphisch und durch Kennwerte
- Eindimensional und mehrdimensional
- Zunächst keine Schlüsse auf die Grundgesamtheit oder allgemeine Phänomene

Rohdaten und Datenmatrix

Die Daten liegen in der Regel als Datenmatrix vor:

- Zeilen entsprechen Untersuchungseinheiten
- Spalten entsprechen Merkmalen
- Elemente der Matrix sind die Merkmalsausprägungen
- Fragen mit Mehrfachnennungen als einzelne binäre Merkmale definieren

Hinweise zur Eingabe unter `http://www.stablab.stat.uni-muenchen.de/Datensaetze_mit_Excel`

Beispiel: Mietspiegel

Nr	nm	nmqm	wfl	rooms	bj	bez	kueche
1	608.40	12.67	48	2	1957	Untergiesing	0
2	780.00	13.00	60	2	1983	Bogenhausen	0
3	822.60	7.48	110	5	1957	Obergiesing	1
4	500.00	8.62	58	2	1957	Schwanthh	0
5	595.00	8.50	70	3	1972	Aubing	0
6	960.00	11.85	81	3	2006	Schwanthh	0

Kodierung

Motivation

Da es nicht möglich ist, mit Zeichenketten zu rechnen, müssen qualitative Merkmale für die statistische Analyse mit einer Statistik-Software geeignet aufbereitet werden.

Definition Kodierung

Der Vorgang, bei dem man Merkmalsausprägungen oder fehlenden Werten Zahlen zuordnet, die die entsprechende Ausprägung repräsentieren, bezeichnet man als Kodierung.

Dateneingabe

Nützliche Werkzeuge

Um die erhobenen Daten von Beobachtungen, Umfragen oder Experimenten auszuwerten, diese gebündelt in einer Datei abzuspeichern. Dafür eignen sich

Tabellenkalkulationsprogramme Excel, Lotus 1-2-3

Datenbaksysteme dBase, Paradox, Access, MySQL

Statistikpakete R, SAS,SPSS,STATA

SPSS

Vorteile

- umfassendes Paket an statistischen Methoden vorhanden
- erarbeitete Skripte können auf neue Daten leicht angepasst werden
- schnelle Einarbeitung in die Benutzeroberfläche
- Konsistenzüberprüfung der Daten vorhanden

Nachteile

- Skripterstellung bedarf Einarbeitung
- Formatierung von Diagrammen in der Benutzeroberfläche zum Teil umständlich
- teure Lizenz

Programmiersprache R

Vorteile

- umfassendes Paket an statistischen Methoden vorhanden
- erarbeitete Skripte können auf neue Daten leicht angepasst werden
- bequeme Schnittstellen zu Dateien/Datenbanken vorhanden
- kostenlos beziehbar unter `http://www.r-project.org`
- Editor R-Studio unter `https://www.rstudio.com/`
- Kurse für Studierende der LMU

Nachteile

- bedarf Einarbeitung
- keine eigene Datenverwaltung vorhanden
- Updates nicht notwendiger Weise kompatibel zu älteren Versionen

Eindimensionale Häufigkeitsverteilung

- Ordnen der Daten nach einem Merkmal
- Auszählen der Häufigkeiten der einzelnen Merkmalsausprägungen
- Relative Häufigkeiten = Häufigkeit/Anzahl der Untersuchungseinheiten

Häufigkeitsverteilung

Im Weiteren:

$X, Y, \ldots$ Bezeichnung für Merkmal

n Untersuchungseinheiten

$x_1, \ldots, x_i, \ldots, x_n, \quad i = 1, \ldots, n$ beobachtete Werte bzw. Merkmalsausprägungen von X

$\{ x_1, \ldots, x_i, \ldots, x_n; \quad i = 1, \ldots, n\}$ Rohdaten

Häufigkeiten I

$a_1 < a_2 < \ldots < a_k$, $k \leq n$ der Größe nach geordnete, *verschiedene* Werte $x_1, \ldots, x_n$

Beispiel: Absolventenstudie
Für die Variable D "Ausrichtung der Diplomarbeit" sind die Daten durch die folgende Tabelle gegeben.

Person i	1	2	3	4	5	6	7	8	9	10	11	12
Variable D	3	4	4	3	4	1	3	4	3	4	4	3

Person i	13	14	15	16	17	18	19	20	21	22	23	24
Variable D	2	3	4	3	4	4	2	3	4	3	4	2

Person i	25	26	27	28	29	30	31	32	33	34	35	36
Variable D	4	4	3	4	3	3	4	2	1	4	4	4

Häufigkeiten II

Ausprägung	absolute Häufigkeit h	relative Häufigkeit f
1	2	$2/36 = 0.056$
2	4	$4/36 = 0.111$
3	12	$12/36 = 0.333$
4	18	$18/36 = 0.500$

Häufigkeitstabelle für die Variable D „Ausrichtung der Diplomarbeit“

Bemerkungen:

- Für Nominalskalen hat die Anordnung „$<$“ keine inhaltliche Bedeutung.
- Bei kategorialen Merkmalen $\Rightarrow$ k = Anzahl der Kategorien
  Bei stetigen Merkmalen $\Rightarrow$ k oft nicht oder kaum kleiner als n.

Absolute und relative Häufigkeiten

$h(a_j) = h_j$ *absolute Häufigkeit* der Ausprägung a_j,

d.h. Anzahl der x_i aus $x_1, \ldots x_n$ mit $x_i = a_j$

$f(a_j) = f_j = h_j/n$ *relative Häufigkeit* von a_j

$h_1, \ldots, h_k$ *absolute Häufigkeitsverteilung*

$f_1, \ldots, f_k$ *relative Häufigkeitsverteilung*

Klassenbildung

Vorgehensweise bei vielen Ausprägungen

Bei stetigen Merkmalen und diskreten Merkmalen mit vielen Ausprägungen (= quasistetig) bedarf es des Zwischenschritts der *Klassenbildung*, um eine überschaubare Verteilung zu erhalten.

Klassenbildung

Als Anhaltspunkt für eine brauchbare Verteilung sollten $\sqrt{n}$ Klassen gebildet werden. Bei der Wahl der Klassen gibt es zwei Möglichkeiten:

1. nach *sachlogischen* Gegebenheiten
2. nach *willkürlichen* Kriterien,

wobei die zweite Gestaltungsart Raum zur Manipulation der Häufigkeitsstruktur schafft und deshalb vermieden werden sollte.

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	
$[900; 1500[$	9,3688 Mio	
$[1500; 2600[$	12,2696 Mio	
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	$\frac{5,7232}{39,2}$
$[900; 1500[$	9,3688 Mio	
$[1500; 2600[$	12,2696 Mio	
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	
$[1500; 2600[$	12,2696 Mio	
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	$\frac{9{,}3688}{39{,}2}$
$[1500; 2600[$	12,2696 Mio	
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	$\frac{12,2696}{39,2}$
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	$\frac{7,4088}{39,2}$
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	0,189
$[4500; \infty[$	4,4296 Mio	
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	0,189
$[4500; \infty[$	4,4296 Mio	$\frac{4,4296}{39,2}$
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	0,189
$[4500; \infty[$	4,4296 Mio	0,113
Σ	39,2 Mio	

Beispiel

Verteilung der monatlichen Haushaltsnettoeinkommen in Deutschland 2005

mon. Einkommen in €	absolute Häufigkeit	relative Häufigkeit
$[0; 900[$	5,7232 Mio	0,146
$[900; 1500[$	9,3688 Mio	0,239
$[1500; 2600[$	12,2696 Mio	0,313
$[2600; 4500[$	7,4088 Mio	0,189
$[4500; \infty[$	4,4296 Mio	0,113
Σ	39,2 Mio	1

Häfigkeitstabelle

Definition Häufigkeitstabelle

Die *tabellarische Zusammenfassung* der

- *Merkmalsausprägungen* a_j
  bei qualitativen und diskreten Merkmalen
- *Klassengrenzen*
  nur bei (quasi-)stetigen Merkmalen
- *Klassenbreiten*
  nur bei stetigen Merkmalen
- *absoluten Häufigkeiten* h_j
- *relativen Häufigkeiten* f_j

für alle Merkmalsausprägungen $j = 1, ..., k$ wird als Häufigkeits-tabelle bezeichnet.

Häfigkeitstabelle

allgemeine Form bei qualitativen und diskreten Merkmalen

j	a_j	h_j	f_j
1	a_1	h_1	f_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	a_k	h_k	f_k
Σ		n	1

Beispiel

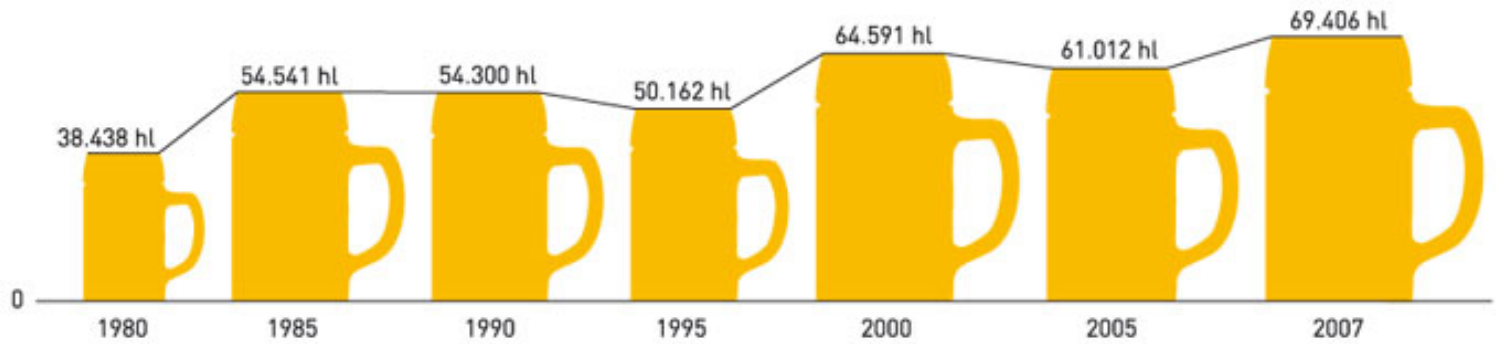
Weltbevölkerung nach Kontinenten im Jahr 2002

lfd Nr. (j)	Kontinent (a_j)	abs. H'keit (h_j)	rel. H'keit (f_j)
1	Asien	3.769 Mio	0,607
2	Afrika	832 Mio	0,134
3	Europa	725 Mio	0,116
4	Lateinamerika	534 Mio	0,086
5	Nordamerika	320 Mio	0,052
6	Ozeanien	31 Mio	0,005
Σ		6.211 Mio	1,000

Quelle: UNFPA Weltbevölkerungsbericht 2002 New York / Stuttgart

Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“



Statistische Grafik

Lit.: Tufte, E. (2001): The visual Display of Information.
Graphic Press 2nd ed.

Allgemeine Kriterien

- Wahl der Skala inkl. Bereich
- Wahl des Prinzips (Längentreue, Flächentreue)
- Einbringen von anderen Visualisierungen (Piktogramme etc.)
- Angemessene Wahl der Variablen
- Angemessene Wahl der Farben

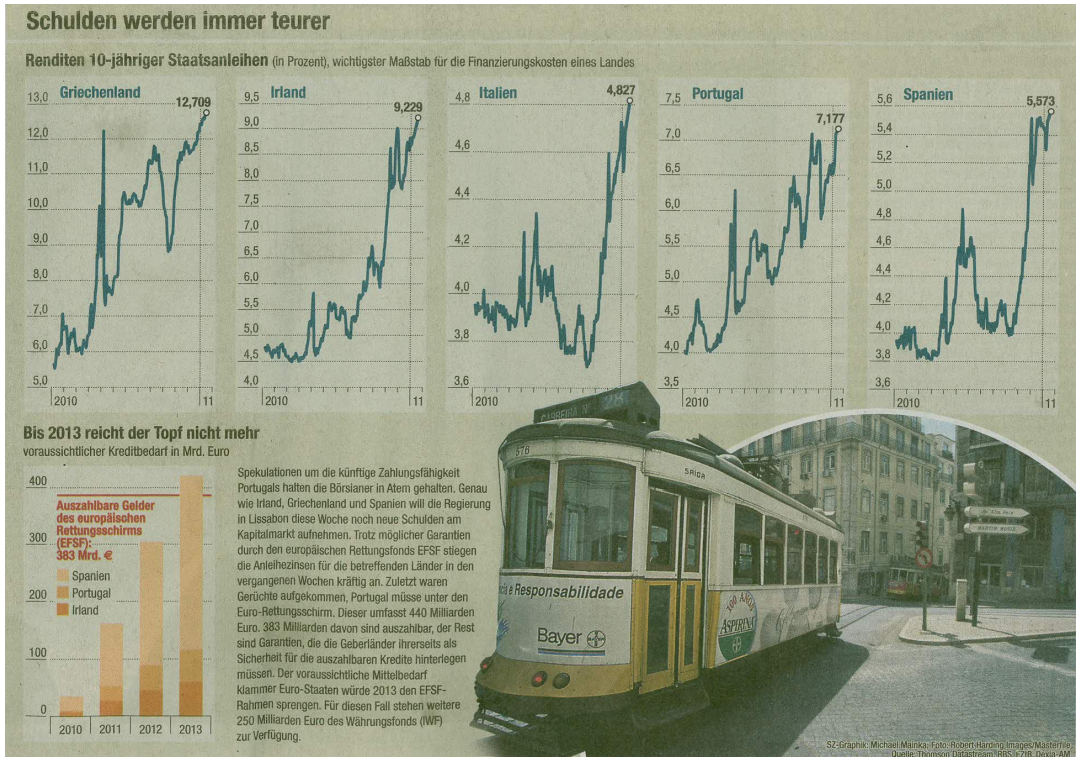
Wahrnehmung von Grafiken

Experimente von Psychologen zeigen Hierarchie der korrekten Interpretation (Cleveland/McGill)

1. Abstände
2. Winkel
3. Flächen
4. Volumen
5. Farbton-Sattheit-Schwärzegrad

Da Abstände am besten wahrgenommen werden, sollten diese bevorzugt verwendet werden.

Grafische Darstellungen



SZ 11.1.11

Entwicklung der weltweiten Kunststoffproduktion

Grafik von Andreas Quatember (Universität Linz)
`http://www.jku.at/ifas/content/e101235/e101334`

Kommentar: Eine waschechte Zeitungsente!

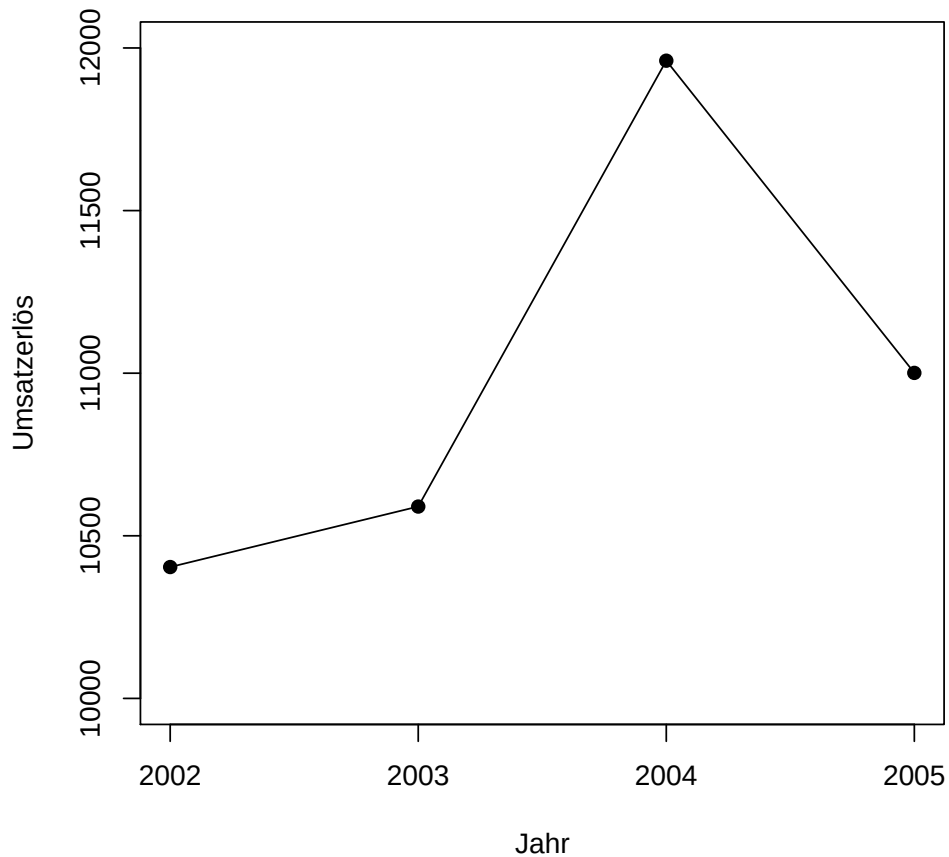
Die mengenmäßige Entwicklung der weltweiten Kunststoffproduktion über drei Zeitpunkte (1976, 2002, 2012) wird hier durch immer größer werdenden „Quietschentchen“ dargestellt.

- Unterschiedlichen Zeiträume zwischen den Jahreszahlen mal hin.
- Eindruck einer Versechsfachung ($288 : 47 = 6{,}1$) falsch !
- Volumina werden miteinander verglichen

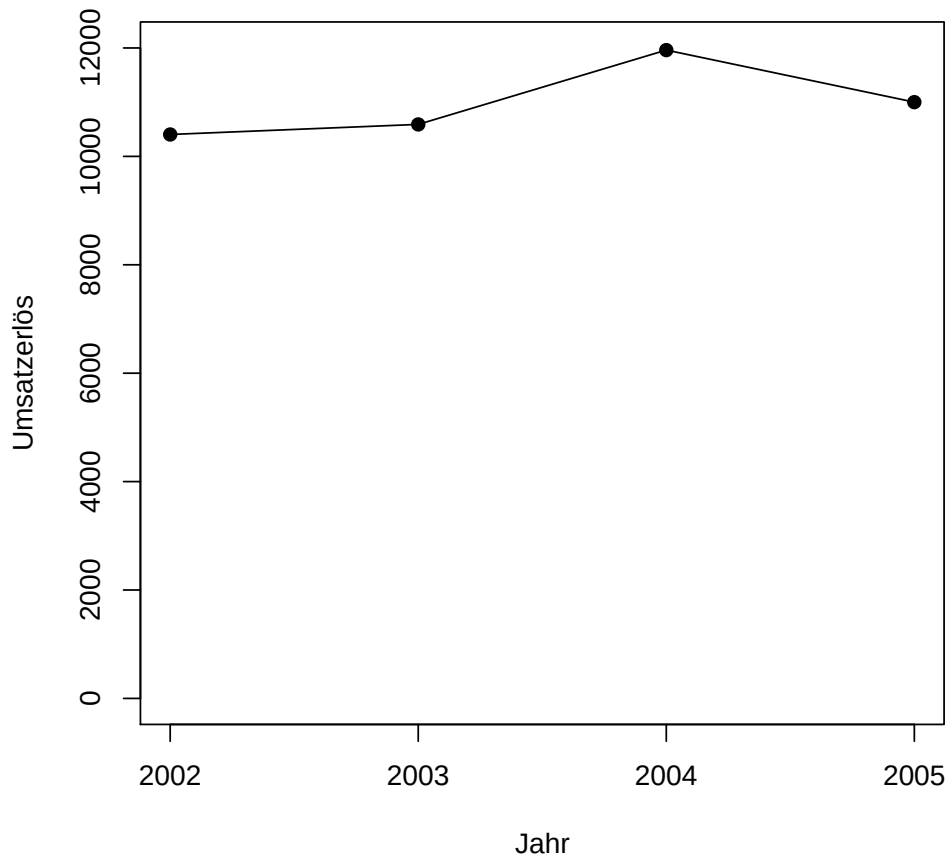
3 D -Darstellungen sind zu vermeiden

Räumliche Darstellungen von eindimensionalen Häufigkeiten führen meist zu verzerrter Wahrnehmung.

Beispiel: Liniendiagramm (??)



Beispiel: Liniendiagramm (!!)



Typen von eindimensionalen Darstellungen

- Stab-, Balken- und Säulendiagramm
- Kreis (Torten)-Diagramm
- Histogramm

Stabdiagramm, Säulen- und Balkendiagramm

- *Stabdiagramm*:
  Trage über $a_1, \ldots, a_k$ jeweils einen zur x-Achse senkrechten Strich (Stab) mit Höhe $h_1, \ldots, h_k$ (oder $f_1, \ldots, f_k$) ab.
- *Säulendiagramm*:
  wie Stabdiagramm, aber mit Rechtecken statt Strichen.
- *Balkendiagramm*:
  wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x-Achse.

Säulendiagramm

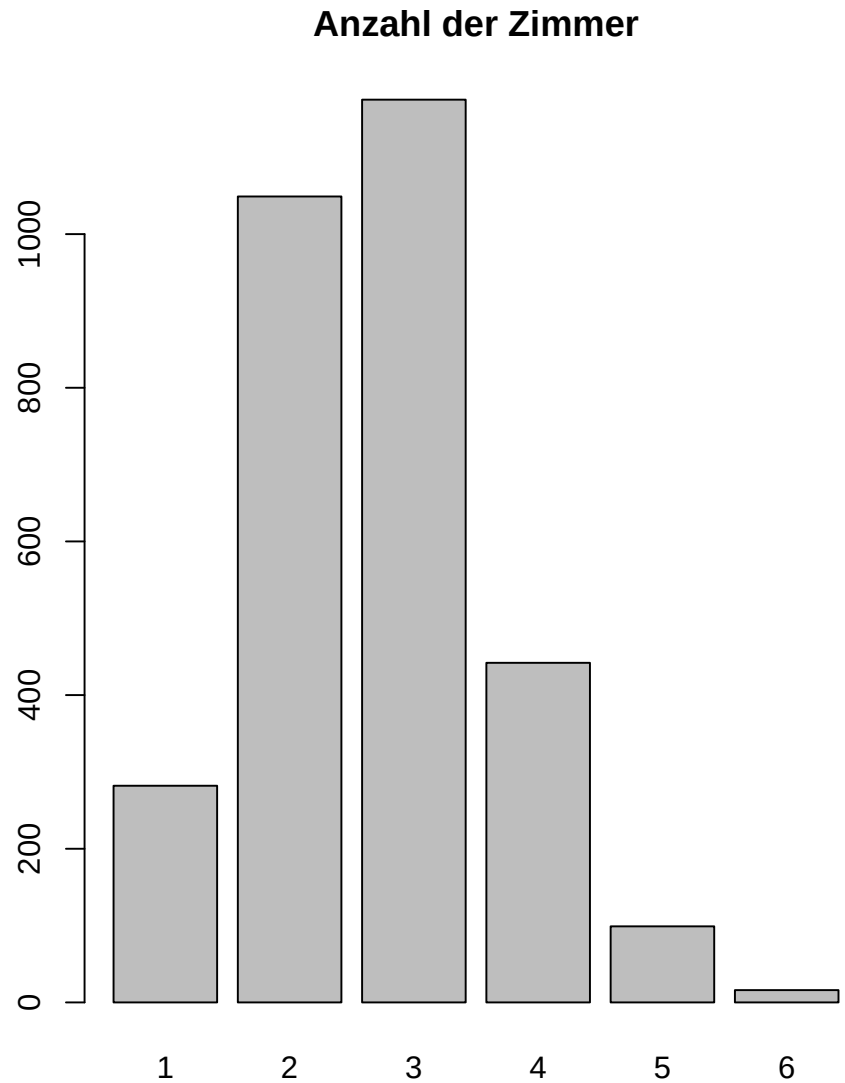
Darstellung der absoluten oder relativen Häufigkeiten als Höhen (Längen)

x-Achse: Ausprägungen des Merkmals
y-Achse: absolute/ relative Häufigkeiten

Anwendungen:

- Ordinale Merkmale
- Metrische Merkmale mit wenigen Ausprägungen
- Nominale Merkmale (Problem: Ordnung nicht vorhanden)

Beispiel : Zahl der Zimmer im Mietspiegel



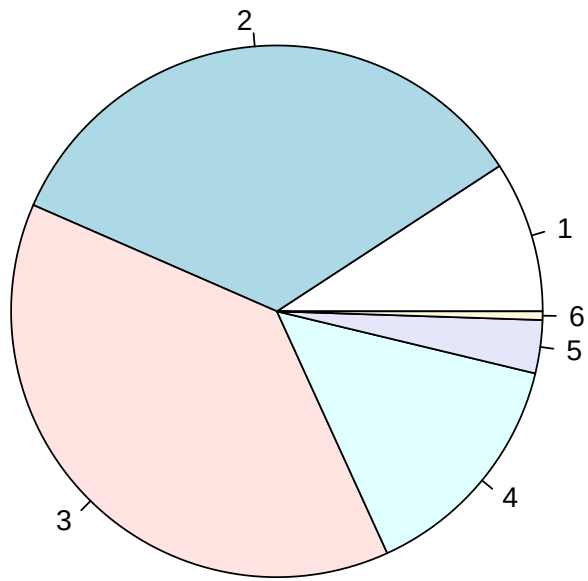
Kreisdiagramm, Tortendiagramm

Darstellung der relativen (absoluten) Häufigkeiten als Fläche eines Kreises

Anwendung:

- Nominale Merkmale
- Ordinale Merkmale (Problem: Ordnung nicht korrekt wiedergegeben)
- Gruppierte Daten

Mietspiegel: Zahl der Zimmer



Stapeldiagramm

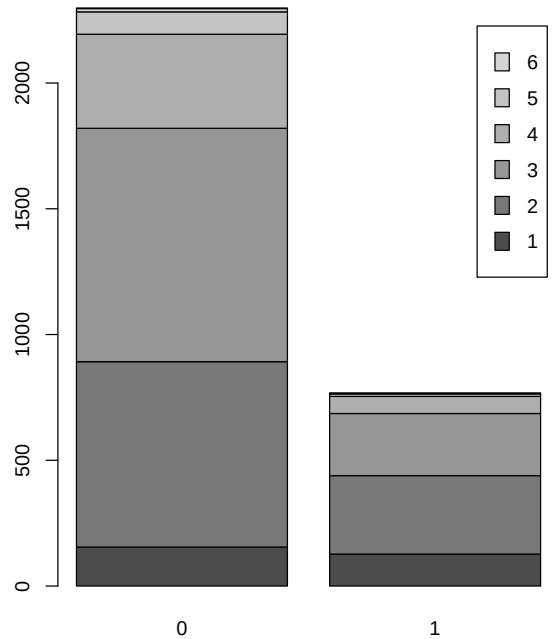
Darstellen der absoluten oder relativen Häufigkeiten als Länge. Die Abschnitte werden übereinander in verschiedenen Farben gestapelt.

Anwendungen:

- Ordinale Daten
- Gruppierte Daten
- Metrische Daten mit wenigen Ausprägungen

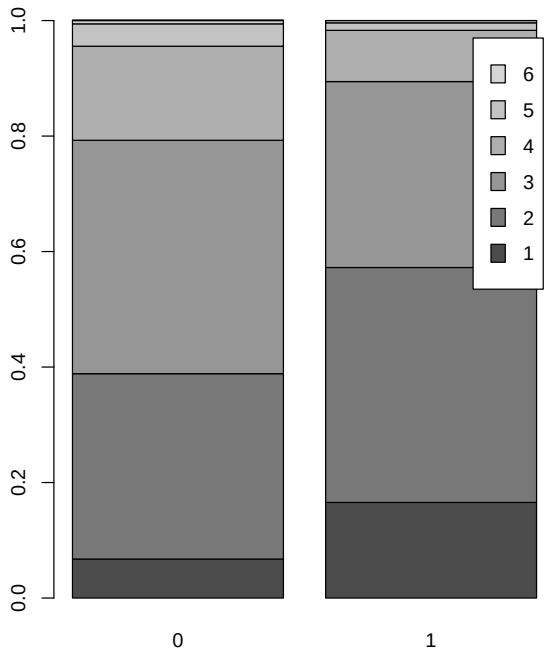
Besonders geeignet für den Vergleich verschiedener Gruppen durch nebeneinander liegende Stapel. Zu beachten ist dann die Unterscheidung: relative Häufigkeit $\leftrightarrow$ absolute Häufigkeit

Beispiel: Zahl der Zimmer/Küchenausstattung



2 Gruppen: 1 = gehobene Ausstattung der Küchen, 0 keine gehobene Ausstattung der Küche

Beispiel: Zahl der Zimmer/Küchenausstattung



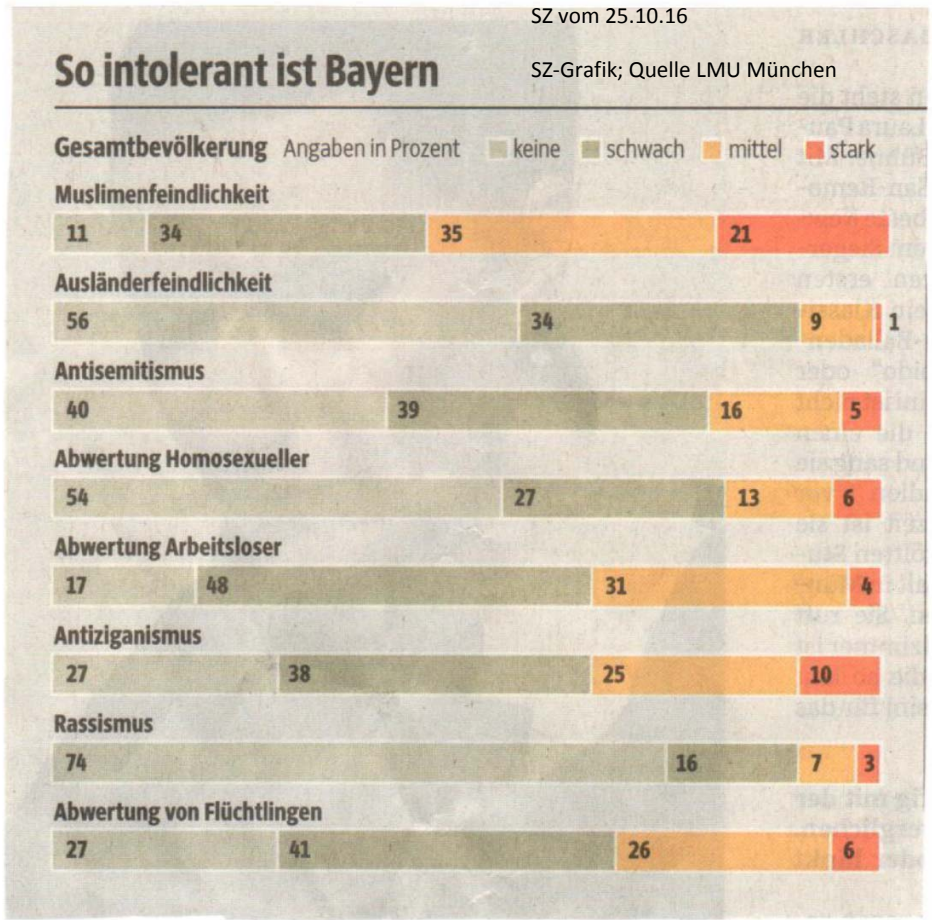
2 Gruppen: 1 = gehobene Ausstattung der Küchen, 0 keine gehobene Ausstattung der Küche

Beispiel: Intoleranz in Bayern SZ 25.10

Studie aus der LMU Soziologie Christian Ganser

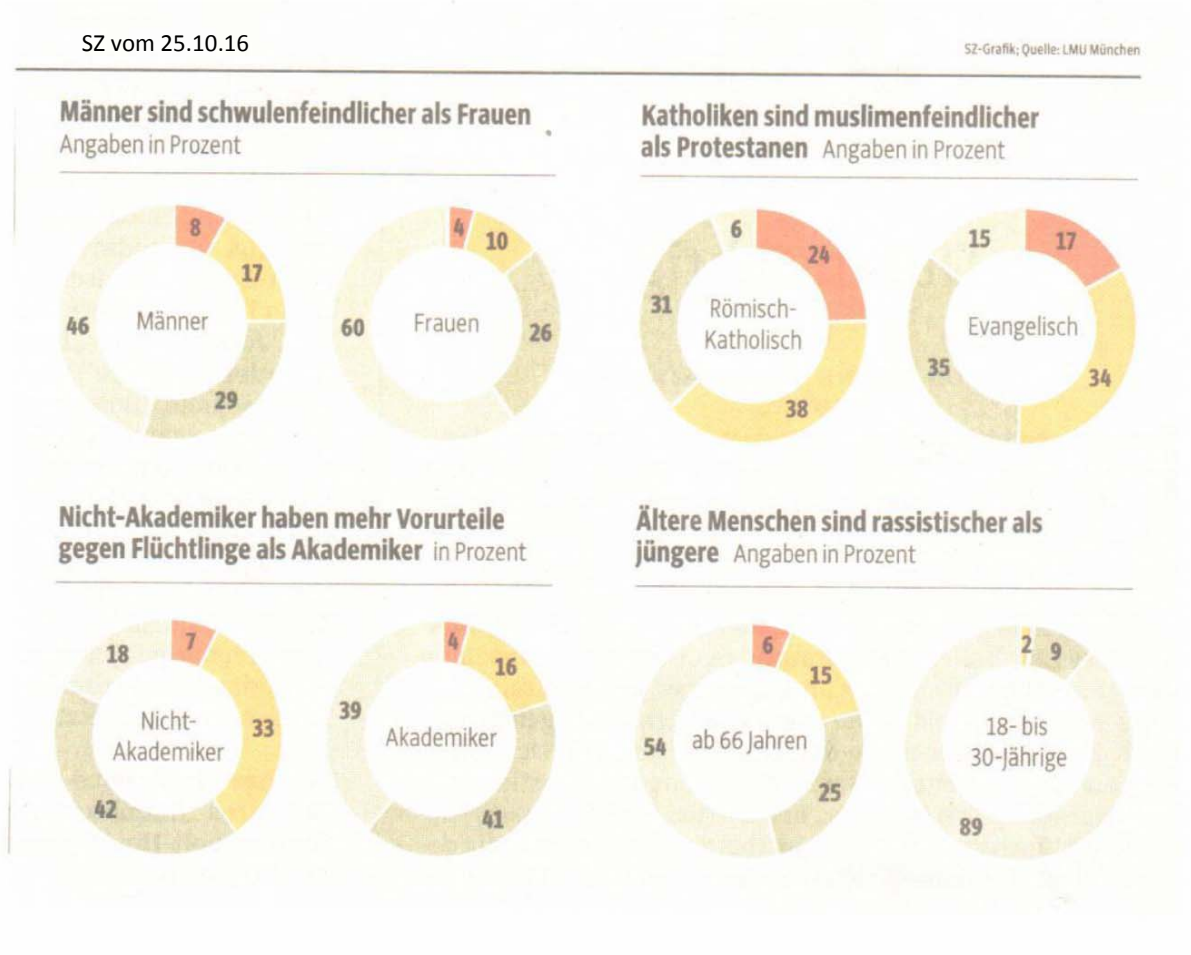
SZ vom 25.10.16

SZ-Grafik; Quelle LMU München



Beispiel: Intoleranz in Bayern SZ 25.10

SZ vom 25.10.16



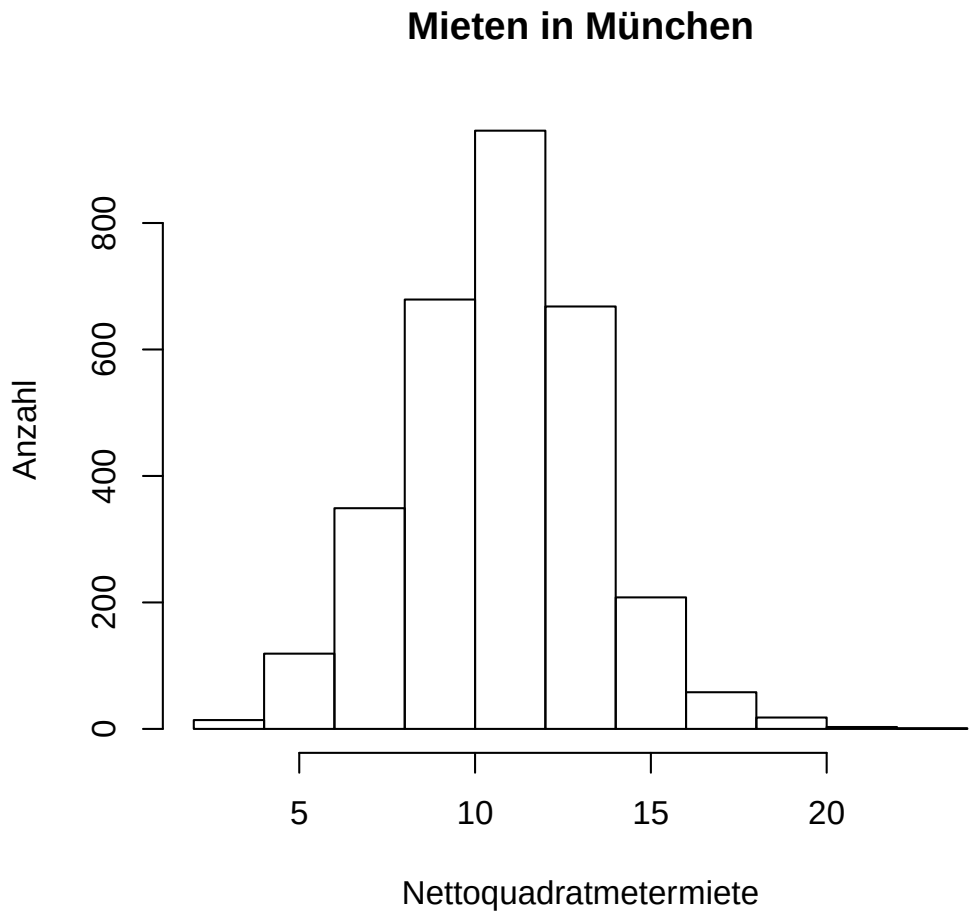
Das Histogramm

Darstellung der relativen Häufigkeiten durch Flächen
(Prinzip der Flächentreue)

Vorgehen:

1. Aufteilung in Klassen (falls die Daten noch nicht gruppiert sind)
2. Bestimmung der relativen Häufigkeiten $f_j = \frac{h_j}{n}$
3. Bestimmung der Höhen l_j, so dass gilt $b_j \cdot l_j = f_j$
   wobei b_j: Breite der Klasse j.
4. Bei gleichen Klassenbreiten $b_j = b$ gilt $l_j = \frac{f_j}{b} = \frac{h_j}{b \cdot n}$. Daher wird bei solchen Histogrammen als Höhe auch h_j gewählt
5. Klassierung sollte möglichst mit gleichen Klassenbreiten erfolgen

Mietspiegel



Histogramm

- Anwendung bei metrischen Daten
- Beachte: Abhängigkeit von der Breite
- Klasse inhaltlich vorgeben, verschiedene Varianten ansehen.
- Vorsicht bei Rändern

Empirische Verteilungsfunktion

$$H(x) := \text{Anzahl der Werte} <= x$$

$$F(x) = H(x)/n = \text{Anteil der Werte } x_i \text{ mit } x_i \leq x$$

bzw.

$$F(x) = f(a_1) + \ldots + f(a_j) = \sum_{i:a_i \leq x} f_i,$$

wobei $a_j \leq x$ und $a_{j+1} > x$ ist.

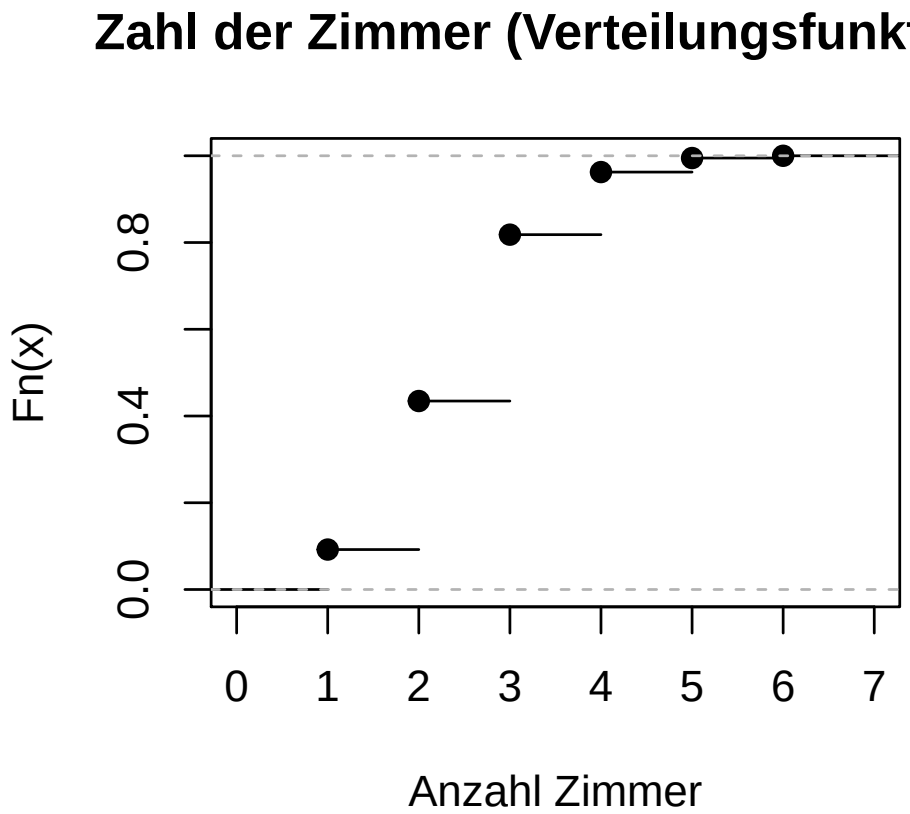
empirische Verteilungsfunktion ordinaler und diskreter Merkmale

Beispiel

Zugrunde liegende Daten entstammen dem Münchner Mietspiegel von 2005.

Anz. Räume	abs. H'keit	rel. H'keit	$F(x)$
1	282	0,092	0,092
2	1049	0,342	0,434
3	1175	0,384	0,818
4	442	0,144	0,962
5	99	0,033	0,995
6	16	0,005	1
Σ	3063	1	

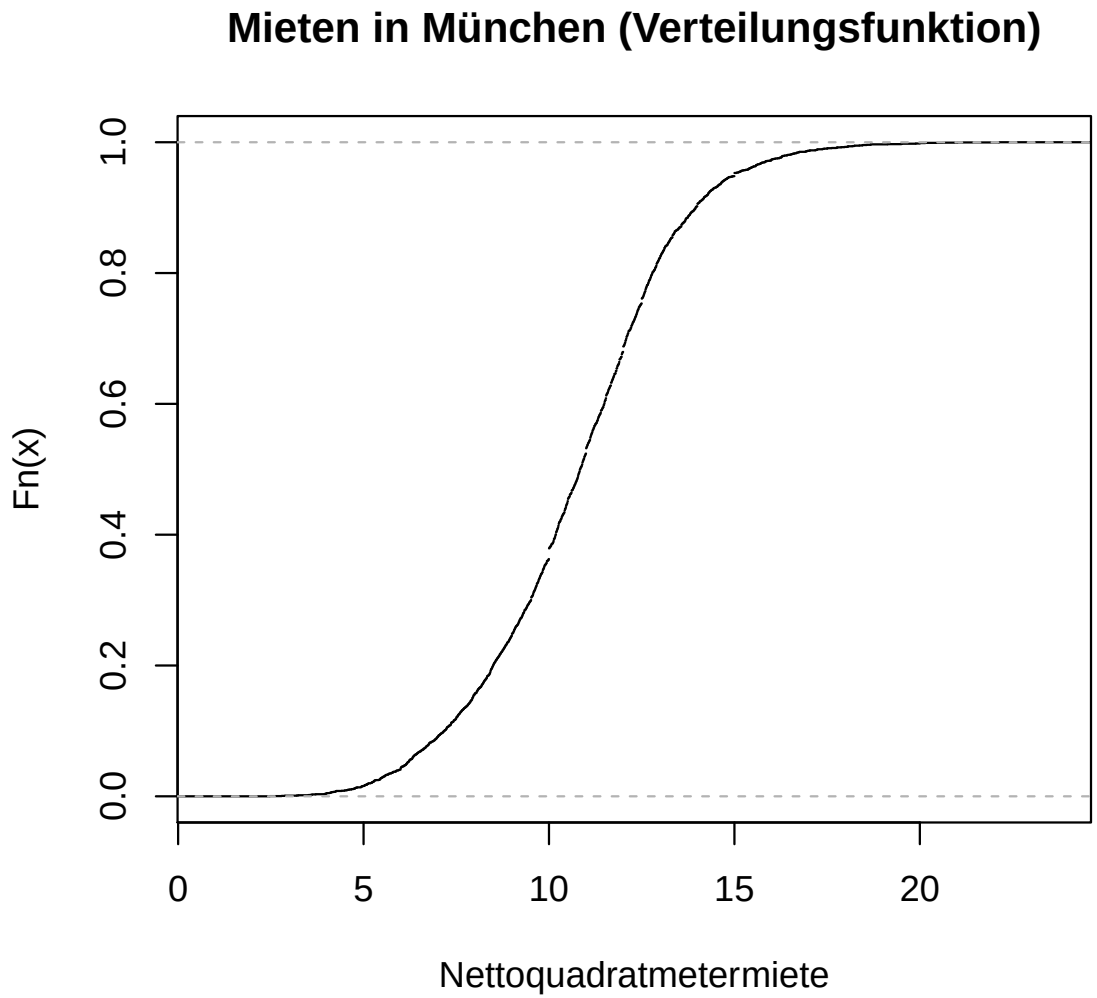
Beispiel für eine Empirische Verteilungsfunktion



Eigenschaften von $F(x)$

- monoton wachsende Treppenfunktionen mit Sprüngen an den Ausprägungen $a_1, \ldots, a_k$
- Sprunghöhen: $f_1, \ldots, f_k$
- rechtsseitig stetig
  $F(x) = 0$ für $x < a_1$, $F(x) = 1$ für $x \geq a_k$

Beispiel für eine Empirische Verteilungsfunktion



3 Lagemaße

- Modus
- Median
- Quantile
- Darstellung: Boxplot
- Mittelwert

Fragen

- Wo liegt die Masse der Daten?
- Wo liegt die Mehrzahl der Daten?
- Wo liegt die Mitte der Daten?
- Welche Merkmalsausprägung ist typisch für die Häufigkeitsverteilung?

Modus

verwendbar bei Merkmalen mit

Nominalskala	Ordinalskala	metrische Skala
x	x	x

Definition Modus

Als Modus oder Modalwert $\bar{x}_M$ bezeichnet man den häufigsten Wert einer Verteilung.

- Bei *diskreten Daten* ist der Modus die Merkmalsausprägung a_j, die am häufigsten auftritt. Bei mehreren Maxima ist der Modus nicht eindeutig definiert.
- Für *gruppierte Daten* ist der Modus definiert als die Klassenmitte der am dichtesten besetzten Gruppe.

Der Modus: Eigenschaften

Eigenschaften:

- oft nicht eindeutig
- nur bei gruppierten Daten oder bei Merkmalen mit wenigen Ausprägungen sinnvoll
- stabil bei allen eindeutigen Transformationen

Beispiel

Gegeben sei folgende Werteliste: 4, 5, 5, 6, 6, 6, 7, 7, 9.
Die zugehörige Häufigkeitstabelle ergibt

Ausprägung	H'keit
4	1
5	2
6	3
7	2
9	1

und somit $\bar{x}_M = 6$.

Bei einer Datentransformation, hier z.B. $Y = X^2$, ergibt sich:

$$\bar{y}_M = 36 = (\bar{x}_M)^2$$

Median

verwendbar bei Merkmalen mit

Nominalskala	Ordinalskala	metrische Skala
	x	x

Definition Median oder Zentralwert

Der Median oder Zentralwert wird aus den geordneten Daten gewonnen. Er wird durch die Forderung bestimmt, dass

- 50% der beobachteten Werte *kleiner oder gleich* und
- 50% der beobachteten Werte *größer oder gleich*

dem Median sein sollen. Er wird mit $\tilde{x}_{0,5}$ bezeichnet.

Eine alternative Formulierung für die Bestimmung des Medians ist über die empirische Verteilungsfunktion durch die Forderung $F(\tilde{x}_{0,5}) = 0,5$ gegeben.

Median

Berechnung des Medians

$$\tilde{x}_{0,5} = \begin{cases} x_{(n+1)/2} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{falls } n \text{ gerade} \end{cases}$$

Eigenschaften des Medians

- anschaulich
- geeignet für ordinale Daten
- stabil gegenüber Ausreißern

Median

Beispiel

Gegeben sei zunächst die Werteliste ohne größeren Ausreißer:

$$4, 5, 5, 6, 6, 6, 7, 7, 9.$$

Da $n = 9$ ungerade ist, gilt:

$$\tilde{x}_{0,5} = x_{(n+1)/2} = x_{(9+1)/2} = x_5 = 6$$

Wird nun die Werteliste mit einem deutlichen Ausreißer versehen, also

$$4, 5, 5, 6, 6, 6, 7, 7, 28,$$

so verbleibt der Median bei 6, denn der neue Extremwert übt keinen Einfluß aus, wie die obige Berechnung aufzeigt.

Das Quantil

Definition: Wert für den gilt:

Mindestens Anteil p der Daten sind kleiner oder gleich x_p
Mindestens Anteil $1-p$ der Daten sind größer oder gleich x_p

$$\begin{cases} x_{(k)} & \text{falls } np \text{ keine ganze Zahl und } k \text{ kleinste Zahl} > np \\ \in \left[x_{(k)}; x_{(k+1)}\right] & \text{falls } k = np \text{ ganze Zahl} \end{cases}$$

Es gibt weitere Definitionen von Quantilen (in R 9 Typen), die sich aber in der Praxis kaum unterschieden.

Quantile: Berechnung

Berechnung von Quantilen

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \text{ keine ganze Zahl ist, } k \text{ ist} \\ & \text{dann die kleinste ganze Zahl } > n\alpha, \\ \frac{1}{2}\left(x_{(n\alpha)} + x_{(n\alpha+1)}\right) & \text{falls } n\alpha \text{ ganzzahlig ist.} \end{cases}$$

Quantile

Besondere Quantile

Bei der Charakterisierung von Verteilungen haben folgende Quantile eine besondere Bedeutung:

Median entspricht dem 50%-Quantil (siehe oben)

unteres Quartil entspricht dem 25%-Quantil

oberes Quartil entspricht dem 75%-Quantil

Fünf-Punkte Zusammenfassung

Minimum, 25%-Quantil, Median, 75%-Quantil, Maximum

Beispiel

Gegeben sei die (altbekannte) Werteliste:

$$4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 9.$$

Für das untere Quartil $\tilde{x}_{0,25}$ ergibt sich der Wert 5, denn

$$n\alpha = 9 \cdot 0,25 = 2,25 \text{ nicht ganzzahlig} \Rightarrow \tilde{x}_{0,25} = x_{(3)} = 5$$

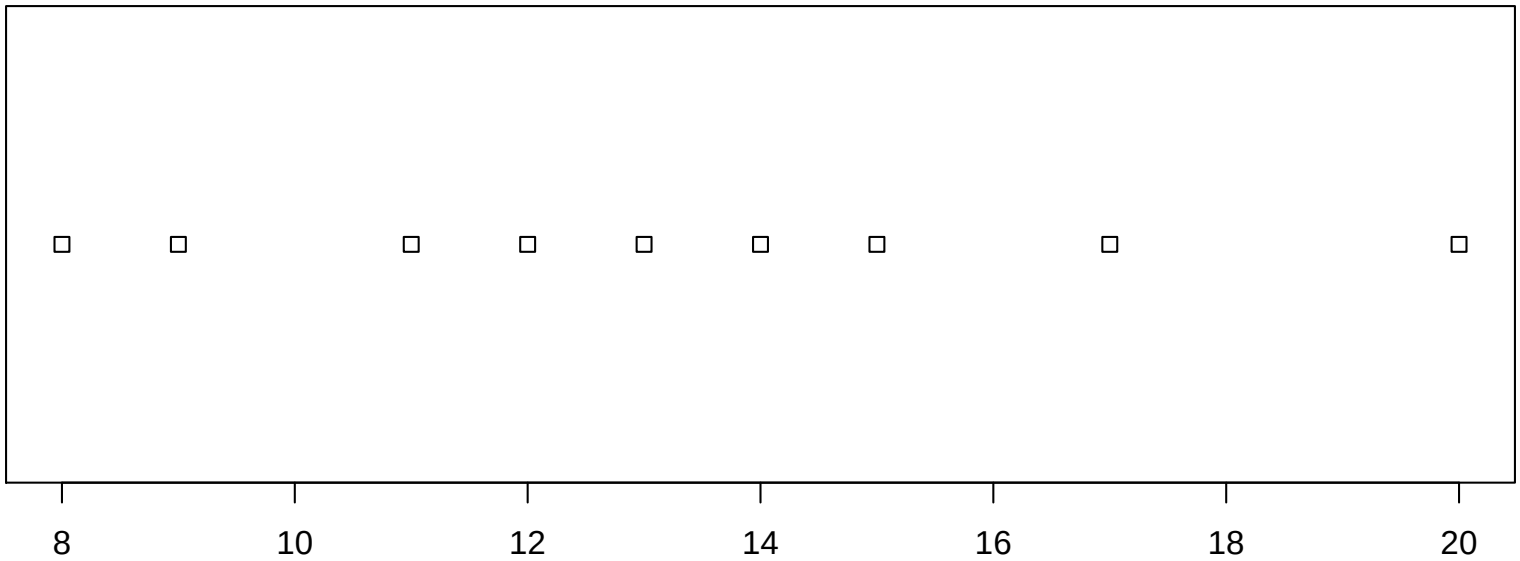
Fünf-Punkte Zusammenfassung: 4,5,6,7,9

Boxplot

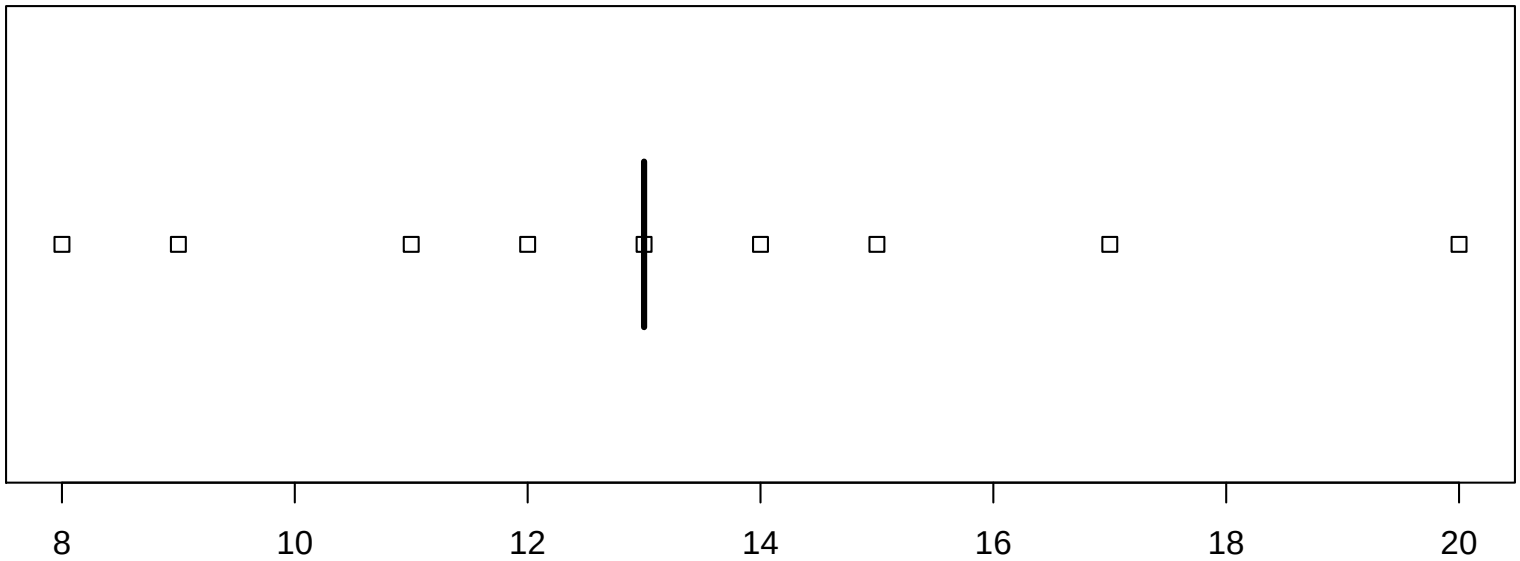
Einfacher Boxplot

- $\tilde{x}_{0.25}$ = Anfang der Schachtel (Box)
  $\tilde{x}_{0.75}$ = Ende der Schachtel
  d_Q = Länge der Schachtel
- Der Median wird durch den Strich in der Box markiert
- Zwei Linien („whiskers“) außerhalb der Box gehen bis zu x_{min} und x_{max}.

Einfacher Boxplot

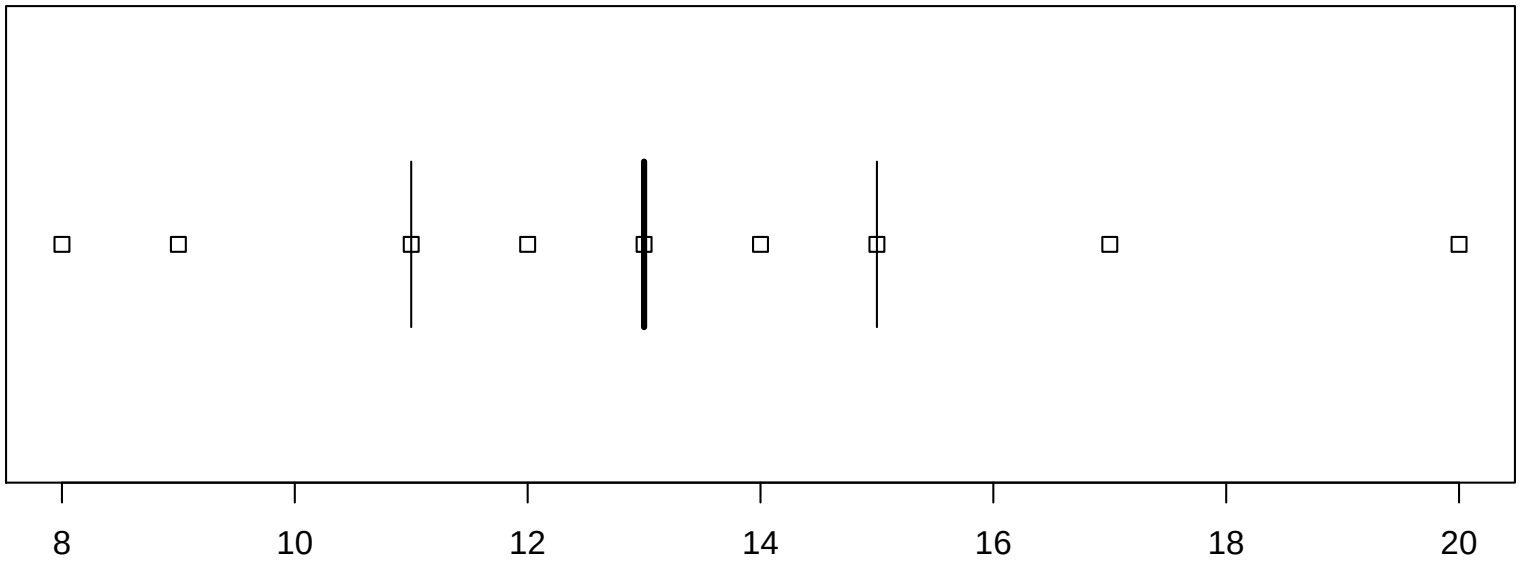


Einfacher Boxplot



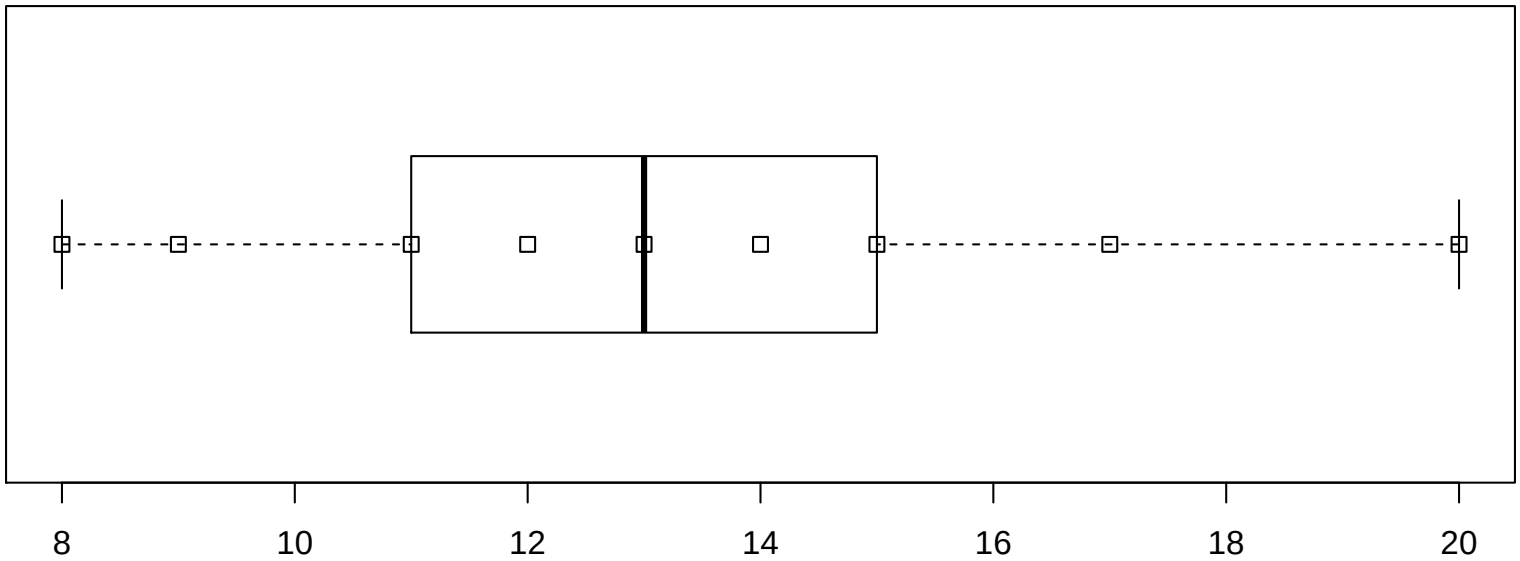
- Median: $x_{(5)}$

Einfacher Boxplot



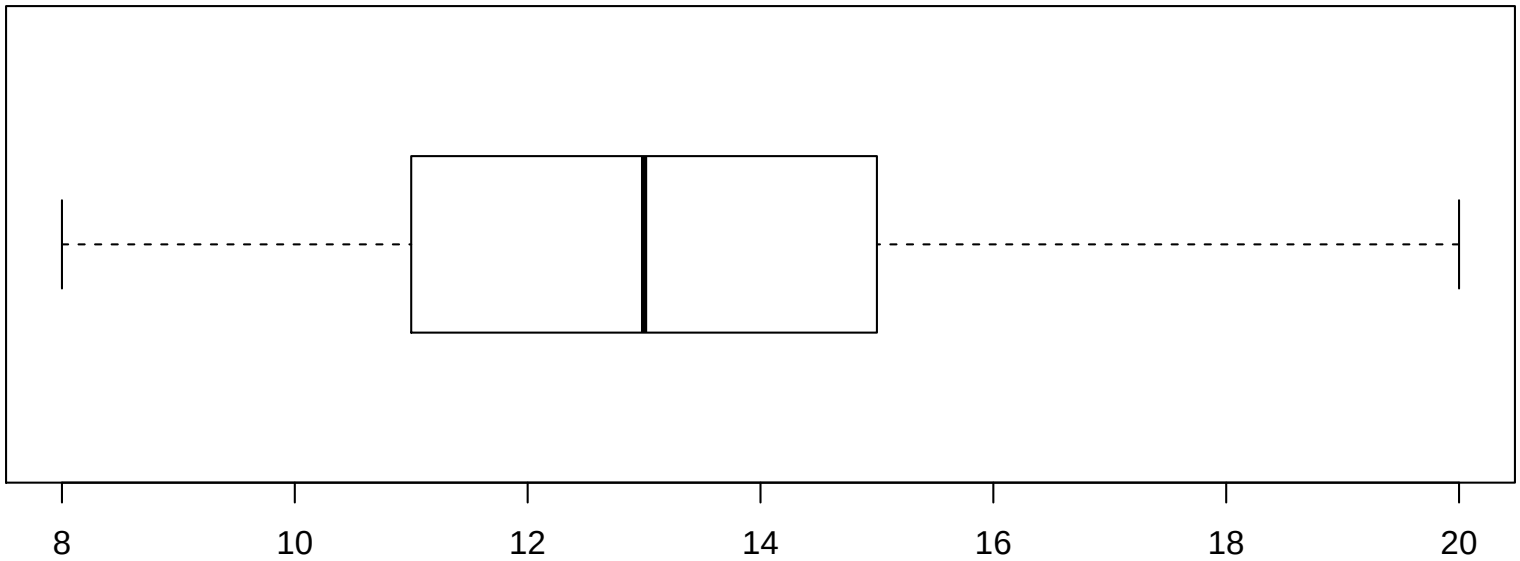
- Median: $x_{(5)}$
- unteres Quartil: $x_{(3)}$
- oberes Quartil: $x_{(7)}$

Einfacher Boxplot



- Median: $x_{(5)}$
- unteres Quartil: $x_{(3)}$
- oberes Quartil: $x_{(7)}$
- Minimum: $x_{(1)}$
- Maximum: $x_{(9)}$

Einfacher Boxplot



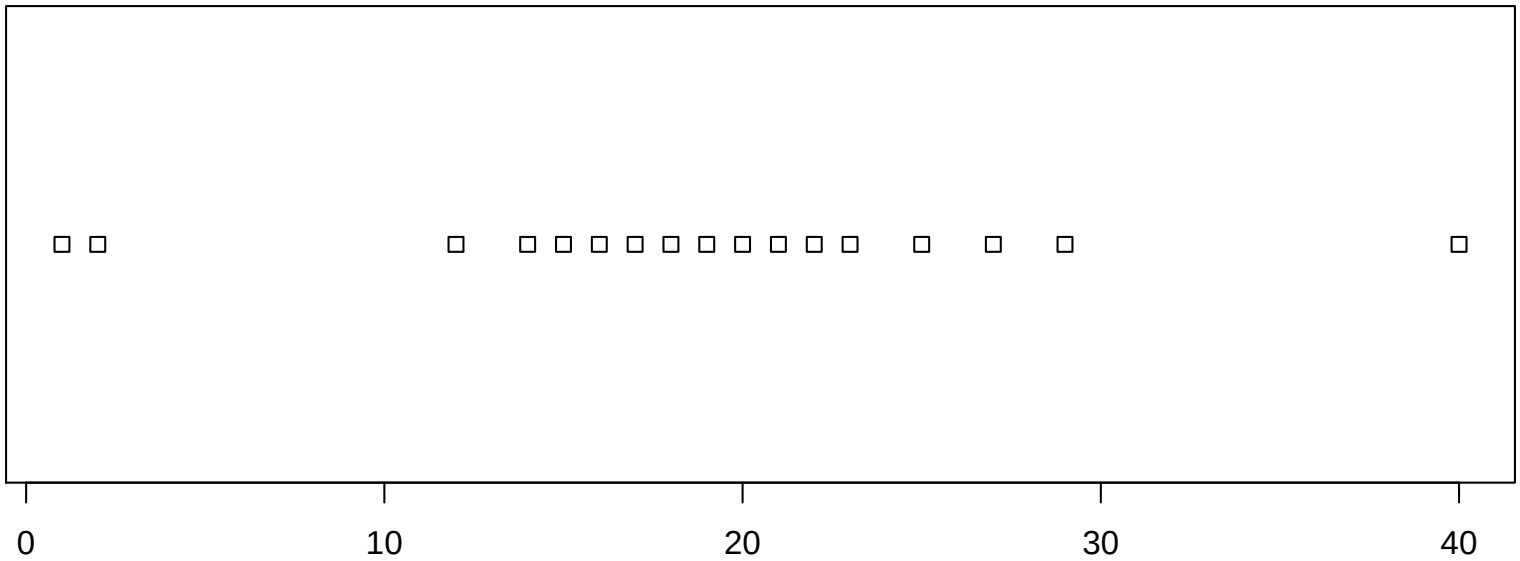
- Median: $x_{(5)}$
- unteres Quartil: $x_{(3)}$
- oberes Quartil: $x_{(7)}$
- Minimum: $x_{(1)}$
- Maximum: $x_{(9)}$

Boxplot

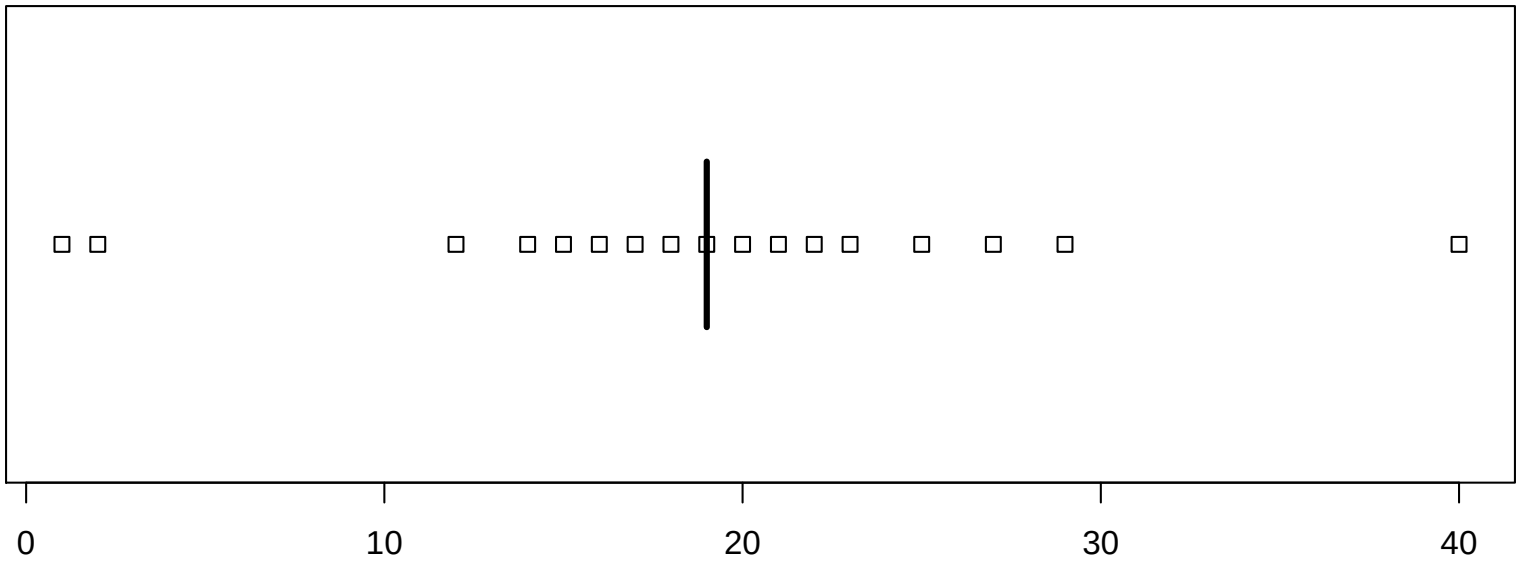
Modifizierter Boxplot

Die Linien außerhalb der Schachtel werden nur bis zu x_{min} bzw. x_{max} gezogen, falls x_{min} und x_{max} innerhalb des Bereichs $[z_u, z_o]$ der Zäune liegen.
$z_u = \tilde{x}_{0.25} - 1,5d_Q$, $z_o = \tilde{x}_{0.75} + 1,5d_Q$
Ansonsten gehen die Linien nur bis zum kleinsten bzw. größten Wert innerhalb der Zäune, die außerhalb liegenden Werte werden individuell eingezeichnet.

Modifizierter Boxplot

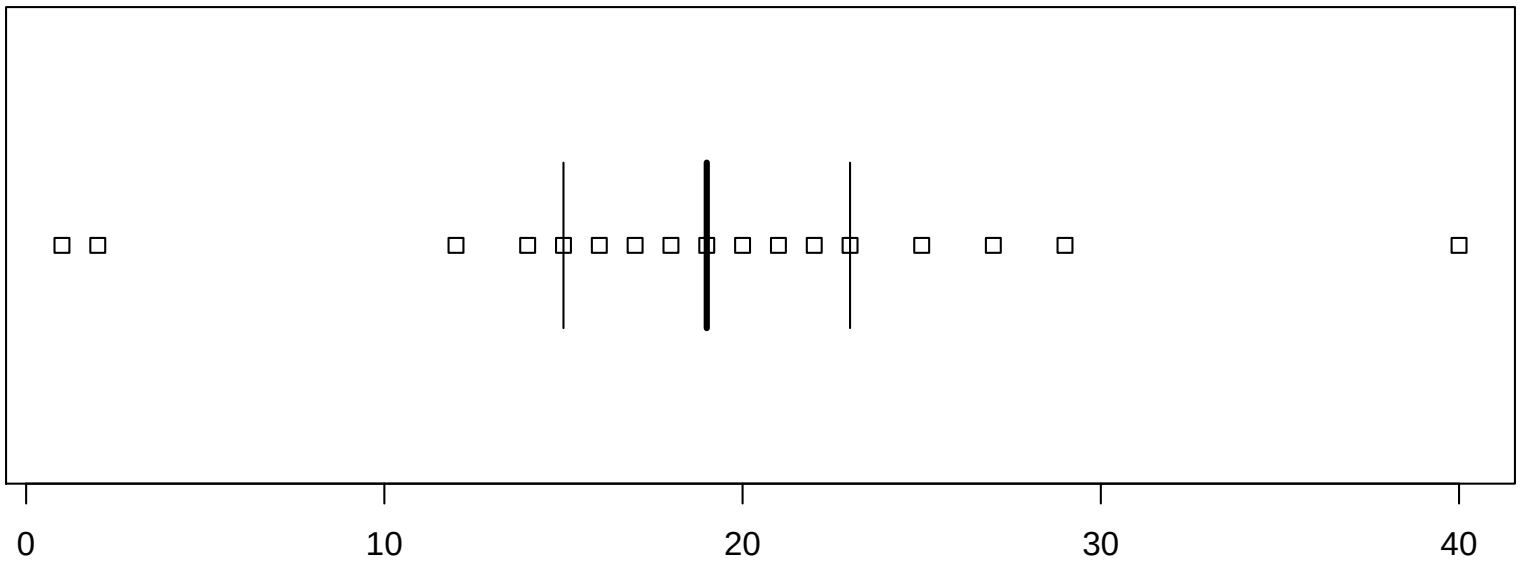


Modifizierter Boxplot



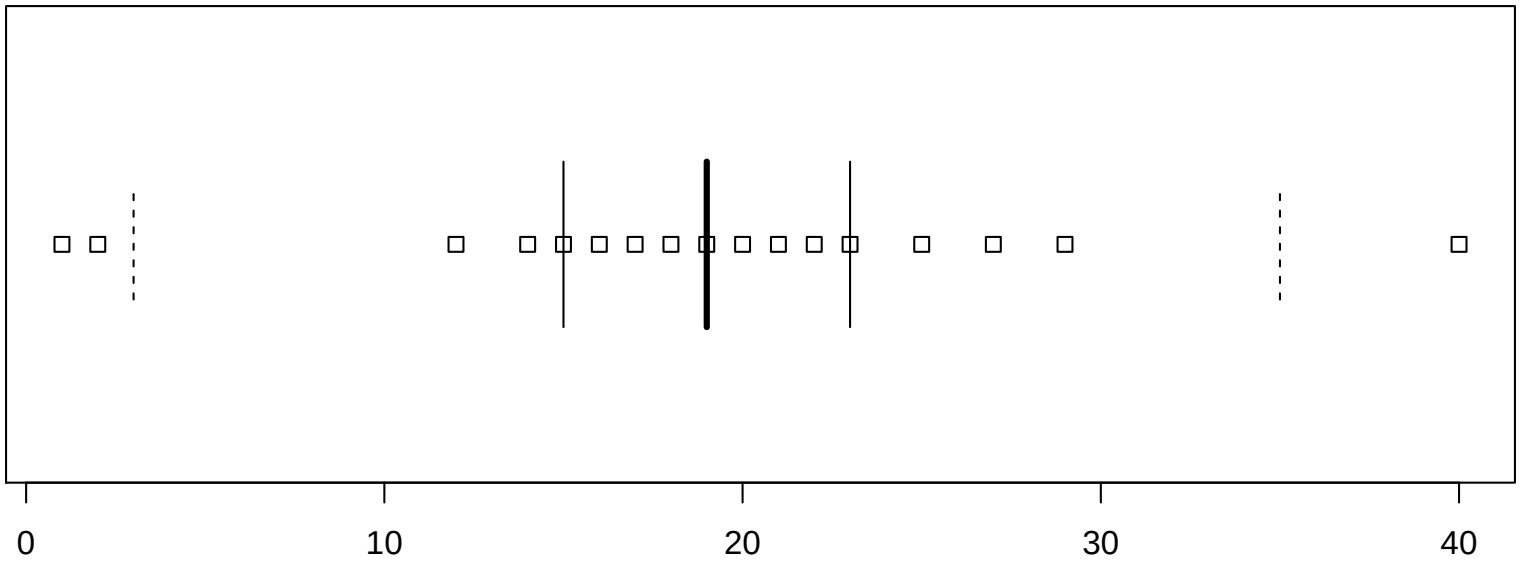
- Median: $x_{(9)}$,

Modifizierter Boxplot



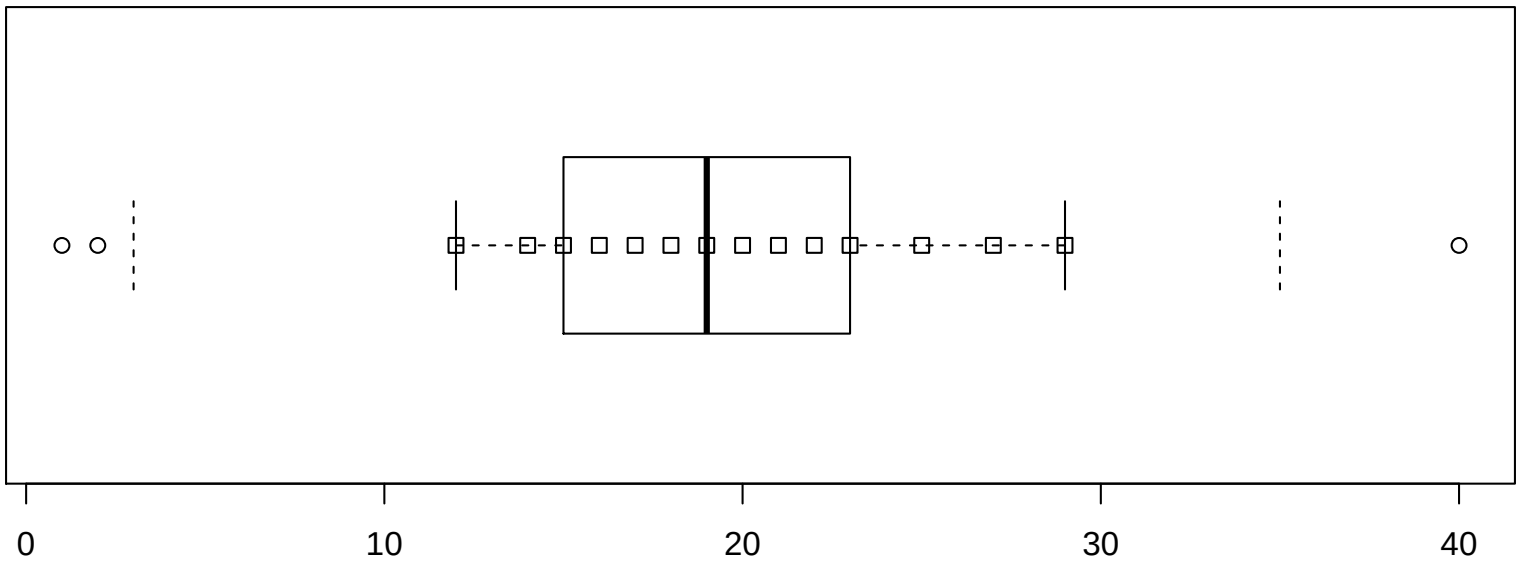
- Median: $x_{(9)}$, unteres Quartil: $x_{(5)}$, oberes Quartil: $x_{(13)}$

Modifizierter Boxplot



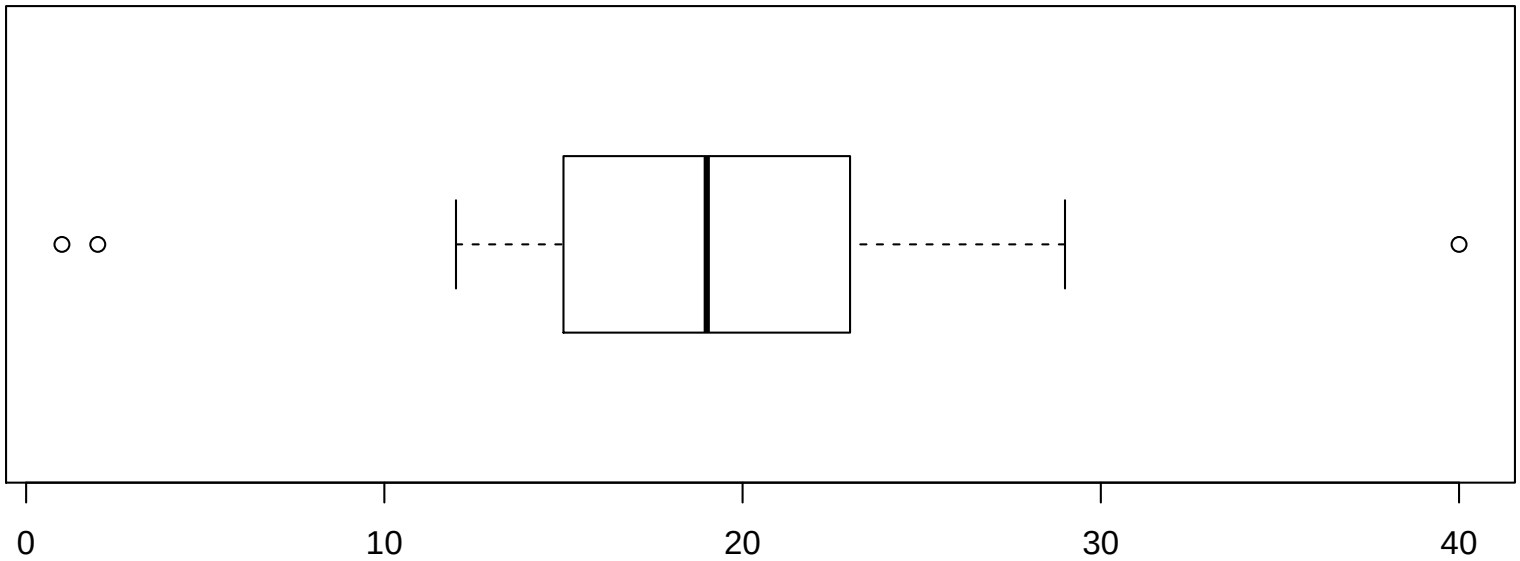
- Median: $x_{(9)}$, unteres Quartil: $x_{(5)}$, oberes Quartil: $x_{(13)}$
- Maximale Whiskerlänge unten: $x_{(5)} - 1.5d = 3$
- Maximale Whiskerlänge oben: $x_{(13)} + 1.5d = 35$

Modifizierter Boxplot



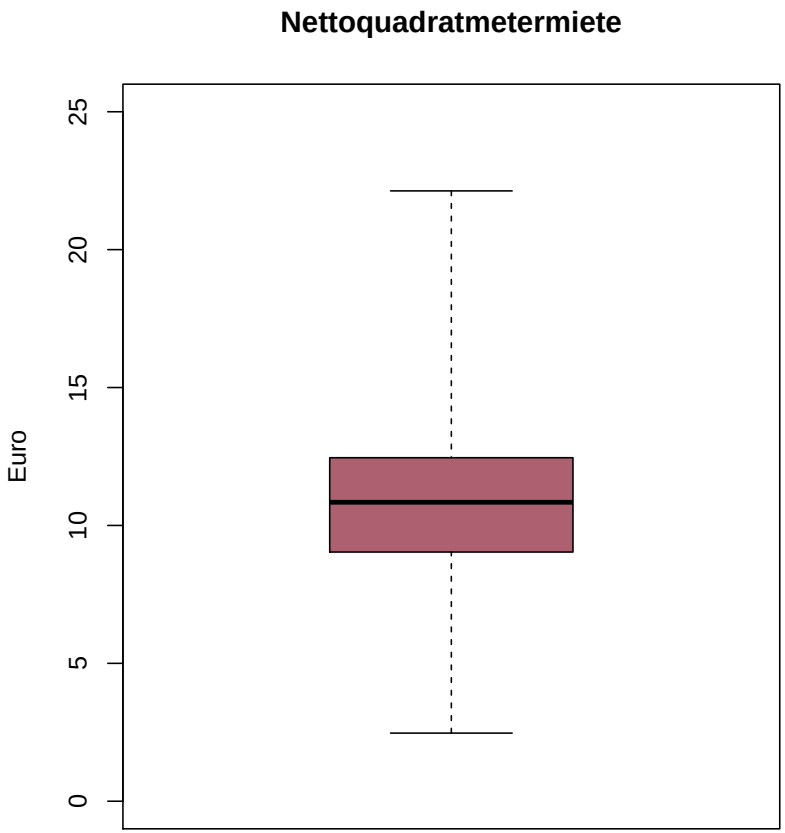
- Median: $x_{(9)}$, unteres Quartil: $x_{(5)}$, oberes Quartil: $x_{(13)}$
- Maximale Whiskerlänge unten: $x_{(5)} - 1.5d = 3$
- Maximale Whiskerlänge oben: $x_{(13)} + 1.5d = 35$
- x_{min} ohne Ausreißer: $x_{(3)} = 12$
- x_{max} ohne Ausreißer: $x_{(16)} = 29$

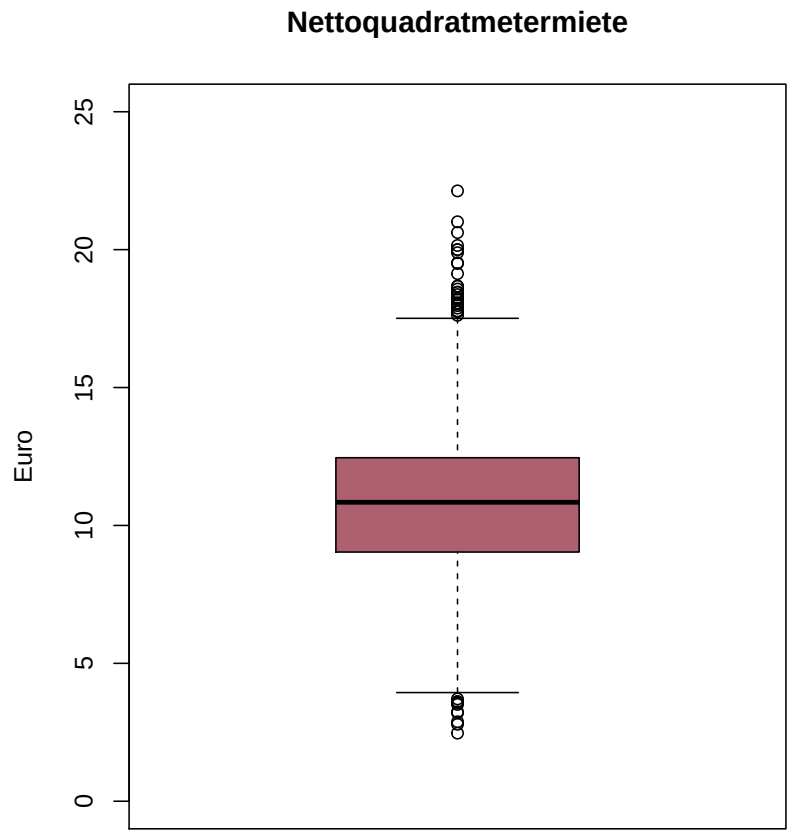
Modifizierter Boxplot



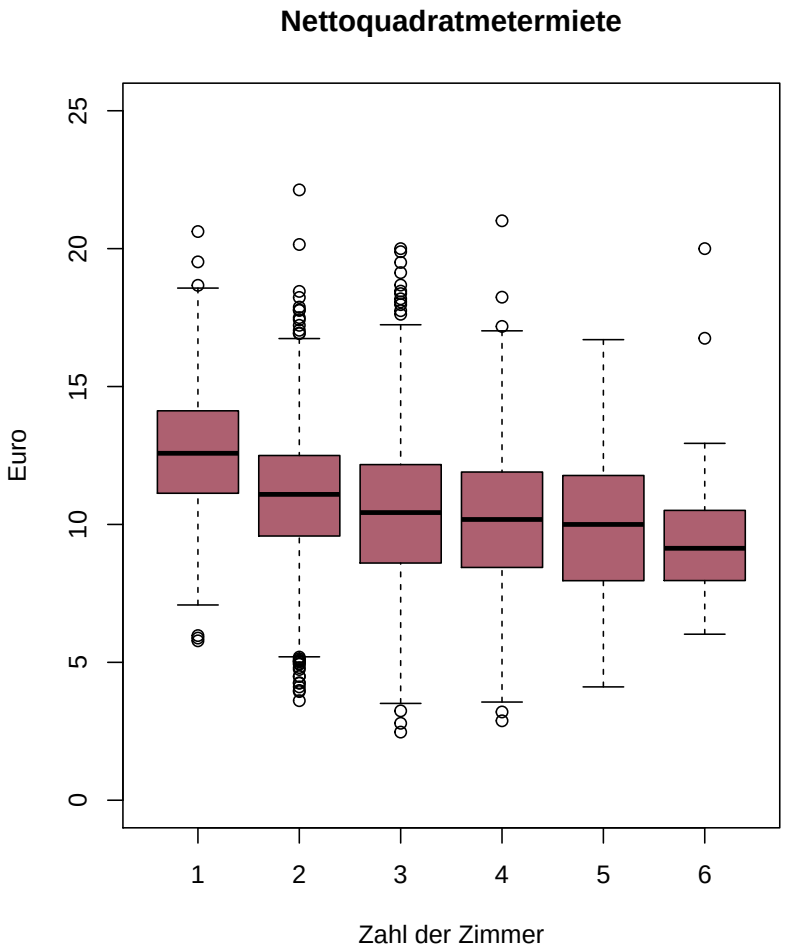
- Median: $x_{(9)}$, unteres Quartil: $x_{(5)}$, oberes Quartil: $x_{(13)}$
- Maximale Whiskerlänge unten: $x_{(5)} - 1.5d = 3$
- Maximale Whiskerlänge oben: $x_{(13)} + 1.5d = 35$
- x_{min} ohne Ausreißer: $x_{(3)} = 12$
- x_{max} ohne Ausreißer: $x_{(16)} = 29$

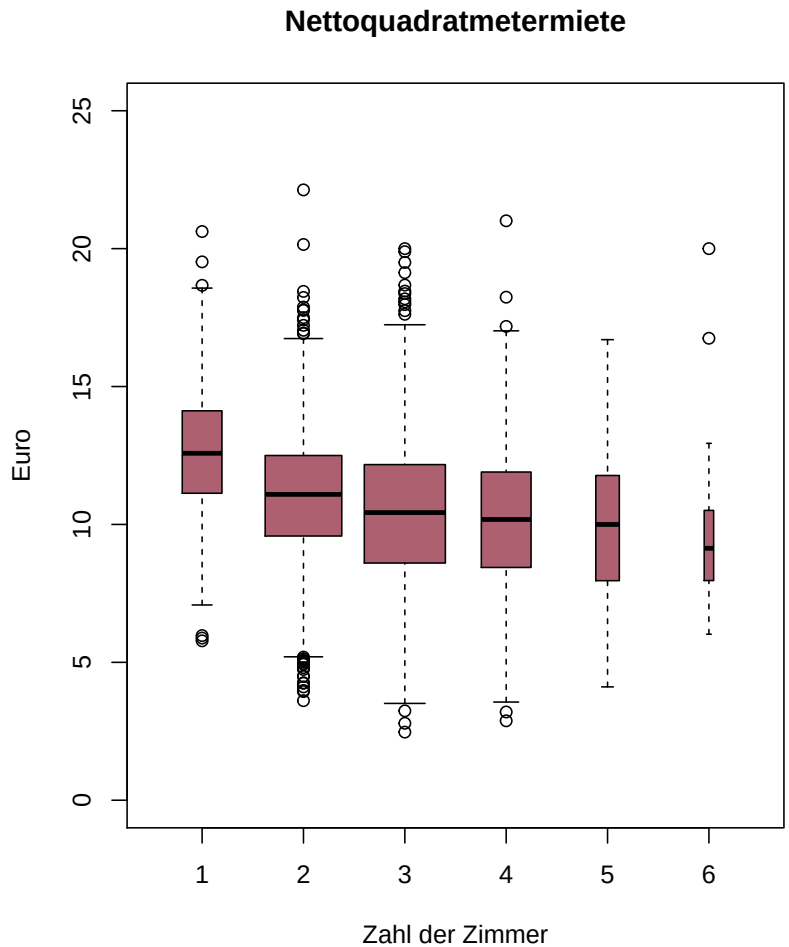
Beispiel: Münchner Mietspiegel



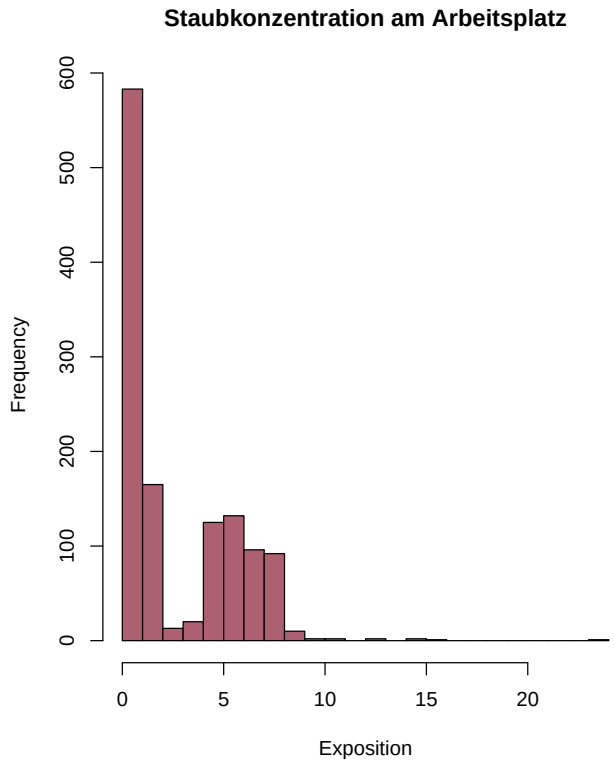


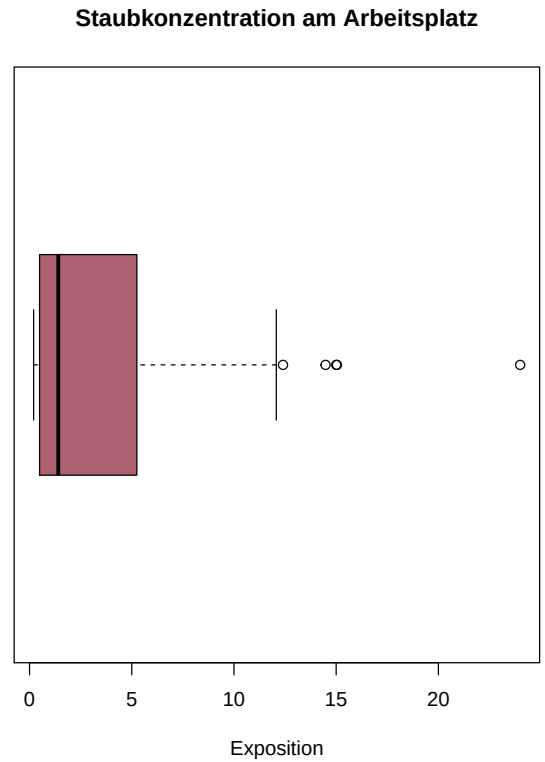
Beispiel: Münchner Mietspiegel





Beispiel: Münchener Staubdaten





- Beachte: Bimodale Verteilung im Boxplot nicht erkennbar.

Boxplot: Vor- und Nachteile

pro:

- kompakt
- geeignet für Vergleiche
- Ausreißer sichtbar
- Schiefe sichtbar

contra

- gegen Intuition (Viel Farbe – wenig Daten)
- Bimodale Verteilungen nicht sichtbar
- Ausreißer sichtbar
- Breite redundant

Der Mittelwert (arithmetisches Mittel)

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

- bekanntestes Lagemaß
- instabil gegen extreme Werte
- geeignet für intervallskalierte Daten

Mittelwert bei gruppierten Daten

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\
&= \frac{1}{n}\sum_{j=1}^{k} h_j a_j
\end{aligned}$$

h_j : Häufigkeit von a_j

Getrimmtes Mittel

Um die Ausreißerempfindlichkeit von $\bar{x}$ abzuschwächen definiert man

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

$x_{(i)}$: geordnete x-Werte
r ist die größte ganze Zahl mit $r \leq n\alpha$

Es wird also der Anteil α der extremsten Werte abgeschnitten.

„α-getrimmtes Mittel“

Winsorisiertes Mittel (gestutztes Mittel)

Der Anteil α der extremsten Werte wird durch das entsprechende Quantil ersetzt.

Das geometrische Mittel

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- arithmetisches Mittel auf der log-Skala

$$x_g = exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)$$

- nur geeignet für positive Werte
- geeignet für intervallskalierte Daten

Das harmonische Mittel

$$\bar{x}_H := \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Das harmonische Mittel entspricht dem Mittel durch Transformation

$$t \to \frac{1}{t} \qquad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Das harmonische Mittel

$$\bar{x}_H := \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Das harmonische Mittel entspricht dem Mittel durch Transformation

$$t \to \frac{1}{t} \qquad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Beispiel:

$x_1, \ldots, x_n$ Geschwindigkeiten, mit denen konstante Wegstrecken L zurückgelegt werden

Gesamt-Geschwindigkeit:

$$\frac{L \cdot n}{\frac{L}{x_1} + \ldots + \frac{L}{x_n}} = \bar{x}_H$$

Ergänzungen zu Mittelwert und Median

Transformationen

Linearität des arithmetischen Mittels

Gegeben sind Daten $x_1, \ldots x_n$ und eine lineare Transformation $y_i = a + b \cdot x_i$
Dann gilt

$$\bar{y} = a + b \cdot \bar{x}$$

Beispiel: x_i Gewinn in €; y_i Gewinn in SFR

Invarianz des Medians bei monotonen Transformationen

Gegeben sind Daten $x_1, \ldots x_n$ und eine streng monoton steigende Transformation $y_i = f(x_i)$, d.h. $x_i < x_j \Rightarrow f(x_i) < f(y_i)$ Dann gilt

$$\tilde{y}_{0,5} = f(\tilde{x}_{0,5})$$

Beispiel: Logarithmierung, Umrechnung von Punkten in Noten.

4 Streumaße

- Spannweite
- Interquartilsabstand
- Standardabweichung und Varianz
- Variationskoeffizient
- MAD

Streumaße

Motivation

Lagemaße allein charakterisieren die Verteilung nur unzureichend!

> *Wenn man den Kopf in der Sauna hat und die Füße im Kühlschrank, sprechen Statistiker von einer angenehmen mittleren Temperatur.*
> *Zwei Männer sitzen im Wirtshaus. Der eine verdrückt eine ganze Kalbshaxe, der andere trinkt zwei Maß Bier. Statistisch gesehen ist das für jeden ein Maß Bier und eine halbe Haxe - aber der eine hat sich überfressen, der andere ist besoffen."*

Die statistische Sicht soll sich also nicht auf den Mittelwert beschränken!!

Maße für die Streuung

- Spannweite
- Interquartilsabstand
- Standardabweichung und Varianz
- Variationskoeffizient

Die Spannweite (Range)

Definition:

$$q = x_{max} - x_{min}$$

- „Bereich in dem die Daten liegen“
- Wichtig für Datenkontrolle

Der Quartilsabstand

Definition:

$$d_Q = x_{0.75} - x_{0.25}$$

- „Größe des Bereichs in dem die mittlere Hälfte der Daten liegt“
- Bei ordinal skalierten Daten Angabe von $x_{0.75}$ und $x_{0.25}$
- Zentraler 50%-Bereich
- Robust gegen Ausreißer

Standardabweichung und Varianz

Definition

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{Varianz}$$

$$s = \sqrt{s^2} \qquad \textit{Standardabweichung}$$

- „Mittlere Abweichung vom Mittelwert“
- Intervallskala Voraussetzung
- Empfindlich gegen Ausreißer
- Verwende $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ für Stichproben

Eigenschaften der Standardabweichung

Die Standardabweichung hat gegenüber der Varianz den Vorteil, dass sie in der *gleichen Einheit* wie die Beobachtungswerte gemessen wird.

Transformationsregel

$y_i = a + bx_i$

$$\begin{aligned} \Rightarrow s_y^2 &= b^2 s_x^2 \\ s_y &= |b| s_x \quad \text{(Analog für } S_x, S_y\text{)} \end{aligned}$$

Varianz und Standardabweichung sind
mit linearen Transformationen verträglich.

Verschiebungssatz

Für jedes $c \in \mathbb{R}$ gilt:

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2$$

$$\begin{aligned} c = 0 \Rightarrow s^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \\ s^2 &= \overline{x^2} - \bar{x}^2 \end{aligned}$$

Beachte:

- Mittelwert minimiert $\sum_{i=1}^{n}(x_i - c)^2$
- Verschiebungssatz für numerische Berechnung mit Computer nicht geeignet.

Streuungszerlegung I

Seien die Daten in r Gruppen (Schichten) aufgeteilt:

$$x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, \ldots, x_{n_r}$$

Gruppenmittelwerte:

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i, \quad \text{usw.}$$

Gruppenvarianzen:

$$s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2, \quad s_2^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2, \quad \text{usw.}$$

Streuungszerlegung II

Dann gilt:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{r} n_j \bar{x}_j$$

$$s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j s_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2$$

$$\text{Gesamtstreuung} = \begin{array}{c}\text{Streuung}\\ \text{innerhalb}\\ \text{der Gruppen}\end{array} + \begin{array}{c}\text{Streuung}\\ \text{zwischen}\\ \text{den Gruppen}\end{array}$$

Variationskoeffizient

Definition

Das Verhältnis von Standardabweichung und Mittelwert ist gegeben durch

$$v = \frac{s}{\bar{x}} \quad \text{mit } \bar{x} > 0$$

Eigenschaften des Variationskoeffizienten

- misst die relative Schwankung um den Mittelwert
- ist nur bei positiven Werten bei Verhältnisskala sinnvoll
- ermöglicht den Vergleich von Streuungen zweier Datensätze mit unterschiedlichen Maßeinheiten

Mittlere absolute Abweichung (MAD)

Definition

Die mittlere absolute Abweichung ist definiert als

$$x_{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Eigenschaften des MAD

- misst direkt die durchschnittliche absolute Abweichung um den Mittelwert
- ist nicht ausreisserempfindlich
- hat nicht so „schöne“Eigenschaften wie die Standardabweichung

Konzentrationsmaße

Motivation

Existiert eine Menge, die auf viele Individuen verteilt ist, kann es hilfreich sein zu wissen, wie diese Menge verteilt ist; ob etwa eher eine Gleichverteilung oder eher ein Monopol vorliegt.

Beispiele

- Vermögensverteilung in einem Staat
- Marktanteile von Firmen in einem Segment

Lorenzkurve

verwendbar bei Merkmalen mit

Nominalskala	Ordinalskala	metrische Skala
		x

Definition Lorenzkurve

- Das Merkmal darf nur *positive* Ausprägungen annehmen
- Die Gesamtsumme aller Merkmalswerte ist $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_{(j)}$
- Die Lorenzkurve verbindet Punktepaare bestehend aus den *kumulierten Summen* der nach Größe geordneten Beobachtungswerte $0 \leq x_{(1)} \leq \ldots \leq x_{(n)}$ und dem *relativen Anteil* der Individuen, die diese kumulierte Summe besitzen.

Lorenzkurve

Gestaltung

- Es wird festgelegt: $u_{(0)} = 0$ und $v_{(0)} = 0$
- Die Abszisse wird in *gleiche Längen* aufgeteilt, deren Anzahl der der Individuen (Merkmalsausprägungen) entspricht:

$$u_i = \frac{i}{n}, \quad i = 1, ..., n$$

- Die Unterteilung der Ordinate berechnet sich wie folgt:

$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \quad i = 1, ..., n,$$

  also dem Quotienten aus der kumulierten Summe und der Gesamtsumme.
- Die so errechneten Koordinatenpunkte werden in den Graphen eingetragen und mit Geraden verbunden.

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20		
2	20		
3	20		
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	$\frac{1}{5}$	$\frac{20}{100}$
2	20		
3	20		
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20		
3	20		
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	$\frac{2}{5}$	$\frac{40}{100}$
3	20		
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20		
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	$\frac{3}{5}$	$\frac{60}{100}$
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20		
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	$\frac{4}{5}$	$\frac{80}{100}$
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	0,8	0,8
5	20		

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	0,8	0,8
5	20	$\frac{5}{5}$	$\frac{100}{100}$

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

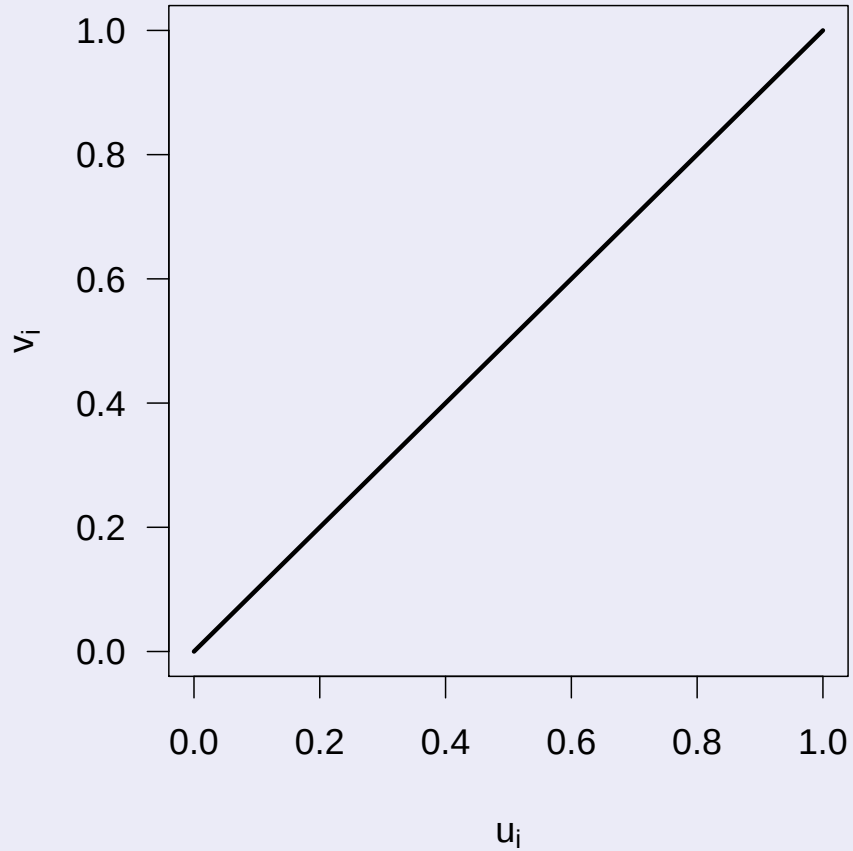
i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	0,8	0,8
5	20	1	1

Lorenzkurve

Beispiel einer Gleichverteilung

5 Bauern teilen sich eine Ackerfläche von 100 ha zu je 20 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	0,8	0,8
5	20	1	1



Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0		
2	0		
3	0		
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	$\frac{1}{5}$	$\frac{0}{100}$
2	0		
3	0		
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0		
3	0		
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	$\frac{2}{5}$	$\frac{0}{100}$
3	0		
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0		
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	$\frac{3}{5}$	$\frac{0}{100}$
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0		
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	$\frac{4}{5}$	$\frac{0}{100}$
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	0,8	0
5	100		

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	0,8	0
5	100	$\frac{5}{5}$	$\frac{100}{100}$

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

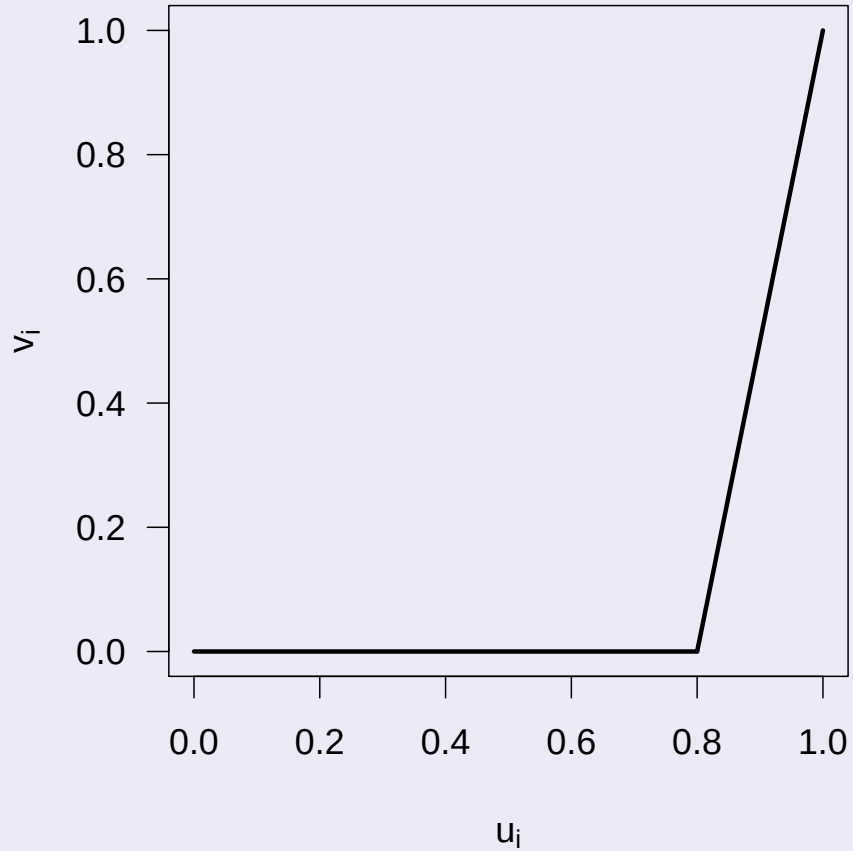
i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	0,8	0
5	100	1	1

Lorenzkurve

Beispiel eines Monopols

Von 5 Bauern besitzt einer die gesamten 100 ha.

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	0,8	0
5	100	1	1



Lorenzkurve

Erscheinungsbild von Lorenzkurven

- Die Kurve bildet auf einen quadratischen Graphen mit Kantenlänge 1 ab.
- Die Koordinate $(u_0; v_0)$ ist *immer* $(0; 0)$.
- Die Koordinate $(u_n; v_n)$ ist *immer* $(1; 1)$.
- Der konstruierte Polygonzug verläuft *immer unterhalb* (im Grenzfall auf) der Winkelhalbierenden.
- Der konstruierte Polygonzug ist *(streng) monoton steigend*.
- Die Steigung des nächsten Polygonsegments ist entweder *gleich groß* oder *größer* als die Steigung des letzten Polygonsegments.

Lorenzkurve

Aussagemöglichkeiten von Lorenzkurven

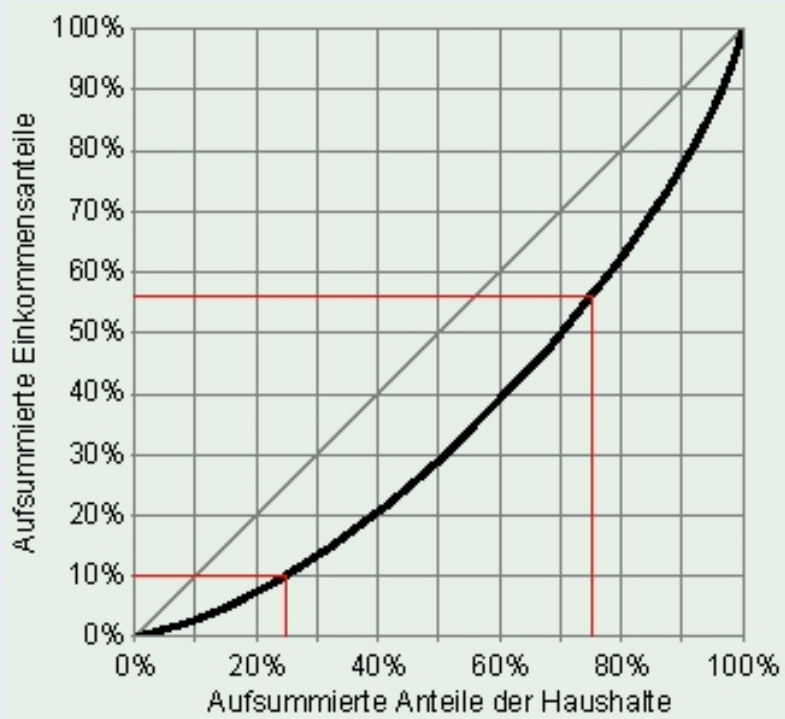
Aufgrund ihrer Struktur kann man anhand einer Lorenzkurve folgende Aussagen verfassen:

- Die „ärmsten“ x% besitzten einen Anteil von y%.
- Die „reichsten“ x% besitzen einen Anteil von y%.

Lorenzkurve

Beispiel

Bruttohaushaltseinkommen 2003 in der Schweiz



Es zeigt sich, dass das ärmste Viertel der Schweizer Bevölkerung nur 10%, das reichste Viertel jedoch über 40% des gesamten Bruttoeinkommens verdient.

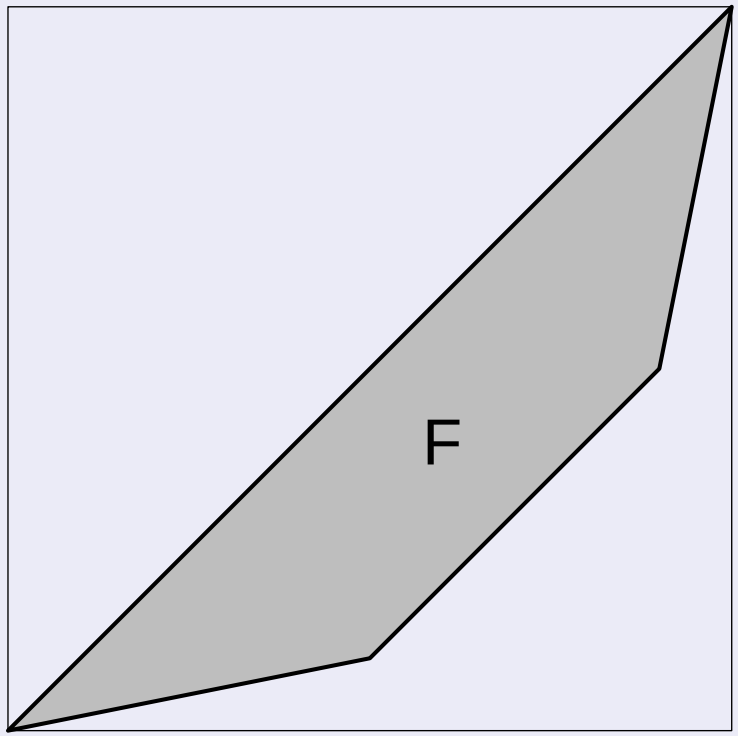
Gini-Koeffizient

Definition Gini-Koeffizient

Der Gini-Koeffizient bzw. das Lorenzsche Konzentrationsmaß ist eine Maßzahl, die das *Ausmaß* der Konzentration beschreibt. Er ist definiert als

$$G = 2 \cdot F,$$

wobei F die Fläche zwischen der Diagonalen und der Lorenzkurve ist.



Gini-Koeffizient

Berechnung des Gini-Koeffizienten

Für die praktische Berechnung von G aus den Wertepaaren $(u_i; v_i)$ stehen folgende alternative Formeln zur Verfügung:

$$G = \frac{2\sum_{i=1}^{n} i \cdot x_{(i)} - (n+1)\sum_{i=1}^{n} x_{(i)}}{n\sum_{i=i}^{n} x_{(i)}}$$

oder alternativ

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}(v_{i-1} + v_i)$$

Wertebereich des Gini-Koeffizienten

$$0 \leq G \leq \frac{n-1}{n}$$

Gini-Koeffizient

der normierte Gini-Koeffizient G^+

Der Gini-Koeffizient wird auf folgende Weise normiert:

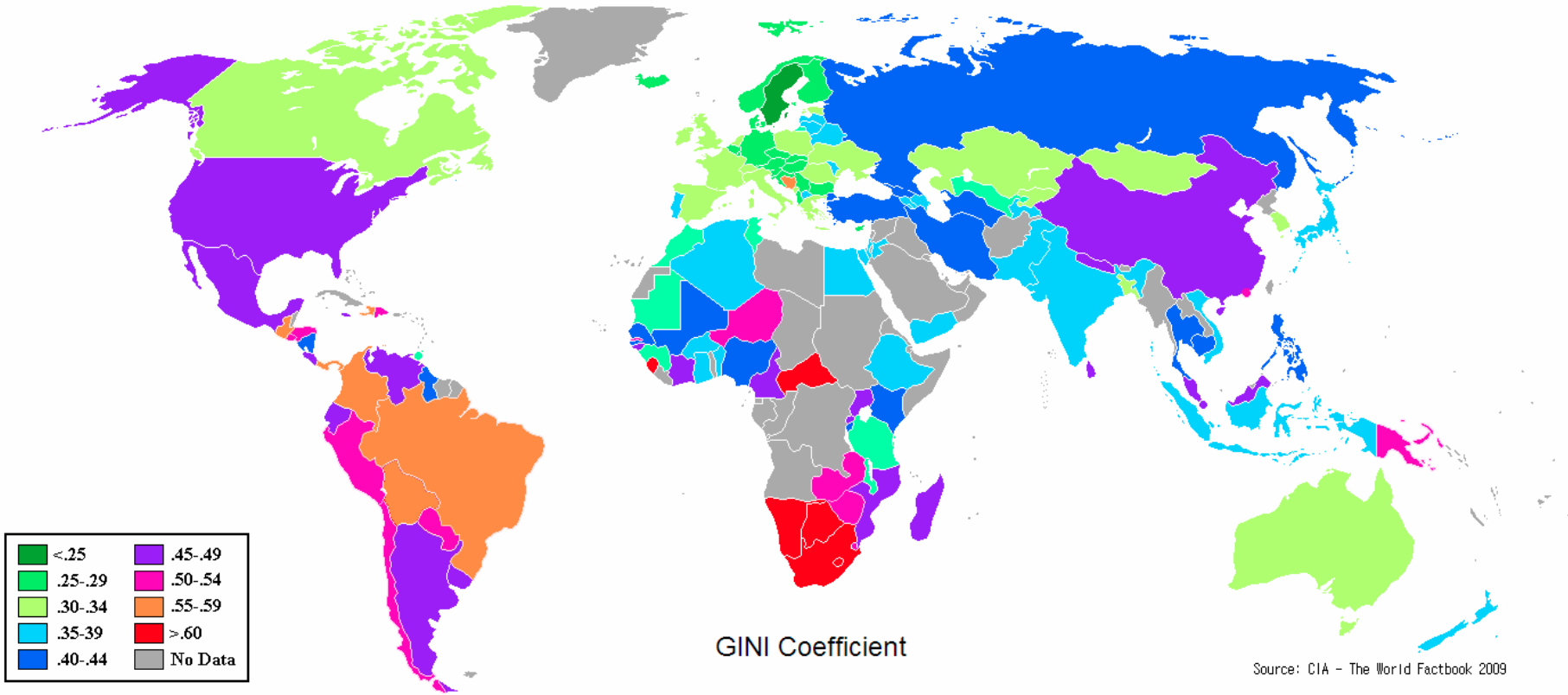
$$G^+ = \frac{n}{n-1} G$$

Er hat somit den Wertebereich

$$0 \leq G^+ \leq 1,$$

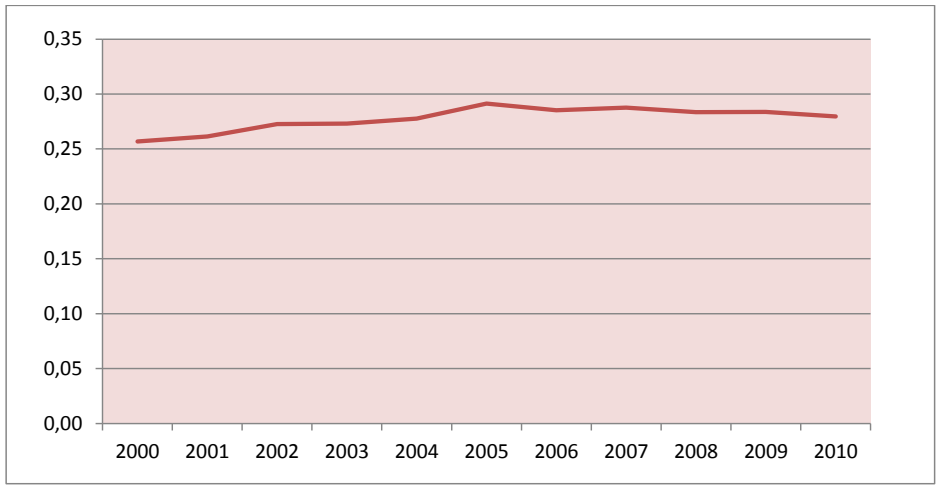
wobei 0 für *keine Konzentration* (Gleichverteilung) und 1 für *vollständige Konzentration* (Monopol) steht.

GINI Einkommen nach CIA report 2009



Bericht der Bundesregierung

Einkommensverteilung (Gini-Koeffizient)



Quelle: Berechnungen des DIW Berlin auf Basis SOEP 2011.

Ein weiteres Verteilungsmaß ist der Gini-Koeffizient. Er beschreibt auf einer Skala von null bis eins die Ungleichheit der Verteilung. Je höher der Wert, umso ungleicher ist die Verteilung. Dieses Maß zeigt eine nach 2007 rückläufige Ungleichheit der Nettoäquivalenzeinkommen auf Haushaltsebene an. Dies umfasst alle Einkommensarten (insbesondere Einkommen aus Erwerb, Renten und Pensionen, aus Vermögen und Sozialtransfers). Der Trend einer Zunahme zwischen 2000 und 2005 hat sich also in der Zeit danach umgekehrt. Die Ungleichheit der Einkommen nimmt derzeit ab.

personenhaushalt berücksichtigt. Die Verteilung der so ermittelten Nettoäquivalenzeinkommen hat sich, gemessen am Gini-Koeffizienten und den Anteilen der Dezile, nach den Daten der EVS zwischen 2003 und 2008 leicht weiter gespreizt.

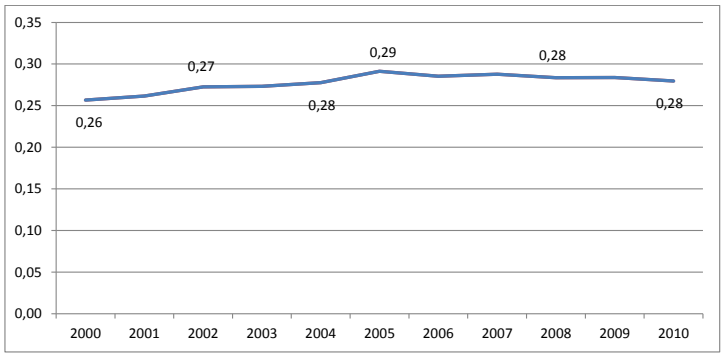
Tabelle C I.1.2:
Verteilung der Nettoäquivalenzeinkommen 2003 und 2008

Jahr	Dezil										Gini-Koeffizient
	1	2	3	4	5	6	7	8	9	10	
	Anteile (%) am Volumen des Nettoäquivalenzeinkommens										
2003	3,9	5,5	6,5	7,5	8,4	9,4	10,5	12,0	14,3	22,0	0,267
2008	3,6	5,1	6,3	7,3	8,3	9,3	10,5	12,2	14,7	22,7	0,284

Quelle: EVS; Statistisches Bundesamt.

Während die unteren sechs Dezile gegenüber 2003 einen geringeren Anteil aufweisen, haben die obersten drei Dezile Zuwächse erfahren. Der Gini-Koeffizient stieg von 0,267 auf 0,284 und damit um rund sechs Prozent (Tabelle C I.1.2). Nach den Daten des SOEP zeigt dieses Maß eine nach 2007 rückläufige Ungleichheit der Nettoäquivalenzeinkommen auf Haushaltsebene an. Der Trend einer Zunahme zwischen 2000 und 2005 hat sich also in der Zeit danach umgekehrt. Die Ungleichheit der Einkommen nimmt derzeit ab (Schaubild C I.1.1).[338]

Schaubild C I.1.1:
Ungleichheit der Einkommensverteilung in Deutschland, 2000-2011 (Gini-Koeffizient)



Quelle: Berechnungen im DIW auf Basis SOEP 2011. Werte auf zwei Nachkommastellen gerundet.

[338] Vgl. Grabka, M. M. u. a. (2012): Höhepunkt der Einkommensungleichheit in Deutschland überschritten? In: DIW Wochenbericht 43/2012.

5 Analyse von Zusammenhängen

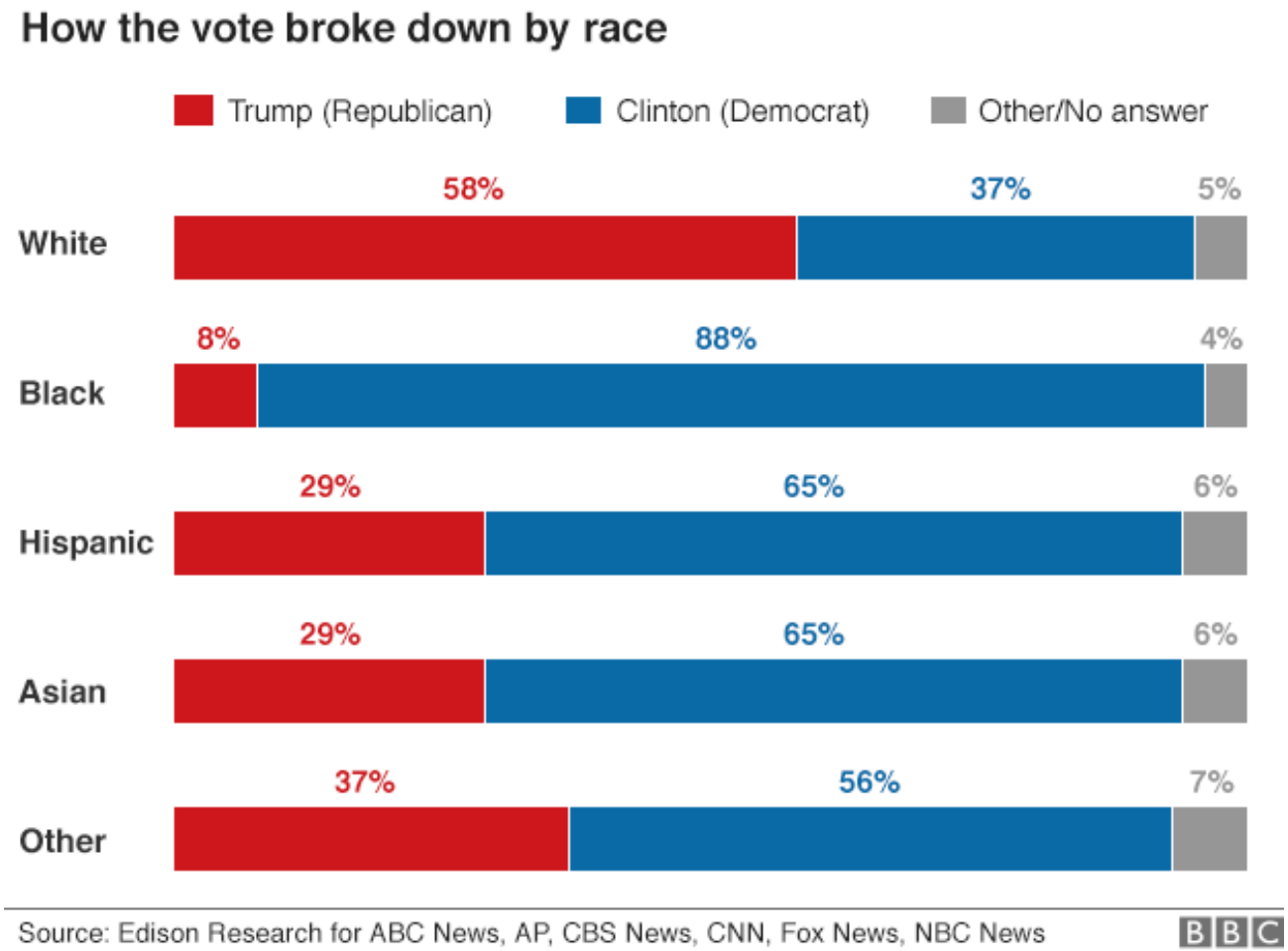
Multivariate beschreibende Statistik

Motivation

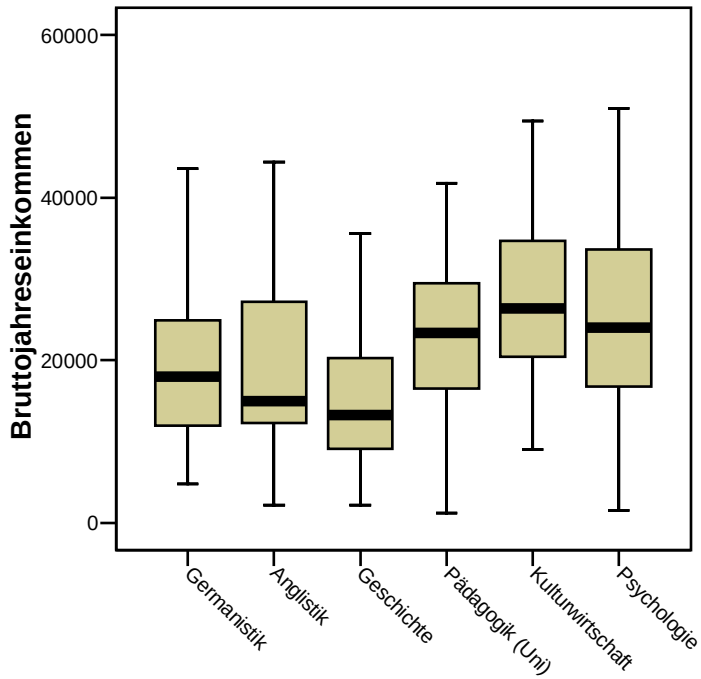
Bei Datenanalysen werden meist mehrere Merkmale X Y Z betrachtet. Fragestellungen

- Gibt es einen Zusammenhang zwischen X und Y ?
- *Wie stark* ist der Zusammenhang ?
- Wird Y von X *beeinflusst* ?
- Kann Y mit Hilfe von X und Z *prognostiziert* werden ?

Beispiel: US Wahl



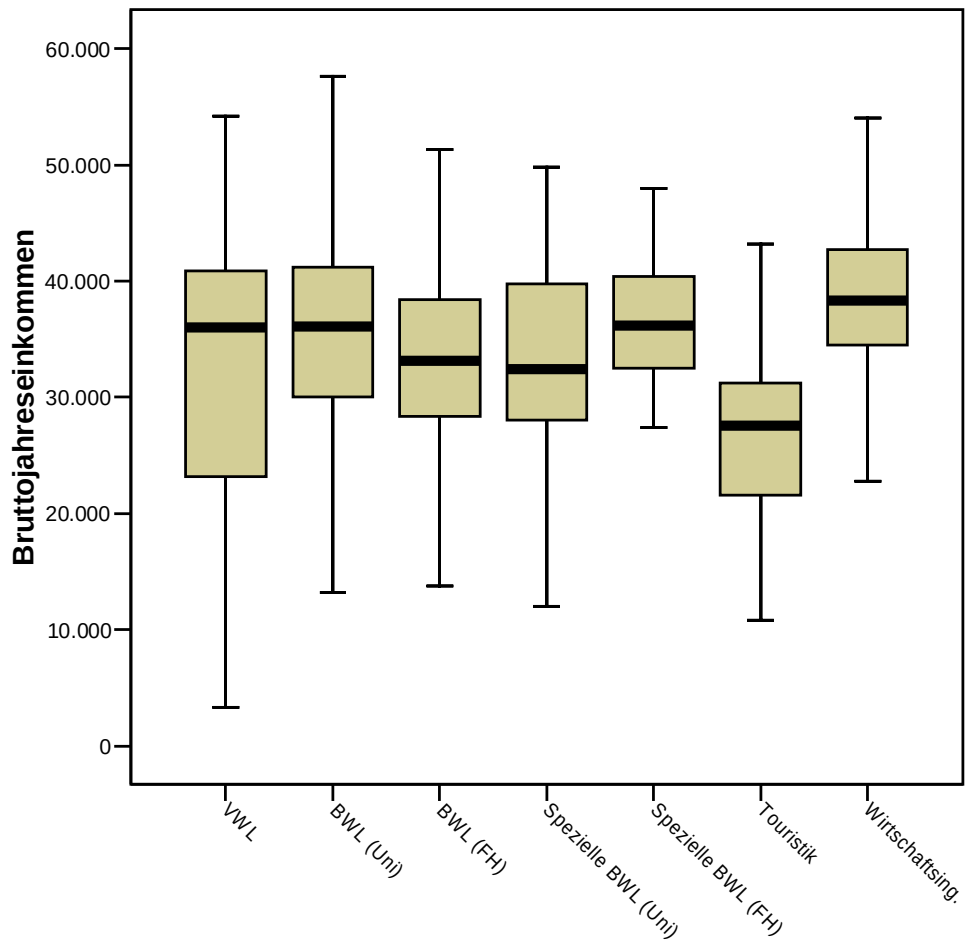
Beispiel: Studienanbschluss und Einstiegsgehalt



Quelle: Bayerisches Absolventenpanel, Befragung Abschlussjahrgang 2004
www.ihf.bayern.de

Beispiel: Studienanbschluss und Einstiegsgehalt (2)

4. Wirtschaftswissenschaften



Mögliche Strukturen

$$X(Studienfach) \rightarrow Y(Gehalt)$$

oder

$$Y(Gehalt) \rightarrow X(Studienfach)$$

oder

$$\begin{aligned} Z(IQ) &\rightarrow X(Studienfach) \\ Z(IQ) &\rightarrow Y(Gehalt) \end{aligned}$$

Interpretation von beobachteten Zusammenhängen

Beachte: Zusammenhänge können verschiedene Ursachen haben

- Kausalität (X hat Effekt auf Y)
- Kausalität in der anderen Richtung (Y hat Effekt auf X)
- Drittvariablen (Confounder), simultane Wirkung von Z auf X und Y
- Zufall und Selektion von Variablen

Notation

Es werden an jeder Einheit *gleichzeitig* mehrere Merkmale $X, Y, Z, \ldots$ erhoben:

$\Rightarrow$ mehrdimensionale oder multivariate Daten

Werte (x_i, y_i, z_i) der Merkmale (X, Y, Z)

Einheit i

Diskrete und gruppierte Merkmale

- Darstellung, Präsentation von (zwei) diskreten Merkmalen X und Y mit den Ausprägungen

$$\begin{array}{ll} a_1, \ldots, a_k & \text{für } X \\ b_1, \ldots, b_m & \text{für } Y \end{array}$$

- Skalenniveau von X, Y beliebig; X, Y können auch gruppierte metrische Merkmale sein.
- Benutzt wird nur das Nominalskalenniveau der Merkmale.

Beispiel: Arbeitslosigkeit

Zwei Merkmale:

- X Ausbildungsniveau mit den Kategorien
  - "keine Ausbildung",
  - "Lehre",
  - "fachspezifische Ausbildung"
  - "Hochschulabschluß"
- Y Dauer der Arbeitslosigkeit mit den Kategorien
  - "Kurzzeitarbeitslosigkeit" ($\leq$ 6 Monate),
  - "mittelfristige Arbeitslosigkeit" (7–12 Monate),
  - "Langzeitarbeitslosigkeit" ($\geq$ 12 Monate)

Arbeitslosigkeit

	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	
K A	86	19	18	123
Lehre	170	43	20	233
Fachspez	40	11	5	56
Hoch	28	4	3	35
	324	77	46	447

Ausbildungsspezifische Dauer der Arbeitslosigkeit für männliche Deutsche

Allgemeine Darstellung

Kontingenztafel der absoluten Häufigkeiten:

Eine $(k \times m)$-*Kontingenztafel der absoluten Häufigkeiten* besitzt die Form

	b_1	$\ldots$	b_m	
a_1	h_{11}	$\ldots$	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	$\ldots$	h_{2m}	$h_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	h_{k1}	$\ldots$	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	n

Notation

$h_{ij} = h(a_i, b_j)$ die absolute Häufigkeit der Kombination (a_i, b_j),

$h_{1\cdot}, \ldots, h_{k\cdot}$ die Randhäufigkeiten von X,

$h_{\cdot 1}, \ldots, h_{\cdot m}$ die Randhäufigkeiten von Y.

Die Kontingenztabelle gibt die gemeinsame Verteilung der Merkmale X und Y in absoluten Häufigkeiten wieder.

Kontingenztafel der relativen Häufigkeiten

Die $(k \times m)$-*Kontingenztafel der relativen Häufigkeiten* hat die Form

	b_1	$\ldots$	b_m	
a_1	f_{11}	$\ldots$	f_{1m}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	f_{k1}	$\ldots$	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	$\ldots$	$f_{\cdot m}$	1

Notation

$f_{ij} = h_{ij}/n$ die relative Häufigkeit der Kombination (a_i, b_j),

$f_{i\cdot} = \sum_{j=1}^{m} f_{ij} = h_{i\cdot}/n, \quad i = 1, \ldots, k$, die relativen Randhäufigkeiten zu X,

$f_{\cdot j} = \sum_{i=1}^{k} f_{ij} = h_{\cdot j}/n, \quad j = 1, \ldots, m$, die relativen Randhäufigkeiten zu Y

Die Kontingenztabelle gibt die gemeinsame Verteilung von X und Y wieder.

Bedingte Häufigkeiten

Zusammenhang zwischen X und Y aus *gemeinsamen* Häufigkeiten h_{ij} bzw. f_{ij} schwer ersichtlich.
Deshalb: Blick auf *bedingte* Häufigkeiten $\Rightarrow$ Verteilung des einen Merkmals für einen festgehaltenen Wert des zweiten Merkmals

Bedingte relative Häufigkeitsverteilung

Die *bedingte Häufigkeitsverteilung von Y unter der Bedingung* $X = a_i$, kurz $Y|X = a_i$, ist bestimmt durch

$$f_Y(b_1|a_i) = \frac{h_{i1}}{h_{i\cdot}}, \ldots, f_Y(b_m|a_i) = \frac{h_{im}}{h_{i\cdot}}.$$

Die *bedingte Häufigkeitsverteilung von X unter der Bedingung*

$Y = b_j$, kurz $X|Y = b_j$, ist bestimmt durch

$$f_X(a_1|b_j) = \frac{h_{1j}}{h_{\cdot j}}, \ldots, f_X(a_k|b_j) = \frac{h_{kj}}{h_{\cdot j}}.$$

Bemerkung

Wegen

$$\frac{h_{i1}}{h_{i\cdot}} = \frac{h_{i1}/n}{h_{i\cdot}/n} = \frac{f_{i1}}{f_{i\cdot}}$$

gilt auch

$$f_Y(b_1|a_i) = \frac{f_{i1}}{f_{i\cdot}}, \ldots, f_Y(b_m|a_i) = \frac{f_{im}}{f_{i\cdot}}$$

$$f_X(a_1|b_j) = \frac{f_{1j}}{f_{\cdot j}}, \ldots, f_X(a_k|b_j) = \frac{f_{kj}}{f_{\cdot j}}.$$

Merksatz:
Bedingte Häufigkeitsverteilungen werden durch Division der h_{ij} bzw. f_{ij} durch die entsprechende Zeilen- bzw. Spaltensumme gebildet.

Beispiel: Arbeitslosigkeit

$$f(\quad|a_i), \quad X = a_i, \;\; i = 1, \ldots, 4 \;\; \text{Ausbildungsniveau}$$

z.B. $\frac{86}{123} = 0.699$, $\frac{19}{123} = 0.154, \ldots$

$\frac{170}{233} = 0.730, \ldots$

usw.

Für festgehaltenes Ausbildungsniveau ($X = a_i$) erhält man die relative Verteilung über die Dauer der Arbeitslosigkeit durch die folgende Tabelle.

Bedingte Verteilung

	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	
Keine Ausb.	0.699	0.154	0.147	1
Lehre	0.730	0.184	0.086	1
Fachspez. Aus.	0.714	0.197	0.089	1
Hochschula.	0.800	0.114	0.086	1

- Bedingen auf das Ausbildungsniveau:
  $\Rightarrow$ Verteilung der Dauer der Arbeitslosigkeit für die Subpopulationen "Keine Ausbildung", "Lehre", usw.
- Verteilungen lassen sich nun miteinander vergleichen

$\Rightarrow$Nun ersichtlich: Relative Häufigkeit für Kurzzeitarbeitslosigkeit ist in der Subpopulation "Hochschulabschluß" mit 0.8 am größten.

Bedingte Verteilungen

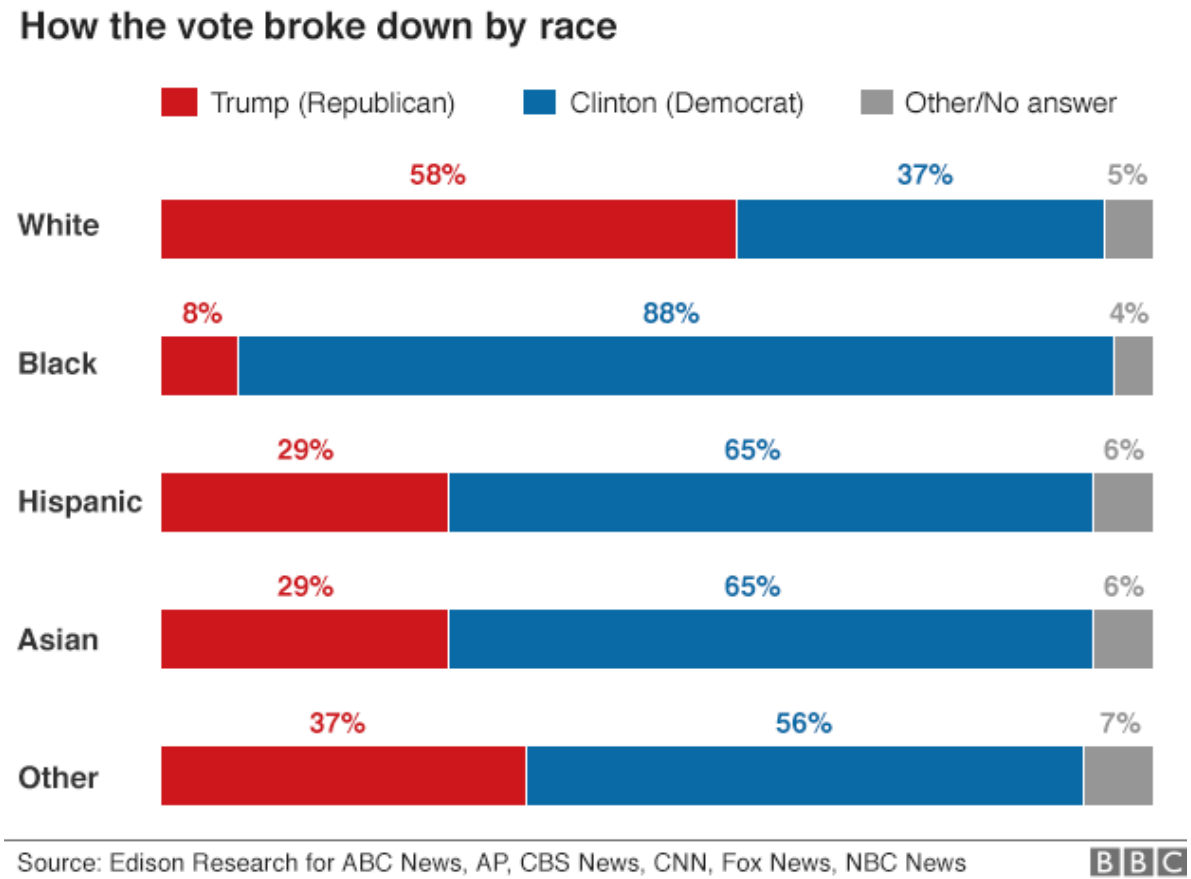
- Bei zwei Merkmalen X und Y kann man die bedingte Verteilung von $X|Y$ und auch von $Y|X$ berechnen.
- Die Wahl hängt von der inhaltlichen Fragestellung ab.
- Typischerweise betrachtet man die bedingte Verteilung $Y|X$, wenn Y eine Zielgröße und X eine Einflussgröße ist. Die Struktur des Zusammenhangs ist dann $X \rightarrow Y$.
- $Y|X$ hilft die Wirkung von X auf Y zu verstehen oder auch Y mit Hilfe von X zu prognostizieren

Beispiel: US Wahl

- Y : Wahlentscheidung
- X1 : Geschlecht , X2 : Hautfarbe
- Y wird durch X1 und X2 beeinflusst (prognostiziert). Betrachte daher $Y|X1$ und $Y|X2$
- $X2|Y$ beantwortet z.B. die Frage: „Wie hoch ist der Anteil der Schwarzen bei den Wählern von Trump? “. Das ist häufig nicht sinnvoll und manchmal auch irreführend.

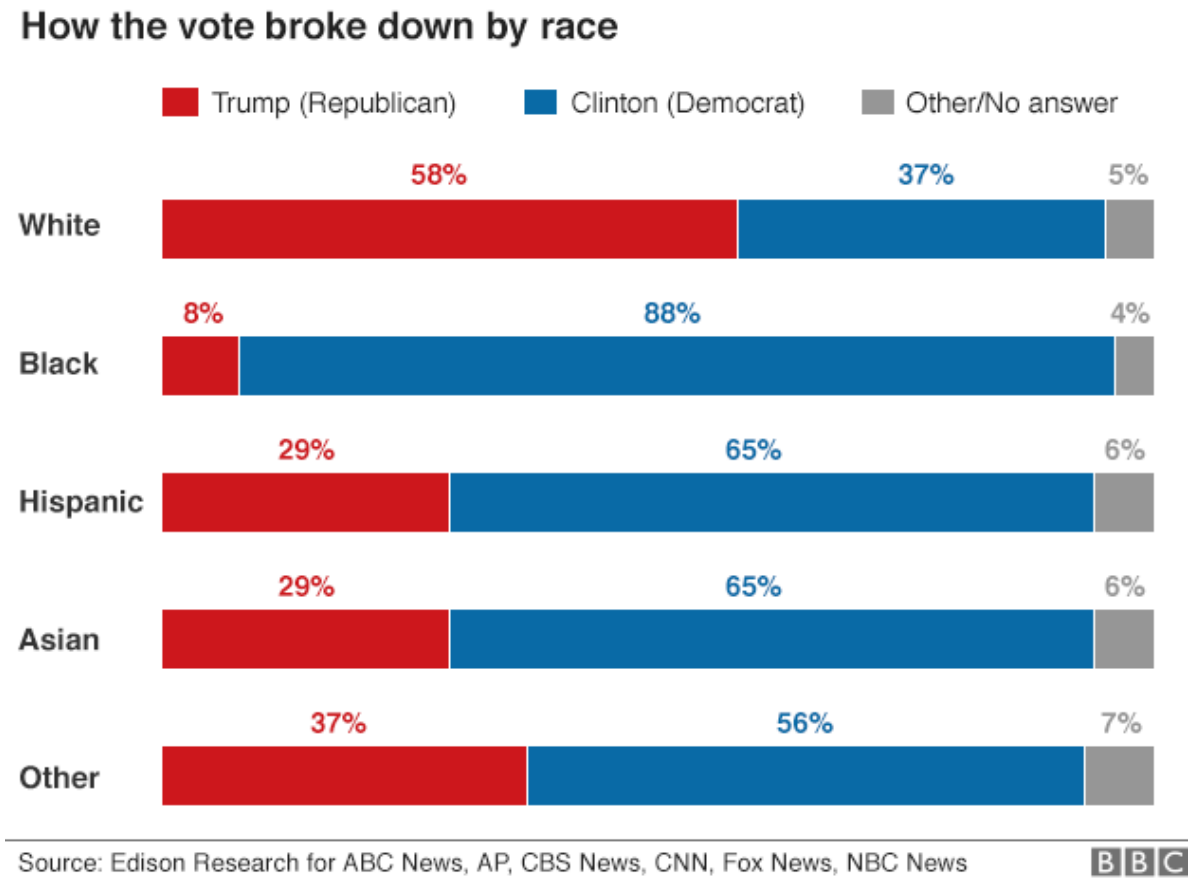
Darstellung der bedingten Verteilung

Balkendiagramme für binäre Zielgrößen und für nicht geordnete Zielgrößen



Darstellung der bedingten Verteilung

Gestapelte Balkendiagramme für binäre ordinale Zielgrößen



Zusammenhangsanalyse in Kontingenztabellen

Bisher: Tabellarische / grafische Präsentation
Jetzt: Maßzahlen für Stärke des Zusammenhangs zwischen X und Y.

Chancen und relative Chancen

- Zunächst 2×2 - Kontingenztafel

		Y		
		1	2	
X	1	h_{11}	h_{12}	$h_{1.}$
	2	h_{21}	h_{22}	$h_{2.}$
		$h_{.1}$	$h_{.2}$	n

Chancen („Odds“)

- Wir betrachten die Merkmale X und Y zunächst asymmetrisch: Die Ausprägungen von X definieren (hier 2) Subpopulationen, Y ist das interessierende binäre Merkmal in diesen Subpopulationen
- Unter einer Chance (“odds”) versteht man nun das Verhältnis zwischen dem Auftreten von $Y = 1$ und $Y = 2$ in einer Subpopulation $X = a_i$.

Odds Ratio

- Die (empirische) bedingte Chance für festes $X = a_i$ ist bestimmt durch

$$\gamma(1, 2|X = a_i) = \frac{h_{i1}}{h_{i2}}.$$

- Ein sehr einfaches Zusammenhangsmaß stellen die empirischen relativen Chancen (Odds Ratio) dar, die gegeben sind durch

$$\gamma(1, 2|X = 1, X = 2) = \frac{\gamma(1, 2|X = 1)}{\gamma(1, 2|X = 2)} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{21}h_{12}},$$

d.h. $\gamma(1, 2|X = 1, X = 2)$ ist das Verhältnis zwischen den Chancen der 1. Population ($X = 1$, 1. Zeile) zu den Chancen der 2. Population ($X = 2$, 2. Zeile).

Beispiel: Dauer der Arbeitslosigkeit

Beschränkt man sich jeweils nur auf zwei Kategorien der Merkmale Ausbildungsniveau und Dauer der Arbeitslosigkeit, erhält man beispielsweise die Tabelle

	Kurzzeit-arbeitslosigkeit	Mittel- und langfristige Arbeitslosigkeit
Fachspezifische Ausbildung	40	16
Hochschulabschluß	28	7

Daraus ergibt sich für Personen mit fachspezifischer Ausbildung die "Chance", kurzzeitig arbeitslos zu sein, im Verhältnis dazu, mittel –oder längerfristig arbeitslos zu sein, durch

$$\gamma(1,2|\text{fachspezifisch}) = \frac{40}{16} = 2.5.$$

Für Arbeitslose mit Hochschulabschluß erhält man

$$\gamma(1, 2|\text{Hochschulabschluß}) = \frac{28}{7} = 4.$$

Für fachspezifische Ausbildung stehen die "Chancen" somit $5 : 2$, für Arbeitslose mit Hochschulabschluß $4 : 1$.

Man erhält für fachspezifische Ausbildung und Hochschulabschluß die relativen Chancen (Odds Ratio)

$$\gamma(1, 2|\text{fachsp. Ausbildung, Hochschule}) = \frac{2.5}{4} = 0.625 = \frac{40 \cdot 7}{16 \cdot 28}$$

Interpretation „Odds Ratio“

- Wegen der spezifischen Form
  $\gamma(1,2|X=1,X=2) = (h_{11}h_{22})/(h_{21}h_{12})$ werden die relativen Chancen auch als Kreuzproduktverhältnis bezeichnet. Es gilt

  $\gamma = 1$ Chancen in beiden Populationen gleich

  $\gamma > 1$ Chancen in Population $X = 1$ besser als in Population $X = 2$

  $\gamma < 1$ Chancen in Population $X = 1$ schlechter als in Population $X = 2$.

- Die relativen Chancen geben somit an, welche der Populationen die besseren Chancen besitzen und um wieviel besser diese Chancen sind.

Symmetrie

- Für die Kontingenztafel

$$\boxed{\begin{array}{cc} h_{11} & h_{12} \\ h_{21} & h_{22} \end{array}}$$

ist das *Kreuzproduktverhältnis* (*relative Chance* oder *Odds Ratio*) bestimmt durch

$$\gamma = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{21}h_{12}}.$$

- Die asymmetrische Betrachtung der Merkmale X und Y wird aufgehoben

Fall - Kontroll - Studien

Beispiel: Morbus Alzheimer und Genetik

Genetik	ApoE3	ApoE4	Summe
Fall	593	620	1213
Kontrolle	2258	803	3061
	2851	1423	4274

$$OR = \frac{593/620}{2258/803} = 0.34$$

$\Rightarrow$ Chance für ApoE3 bei Fällen um den Faktor 3 niedriger als bei Kontrollen

$\overset{?}{\Rightarrow}$ ApoE4 Risiko-Faktor für Morbus Alzheimer

Fall - Kontroll - Studien

Zentrale Argumentation:

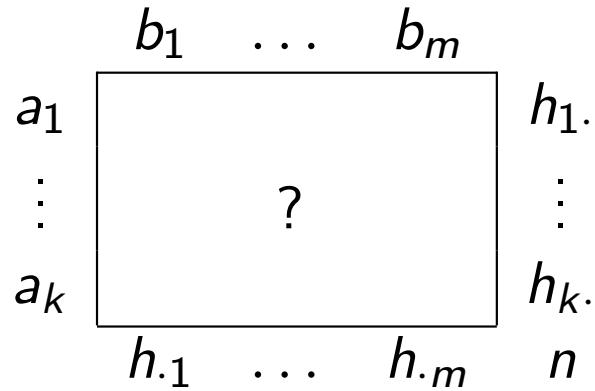
Odds Ratio ist symmetrisches Maß
d.h. Chancenverhältnis für Auftreten von ApoE4 bei Kontrolle zu Auftreten von ApoE4 bei Fällen

Person ist krank bei ApoE3
zu
Person ist krank bei ApoE4

$\Rightarrow$ Interpretation als **Risikofaktor** zulässig

Kontingenz- und χ^2-Koeffizient

Ausgangspunkt: Wie sollten gemeinsame Häufigkeiten $\tilde{h}_{ij}$ bzw. $\tilde{f}_{ij}$ verteilt sein, damit - bei vorgegebenen Randverteilungen - die Merkmale X und Y als „empirisch unabhängig“ angesehen werden können?



Empirische Unabhängigkeit

Idee: X und Y „empirisch unabhängig“

$\Leftrightarrow$ Bedingte relative Häufigkeiten

$$f_Y(b_1|a_i), \ldots, f_Y(b_m|a_i), \quad i = 1, \ldots, k$$

sind in jeder Schicht $X = a_i$ identisch, d.h. unabhängig von a_i.
Formal:

$$\begin{aligned}
f_Y(b_1|a_1) &= f(b_1), \ldots, f_Y(b_m|a_1) = f_Y(b_m) \\
f_Y(b_1|a_2) &= f(b_1), \ldots, f_Y(b_m|a_2) = f_Y(b_m) \\
\vdots &= \vdots \\
f_Y(b_1|a_k) &= f(b_1), \ldots, f_Y(b_m|a_k) = f_Y(b_m)
\end{aligned}$$

Kunstbeispiel:

	b_1	b_2	b_3	
a_1	10	20	30	60
a_2	20	40	60	120
	30	60	90	180

$$
\begin{aligned}
f_Y(b_1|a_1) &= f_Y(b_1|a_2) = f_Y(b_1) = \frac{1}{6} \\
f_Y(b_2|a_1) &= f_Y(b_2|a_2) = f_Y(b_2) = \frac{1}{3} \\
f_Y(b_3|a_1) &= f_Y(b_3|a_2) = f_Y(b_3) = \frac{1}{2}
\end{aligned}
$$

Wie sehen die "erwarteten" (absoluten und relativen) Häufigkeiten $\tilde{h}_{ij}$ und $\tilde{f}_{ij}$ also aus?

$$f_Y(b_1|a_i) = f(b_1), \ldots, f_Y(b_m|a_i) = f_Y(b_m), \; i = 1, \ldots, k$$

$$\Leftrightarrow \frac{\tilde{h}_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n}$$

$$\Leftrightarrow \tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$$

$$\Leftrightarrow \tilde{f}_{ij} = f_{i\cdot} f_{\cdot j}$$

„Unabhängigkeitstabelle“

Idee: Vergleiche für jede Zelle (i, j) $\tilde{h}_{ij}$ mit tatsächlich beobachteten h_{ij}

$\Rightarrow$ χ^2-**Koeffizient**

Der χ^2-*Koeffizient* ist bestimmt durch

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{h_{i\cdot} h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot} h_{\cdot j}}{n}} = n \sum_{i} \sum_{j} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}$$

Eigenschaften des χ^2-Koeffizienten:

- $\chi^2 \in [0, \infty)$
- $\chi^2 = 0 \Leftrightarrow X$ und Y „empirisch unabhängig“
- χ^2 groß $\Leftrightarrow$ starker Zusammenhang
- χ^2 klein $\Leftrightarrow$ schwacher Zusammenhang
- Nachteil: χ^2 hängt vom Stichprobenumfang n und von der Dimension der Tafel ab.

Kontingenzkoeffizient und korrigierter Kontingenzkoeffizient

Weitere Normierung $\Rightarrow$ Kontingenzkoeffizient

Der Kontingenzkoeffizient ist bestimmt durch

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

und besitzt den Wertebereich $K \in \left[0, \sqrt{\frac{M-1}{M}}\right]$, wobei $M = \min\{k, m\}$.

Der korrigierte Kontingenzkoeffizient ergibt sich durch

$$K^* = K / \sqrt{\frac{M-1}{M}}$$

mit dem Wertebereich $K^* \in [0, 1]$.

Eigenschaften des Kontingenzkoeffizienten

- Es wird nur die *Stärke* des Zusammenhangs gemessen, nicht die Richtung wie beim Odds Ratio.
- Vorsicht ist geboten bei einem Vergleich von Kontingenztafeln mit stark unterschiedlichen Stichprobenumfängen, da χ^2 mit wachsendem Stichprobenumfang wächst, beispielsweise führte eine Verzehnfachung von h_{ij} und $\tilde{h}_{ij}$ zu zehnfachem χ^2.
- Sämtliche Maße benutzen nur das Nominalskalenniveau von X und Y.

Beispiel: Sonntagsfrage

Für die Kontingenztafel aus Geschlecht und Parteipräferenz für das Beispiel der Sonntagsfrage erhält man die in der folgenden Tabelle wiedergegebenen zu erwartenden Häufigkeiten $\tilde{h}_{ij}$.

	CDU/CSU	SPD	FDP	Grüne	Rest	
Männer	160.73 (144)	139.24 (153)	21.96 (17)	35.51 (26)	77.56 (95)	435
Frauen	183.27 (200)	158.76 (145)	25.04 (30)	40.49 (50)	88.44 (71)	496
	344	298	47	76	166	

Zu erwartende Häufigkeiten $\tilde{h}_{ij}$ und tatsächliche Häufigkeiten h_{ij} (in Klammern)

Interpretation:

- Wenn Geschlecht und Parteipräferenz keinen Zusammenhang aufweisen, wären 160.73 die CDU/CSU präferierende Männer zu erwarten.
- Tatsächlich wurden aber nur 144 beobachtet.

$$\Rightarrow \quad \chi^2\text{-Wert von } 20.065,$$

$$K = 0.145,$$

$$K^* = 0.205$$

Spezialfall: (2×2)-Tafel

Für den Spezialfall einer (2×2)-Tafel

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	

erhält man χ^2 aus

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Beispiel: Arbeitslosigkeit

Aus der Kontingenztafel

	Mittelfristige Arbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

erhält man also unmittelbar

$$\chi^2 = \frac{100(19 \cdot 20 - 18 \cdot 43)^2}{37 \cdot 63 \cdot 62 \cdot 38} = 2.826$$

und $K = 0.165,\ K^* = 0.234.$

Mehrdimensionale Kontingenztabellen

Beispiel: Fluglinien und Verspätung

- Mehrere diskrete Merkmale: Fluglinie, Ort, Verpätung (Ja/Nein)
- Darstellung durch geeignete bedingte und marginale Verteilungen
- Berechnung von Odds-Ratio zweier Merkmale bedingt auf ein drittes Merkmal
- Graphische Darstellung durch Mosaik-Plot

Verspätung von Flügen

Flüge mit (sp) und ohne Verspätung (ok)

Fluglinie	AW	AA	Summe
ok	6438	3274	9712
Verspätung	787	501	1288
Summe	7225	3775	11000

Welche Fluglinie nehmen Sie ?

Verspätung von Flügen

Flüge mit (sp) und ohne Verspätung (ok)

Fluglinie	AW	AA	Summe
ok	6438	3274	9712
Verspätung	787	501	1288
Summe	7225	3775	11000

Fluglinie	AW	AA
ok	0.89	0.87
sp	0.11	0.13
Summe	1	1

Wie entscheiden Sie ?

Verspätung von Flügen

Sie starten in LA :

Fluglinie	AW	AA
ok	694	497
sp	117	62

	AW	AA
ok	0.86	0.89
sp	0.14	0.11

Wie entscheiden Sie jetzt ?

Verspätung von Flügen

Sie starten in San Francisco :

	AW	AA
ok	320	503
sp	129	102

	AW	AA
ok	0.71	0.83
sp	0.29	0.17

Wie entscheiden Sie jetzt ?

Verspätung von Flügen

Sie starten in Seattle

	AW	AA
ok	201	1841
sp	61	305

	AW	AA
ok	0.77	0.86
sp	0.23	0.14

Wie entscheiden Sie jetzt ?

Verspätung von Flügen

Sie starten in Phoenix

	AW	AA
ok	4840	221
sp	415	12

	AW	AA
ok	0.92	0.95
sp	0.08	0.05

Wie entscheiden Sie jetzt ?

Verspätung von Flügen

Sie starten San Diego

	AW	AA
ok	383	212
sp	65	20

	AW	AA
ok	0.85	0.91
sp	0.15	0.09

Wie entscheiden Sie jetzt ?

Das Simpsonsche Paradoxon

Betrachte beide Fluglinien

	SanF	Seattle	LA	San Diego	Phoenix
ok	823	2042	1191	595	5061
sp	231	366	179	85	427

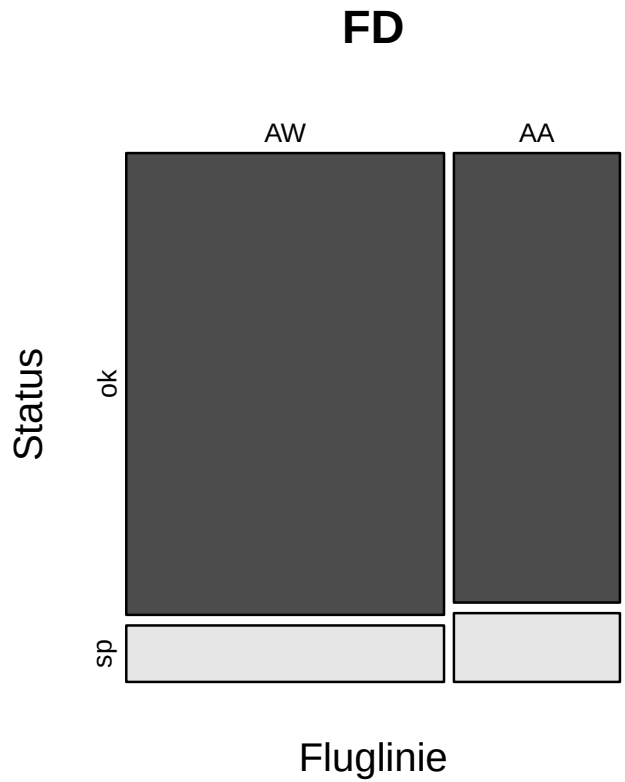
	SanF	Seattle	LA	San Diego	Phoenix
ok	0.78	0.85	0.87	0.88	0.92
sp	0.22	0.15	0.13	0.12	0.08

AW startet häufiger von Phoenix!!!

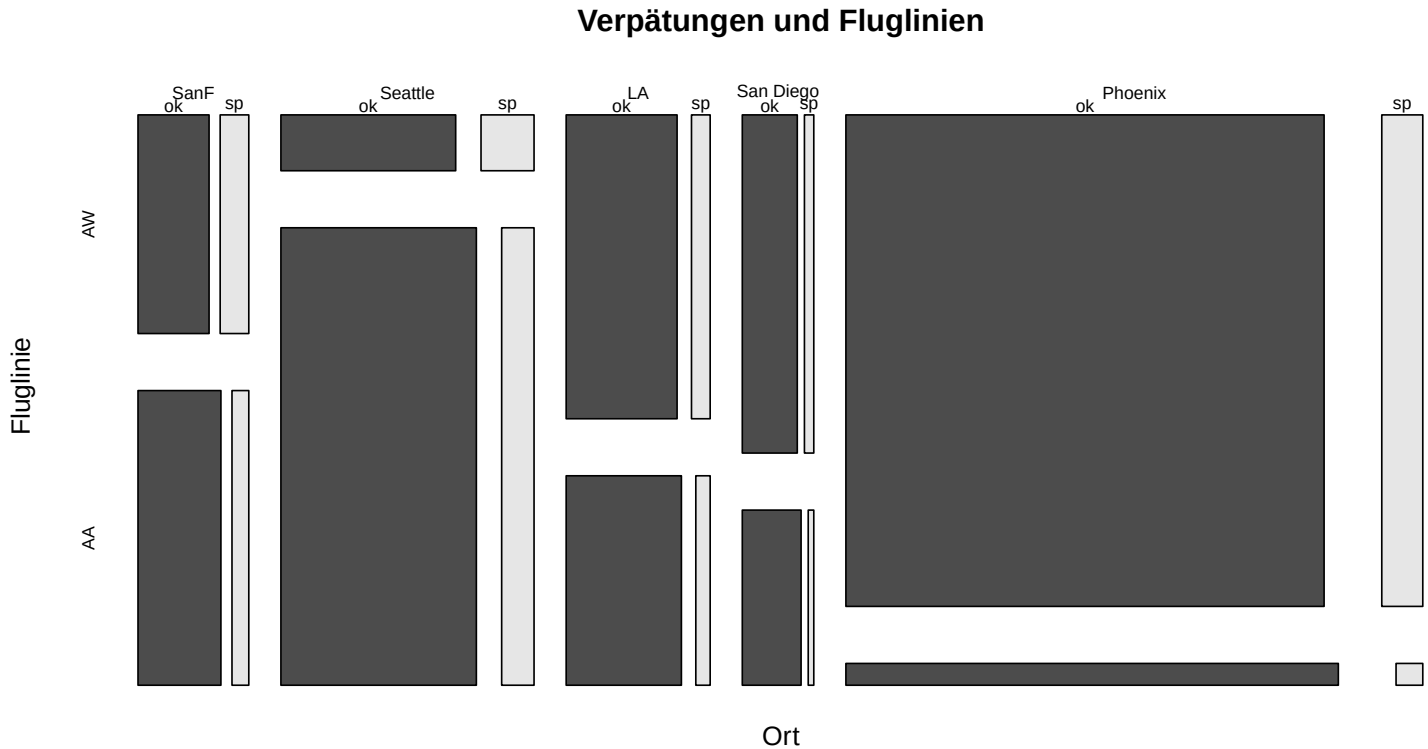
Mosaik-Plot

- Flächentreue Darstellung von Häufigkeiten
- Aufteilung schrittweise
- Zuerst Einflussgröße, zum Schluss nach Zielgröße aufteilen
- Gut geeignet für mehrkategoriale ordinale Daten
- Auch für höhere Dimensionen geeignet

Beispiel: Fluglinien



Beispiel: Fluglinien nach Ort



Ergänzung: Relatives Risiko

Gegeben sei eine 2×2 - Kontingenztafel

		Y		
		1	2	
X	1	h_{11}	h_{12}	$h_{1\cdot}$
	2	h_{21}	h_{22}	$h_{2\cdot}$
		$h_{\cdot 1}$	$h_{\cdot 2}$	n

X: Gruppe Y: Zielgröße, z.B. Krankheit, Insolvenz

- Odds Ratio (Chancenverhältnis)

$$\frac{h_{11}/h_{12}}{h_{21}/h_{22}}$$

- Relatives Risiko

$$\frac{h_{11}/h_{1\cdot}}{h_{21}/h_{2\cdot}}$$

6 Zusammenhänge von metrischen Variablen

Zusammenhänge zwischen metrischen Merkmalen

Darstellung des Zusammenhangs, Korrelation und Regression

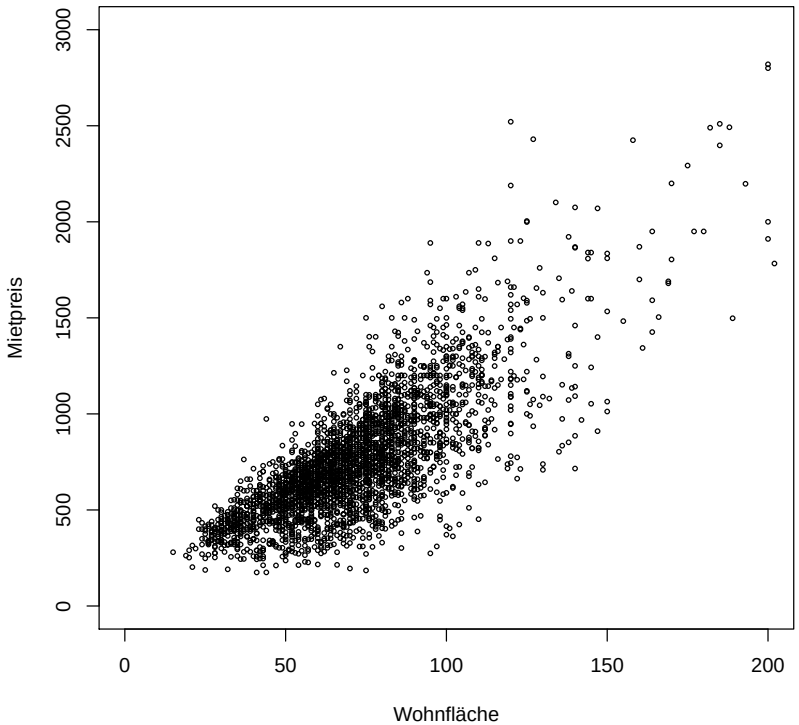
Daten liegen zu zwei metrischen Merkmalen vor:
Datenpaare $(x_i, y_i), \quad i = 1, \ldots, n$

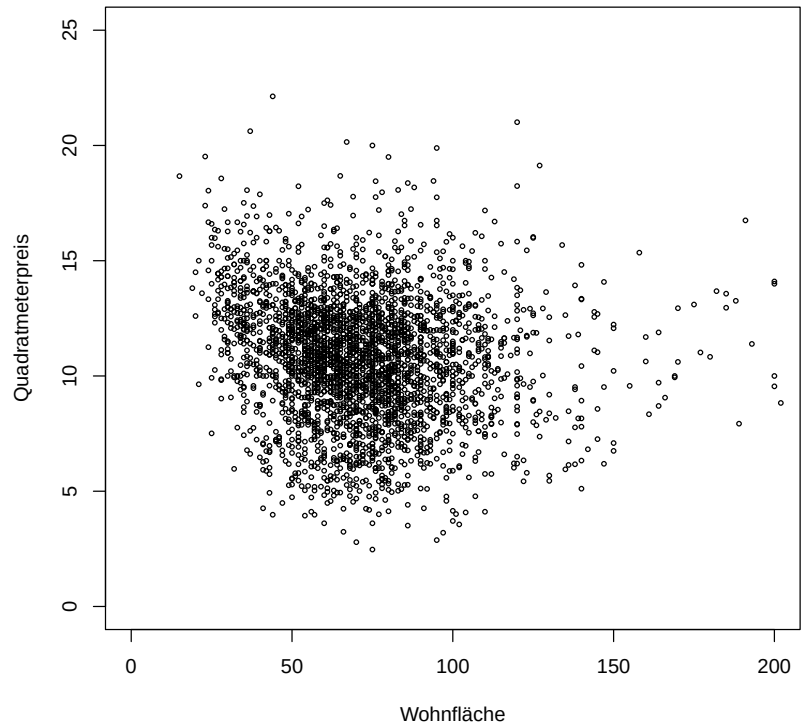
Beispiel:
x: Wohnfläche y: Quadratmeterpreis

Frage:
Gibt es einen Zusammenhang zwischen diesen Merkmalen?
Wie lässt sich dieser Zusammenhang beschreiben?

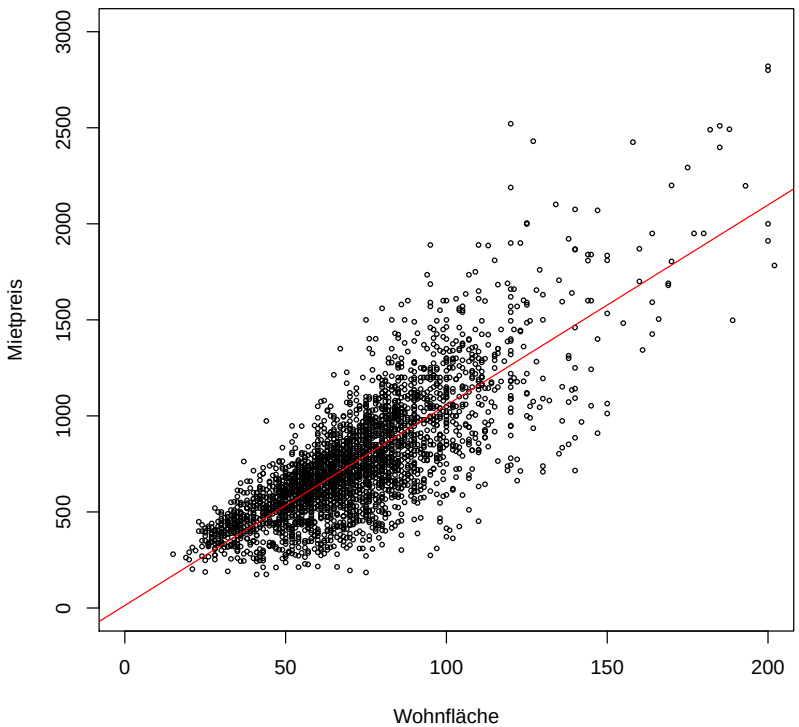
Einfachste graphische Darstellung: Streudiagramm.
Die Datenpaare entsprechen Punkten in der Ebene (Punktwolke)

Beispiel: Streudiagramm (Mietspiegel 2015)

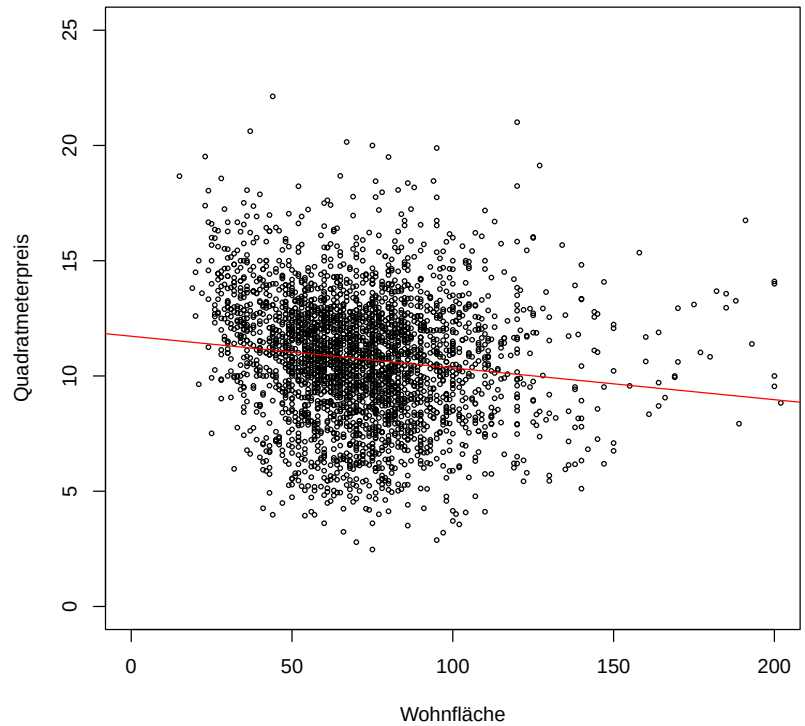




Beispiel: Streudiagramm (Mietspiegel 2015)

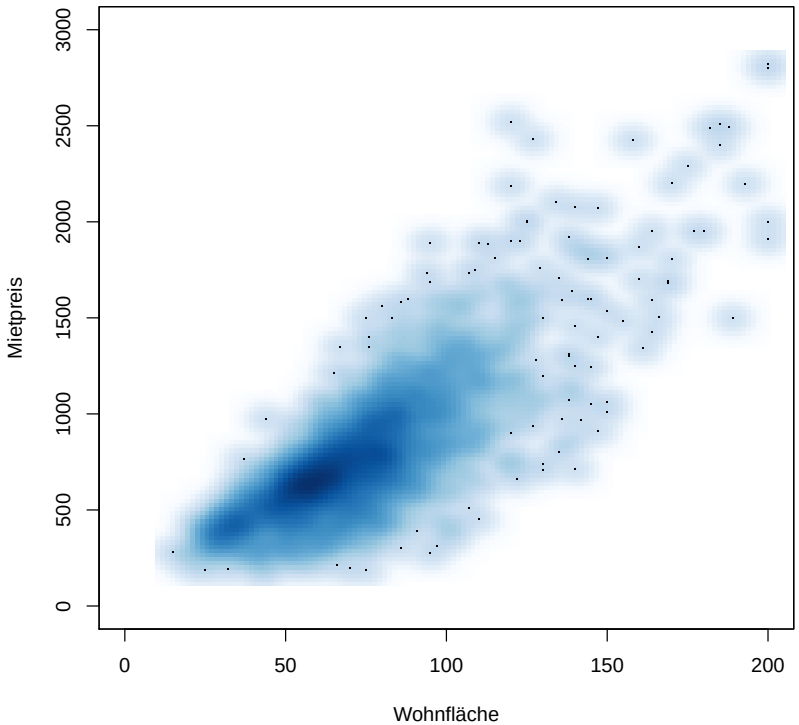


- Zusammenhang erkennbar

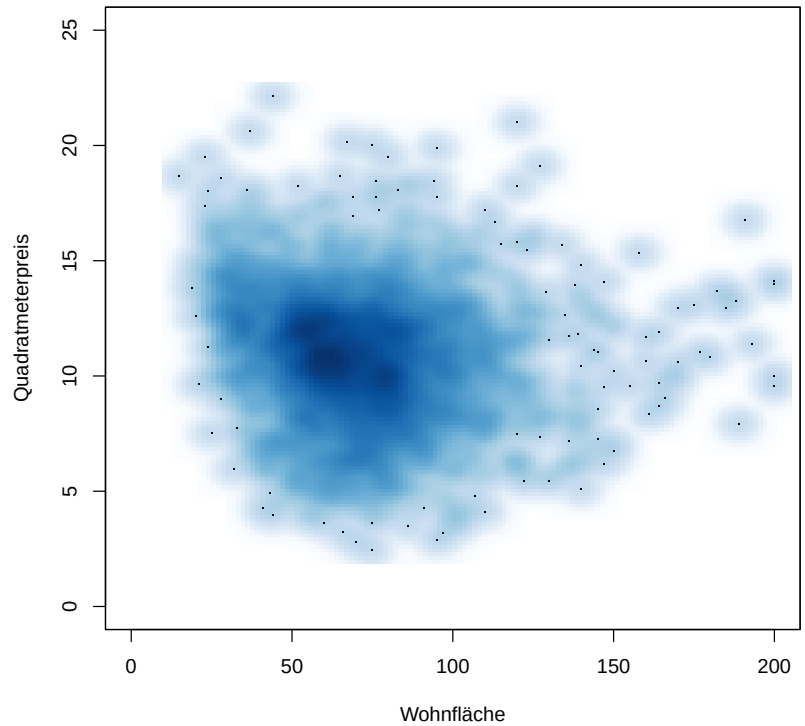


- Kaum ein Zusammenhang zu sehen

Beispiel: Streudiagramm (Mietspiegel 2015)



- Zusammenhang erkennbar



- Kaum ein Zusammenhang zu sehen

Kovarianz

Maß für den Zusammenhang der beiden Merkmale:

Daten: $(x_i, y_i), \quad i = 1, \ldots, n$

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

Beachte:

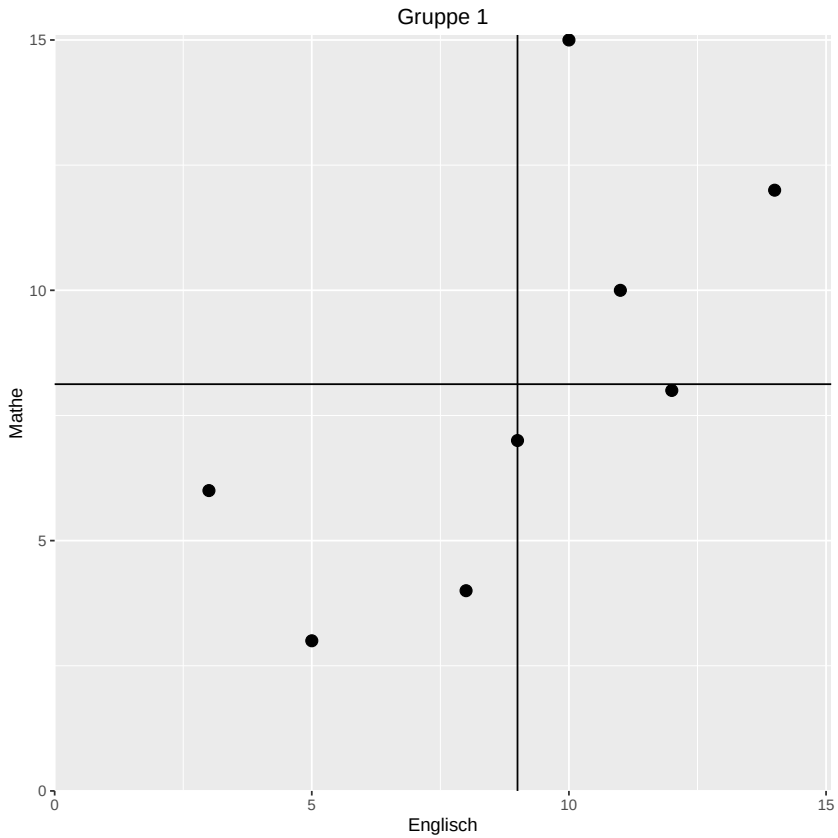
- Summand i positiv, falls x_i und y_i relativ zum Mittelwert das gleiche Vorzeichen haben.
- Für s_{xx} ergibt sich die Varianz von X.
- Die Kovarianz hängt sowohl von der Streuung als auch von dem Zusammenhang der beiden Merkmale ab.

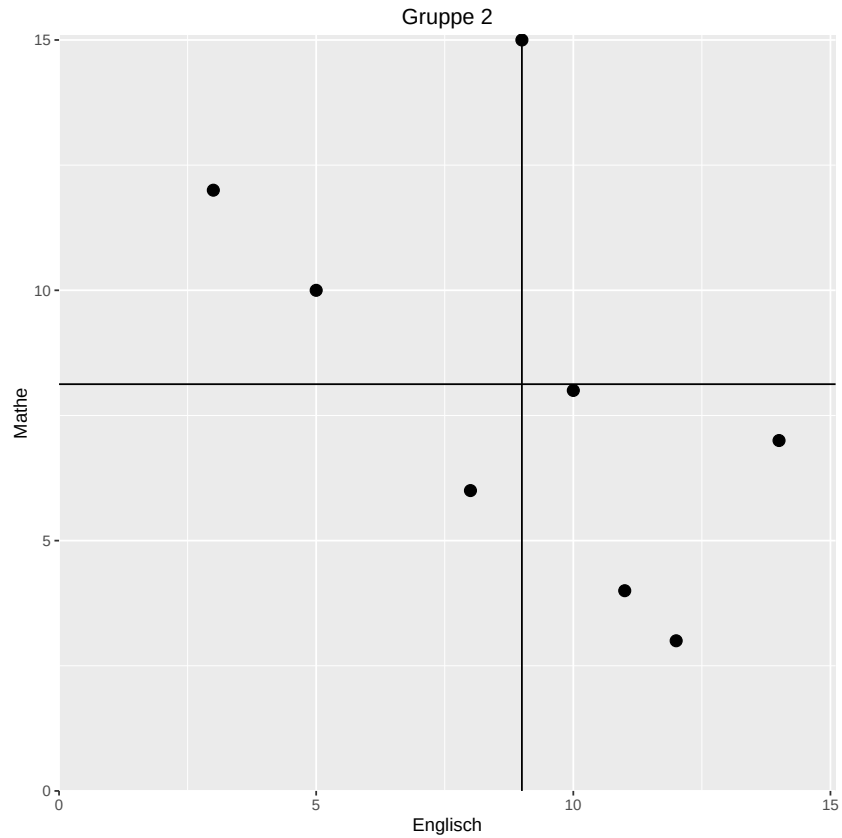
Beispiel: Kovarianz

Punkte in Englisch und Mathematik

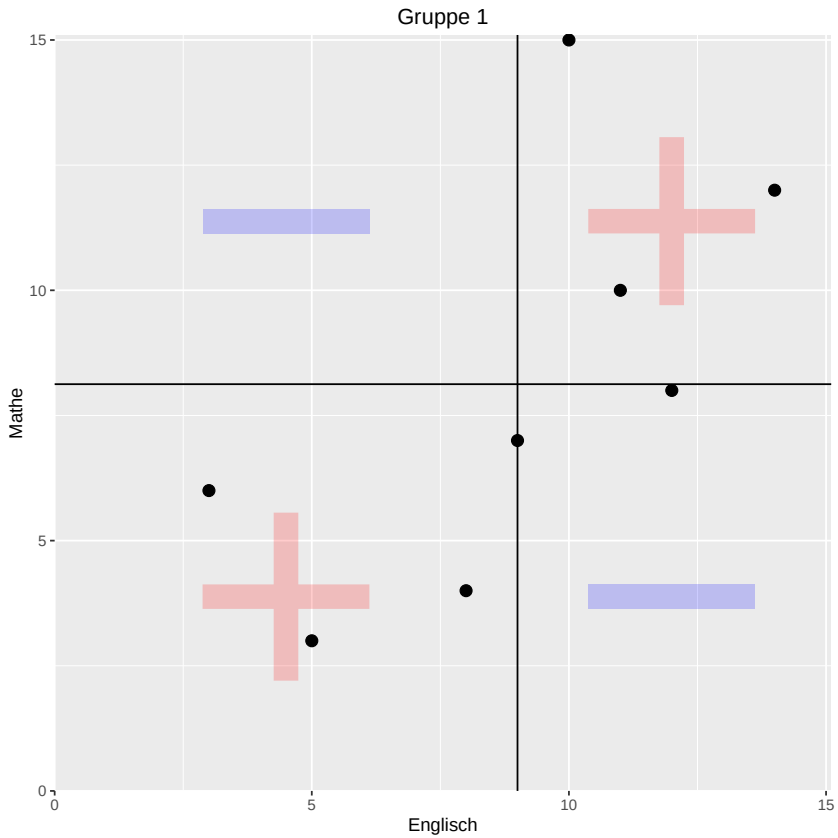
Schüler	Gruppe 1		Gruppe 2	
	Englisch	Mathe	Englisch	Mathe
1	14	12	10	8
2	9	7	8	6
3	5	3	3	12
4	3	6	5	10
5	11	10	14	7
6	8	4	9	15
7	10	15	11	4
8	12	8	12	3
Mittelwert	9.0	8.1	9.0	8.1
Standardabweichung	3.6	4.1	3.6	4.1

Beispiel: Kovarianz

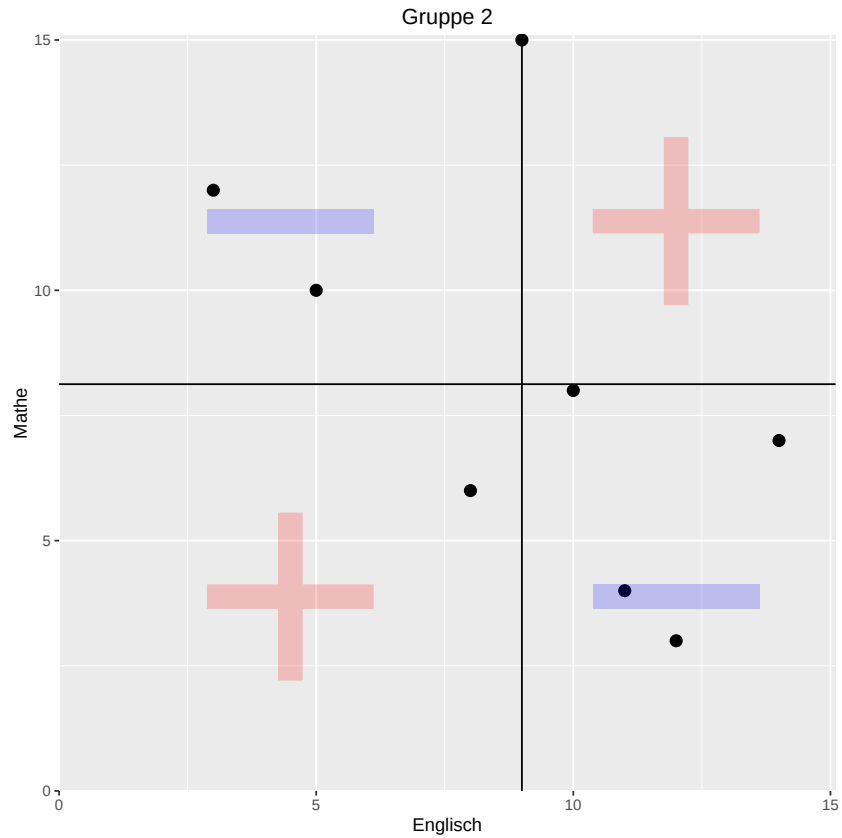




Beispiel: Kovarianz



- Kovarianz: 9.57



- Kovarianz: -8.29

Bravais-Pearson-Korrelationskoeffizient

Der Bravais-Pearson-Korrelationskoeffizient ergibt sich aus den Daten $(x_i, y_i), i = 1, \ldots, n$ durch

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{S_{xy}}{S_x S_y}$$

Wertebereich: $-1 \leq r \leq 1$

$r > 0$	positive Korrelation Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
$r < 0$	negative Korrelation Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
$r = 0$	keine Korrelation, kein linearer Zusammenhang

Punkte in Englisch und Mathematik

Gruppe 1:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{9.57}{3.641} = 0.65$$

Gruppe 2:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-8.29}{3.6 \cdot 4.1} = -0.56$$

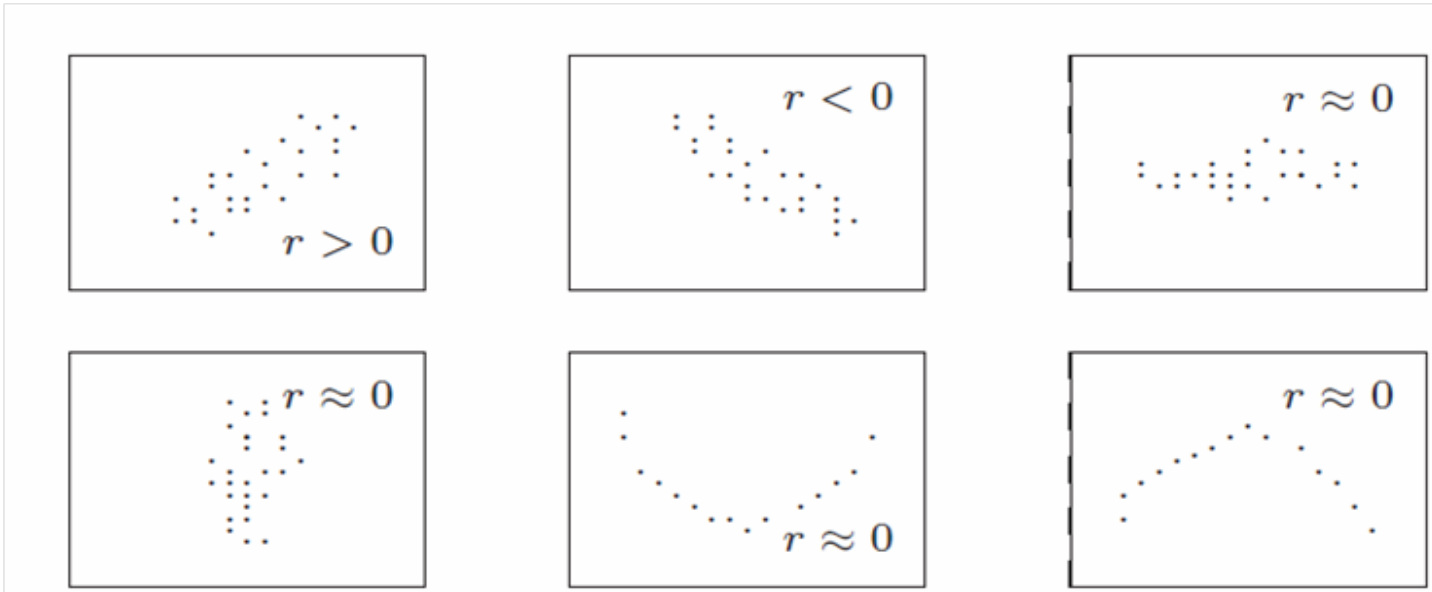
Gruppe 1: positiver linearer Zusammenhang
Gruppe 2: negativer linearer Zusammenhang

Eigenschaften des Korrelationskoeffizienten

- Maß für den linearen Zusammenhang
- Ändert sich nicht bei linearen Transformationen
- Symmetrisch (Korrelation zwischen x und y = Korrelation zwischen y und x)
- Positive Korrelation bedeutet: Je größer x, desto größer im Durchschnitt y
- Korrelation = +1 oder -1, falls die Punkte genau auf einer Geraden liegen
- Korrelation = 0 bedeutet keinen linearen Zusammenhang, aber nicht Unabhängigkeit
- Korrelation empfindlich gegenüber Ausreißern

Eigenschaften von r

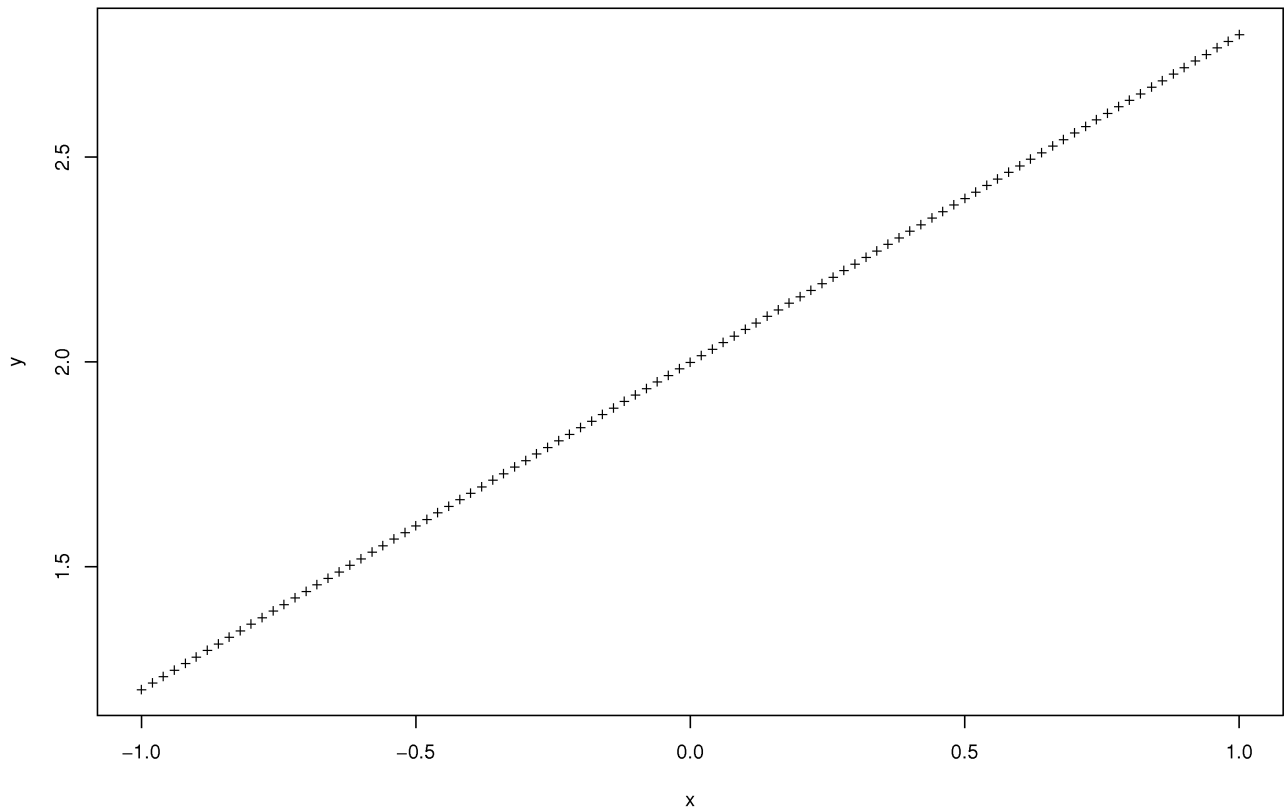
- r misst Stärke des *linearen* Zusammenhangs.



Punktkonfigurationen und Korrelationskoeffizienten (qualitativ)

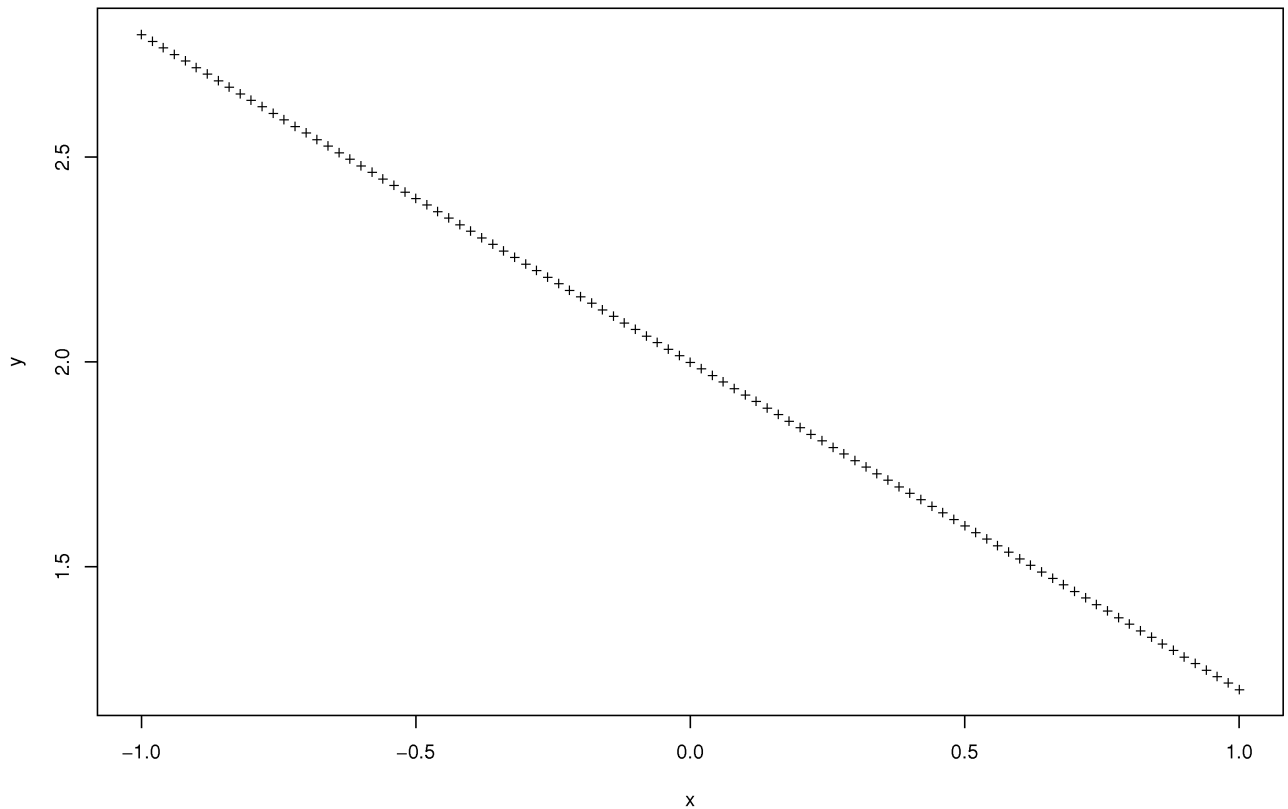
Einige Beispiele von Zusammenhängen

Beispiel 1: Lineare (unverrauschte) Funktion, $y = 0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



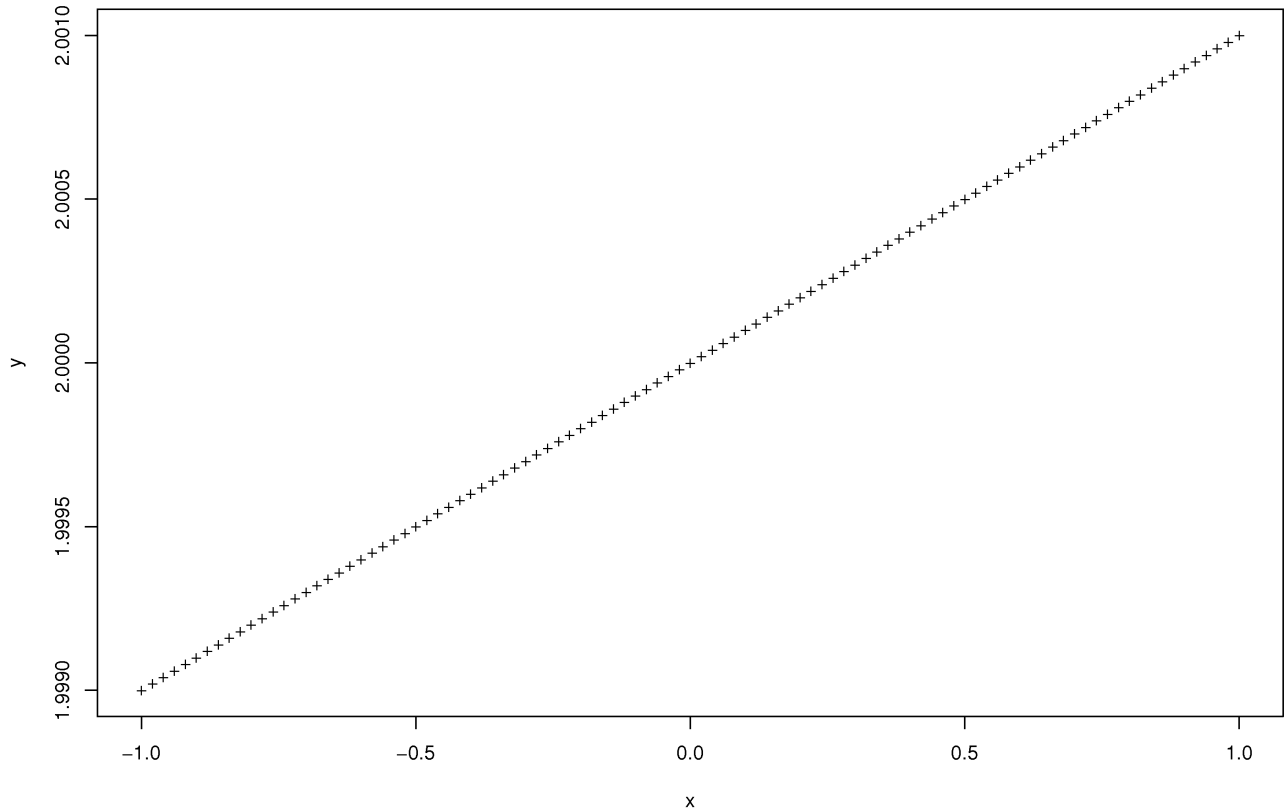
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 2: Lineare (unverrauschte) Funktion, $y = -0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



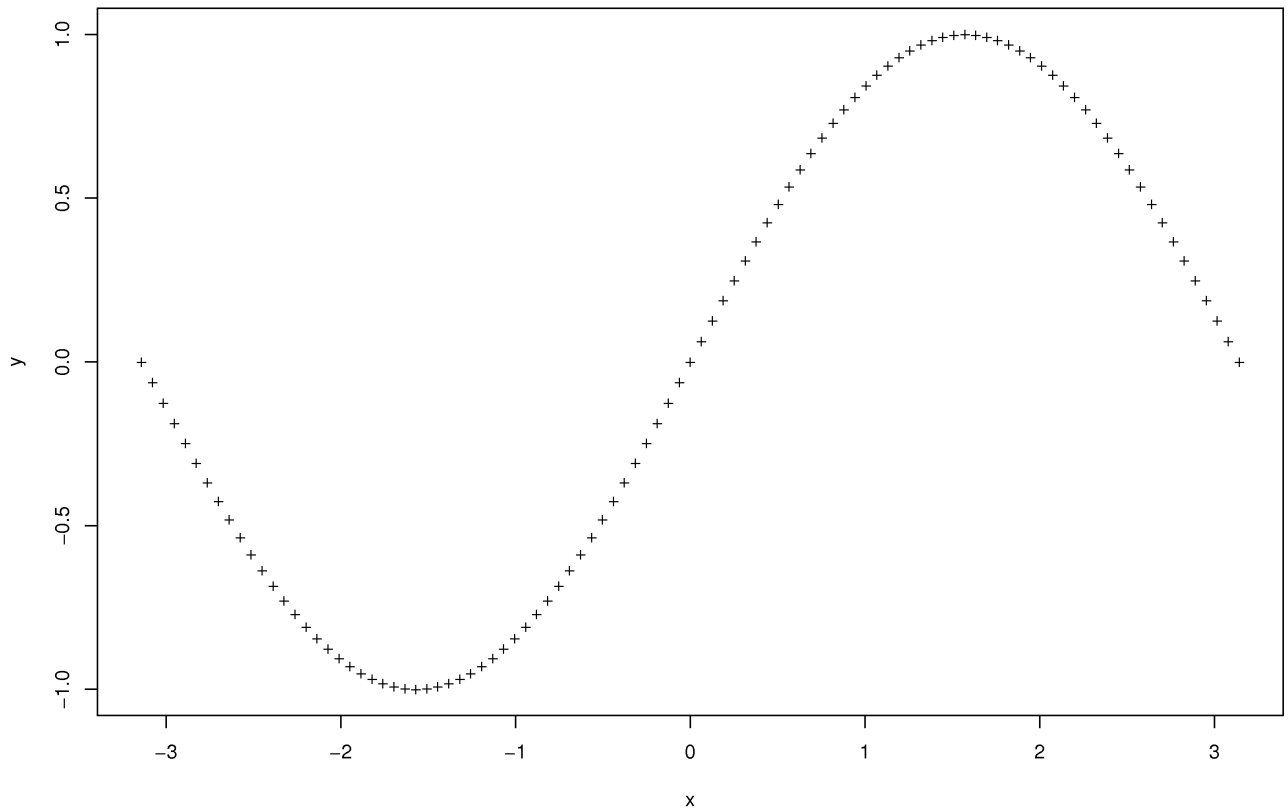
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 3: Lineare (unverrauschte) Funktion, $y = 0.001x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



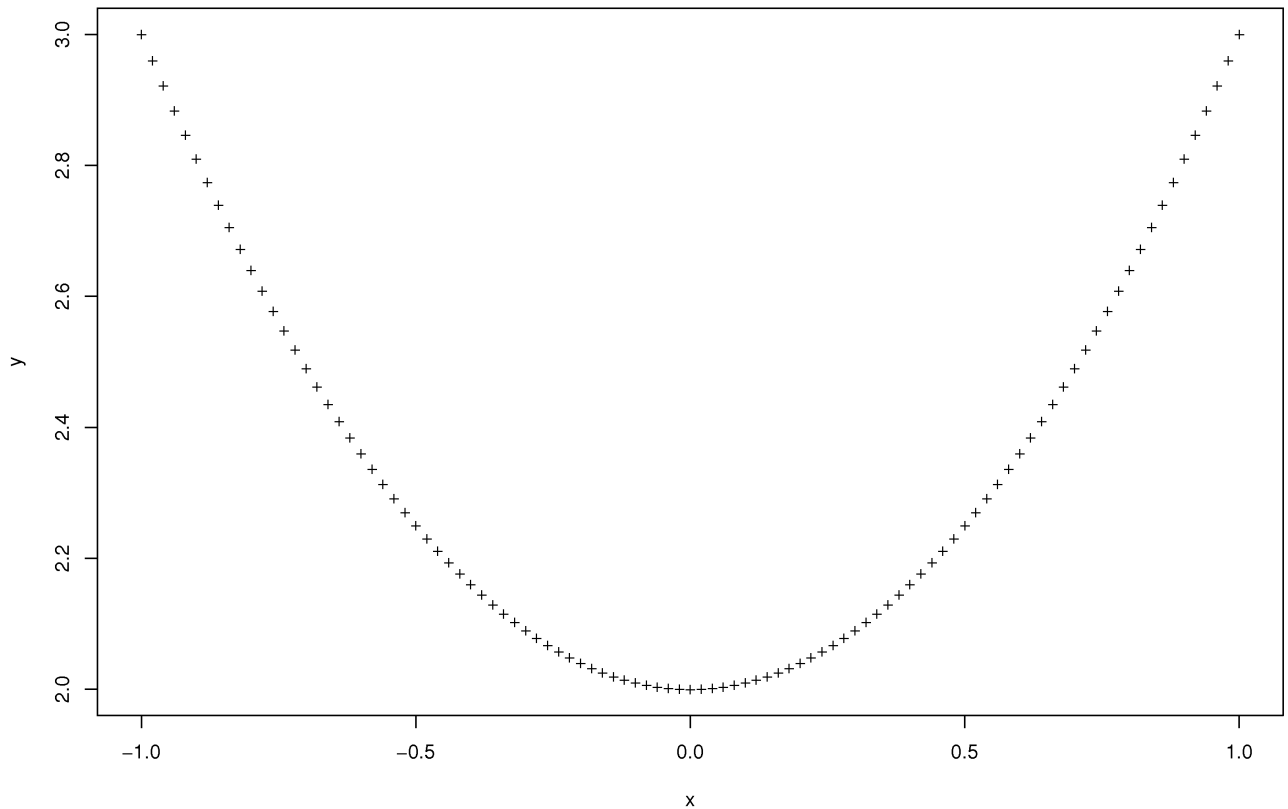
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 4: Periodische (unverrauschte) Funktion, $y = \sin(x)$, 101 equidistante Stützstellen im Intervall $[-\pi, \pi]$, $r =$



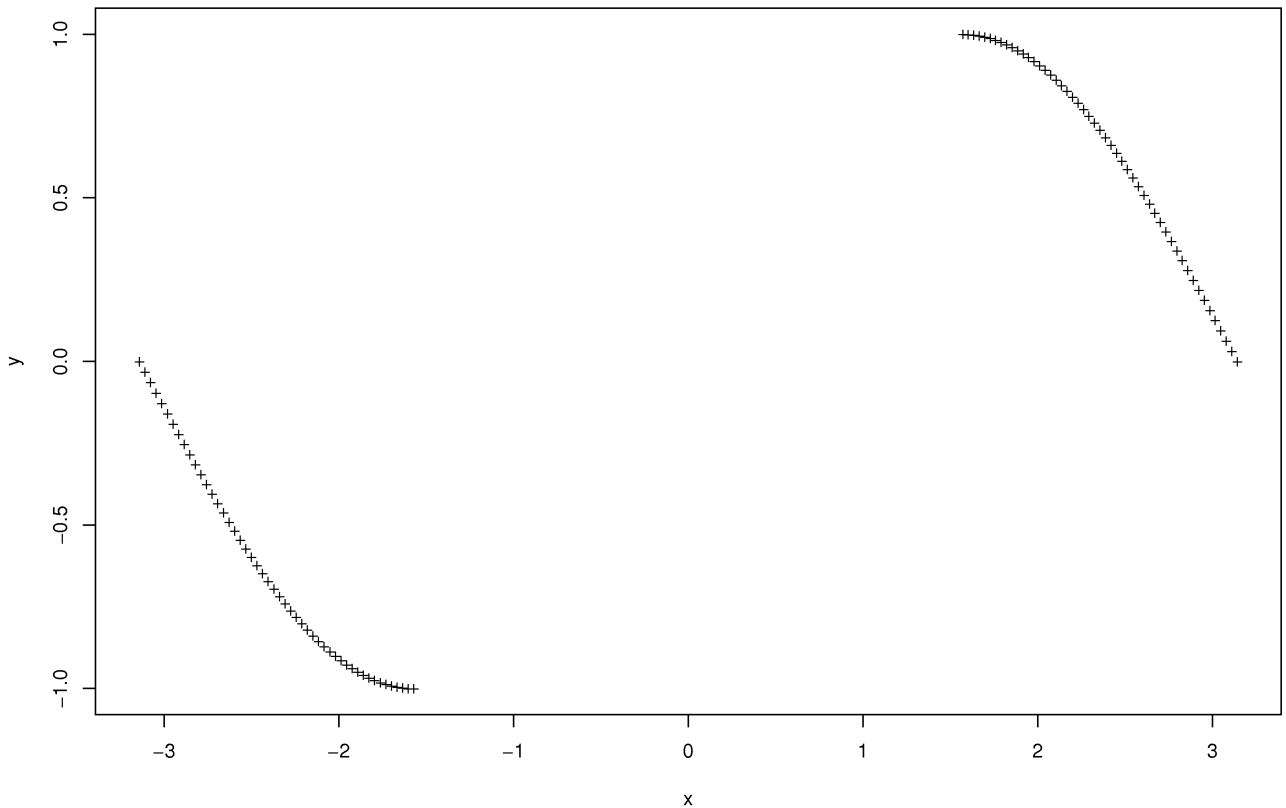
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 5: Quadratische (unverrauschte) Funktion, $y = x^2 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r =$



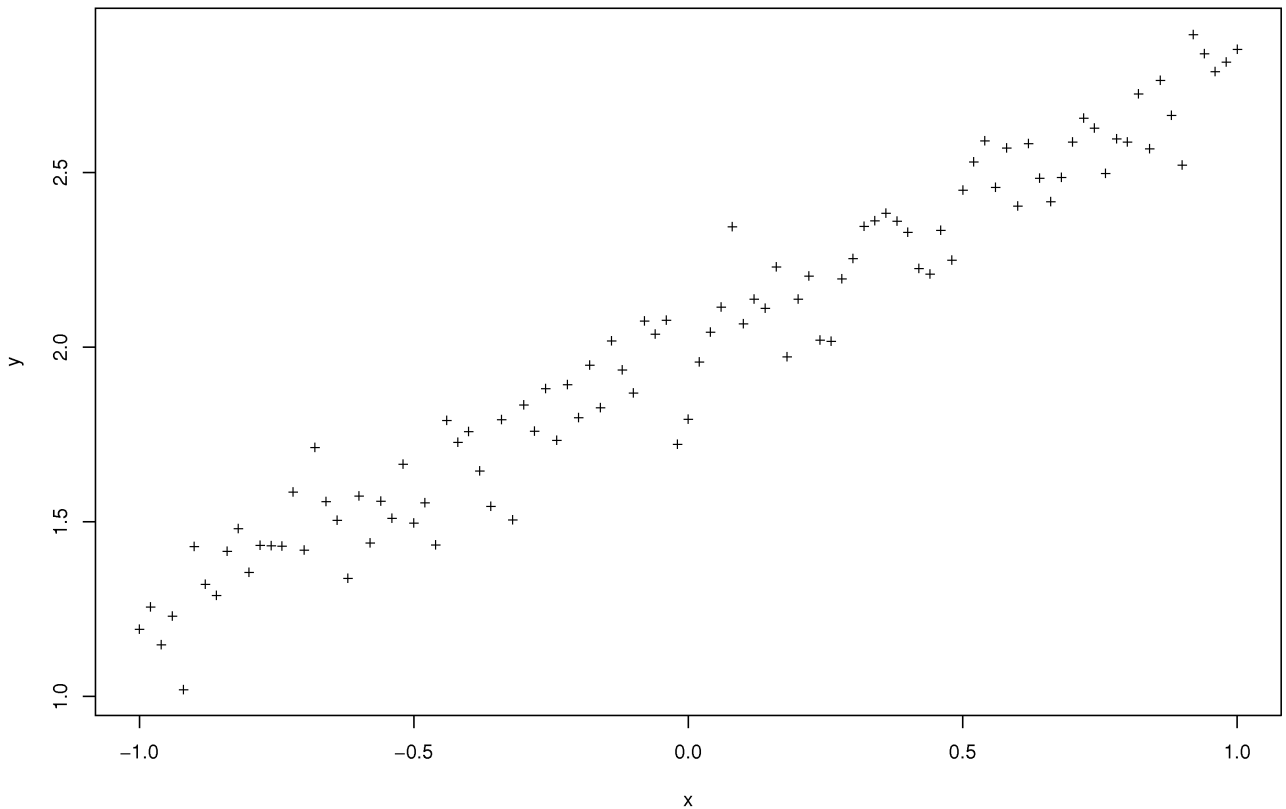
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 6: Abschnittweise definierte (unverrauschte) Funktion $y = \sin(x)$, 50 und 51 equidistante Stützstellen in den Intervallen $[-\pi, -\frac{\pi}{2}]$ und $[\frac{\pi}{2}, \pi]$, $r =$



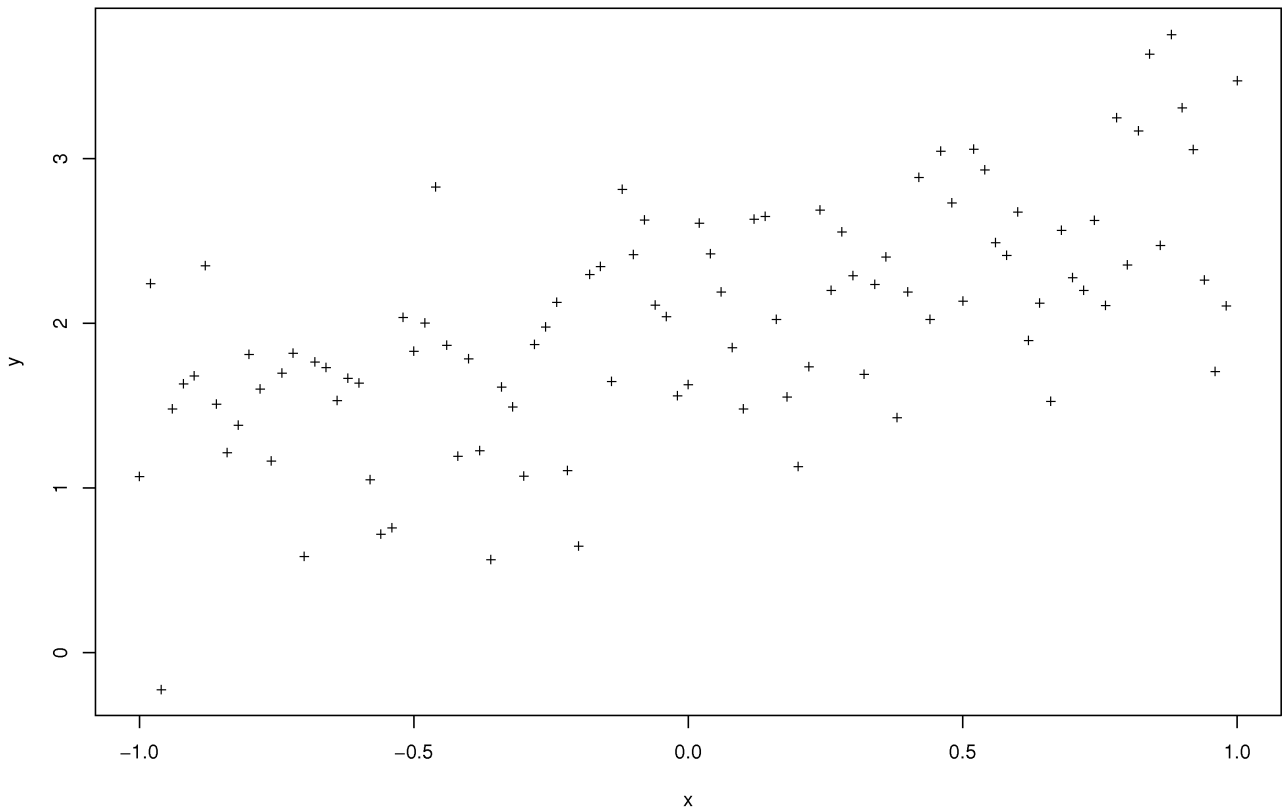
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 7: Lineare, schwach verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 8: Lineare, stärker verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Lineare Transformationen

- Bei exakten lineare Zusammenhängen gilt:

$$r = +1 \text{ bzw. } -1 \quad \Leftrightarrow \quad Y = aX + b \text{ mit } a > 0 \text{ bzw. } a < 0$$

- Lineare Transformationen

$\tilde{X} = a_X X + b_X, \tilde{Y} = a_Y Y + b_Y, a_X, a_Y \neq 0$
r Korrelationskoeffizient zwischen X und Y
$\tilde{r}$ Korrelationskoeffizient zwischen $\tilde{X}$ und $\tilde{Y}$

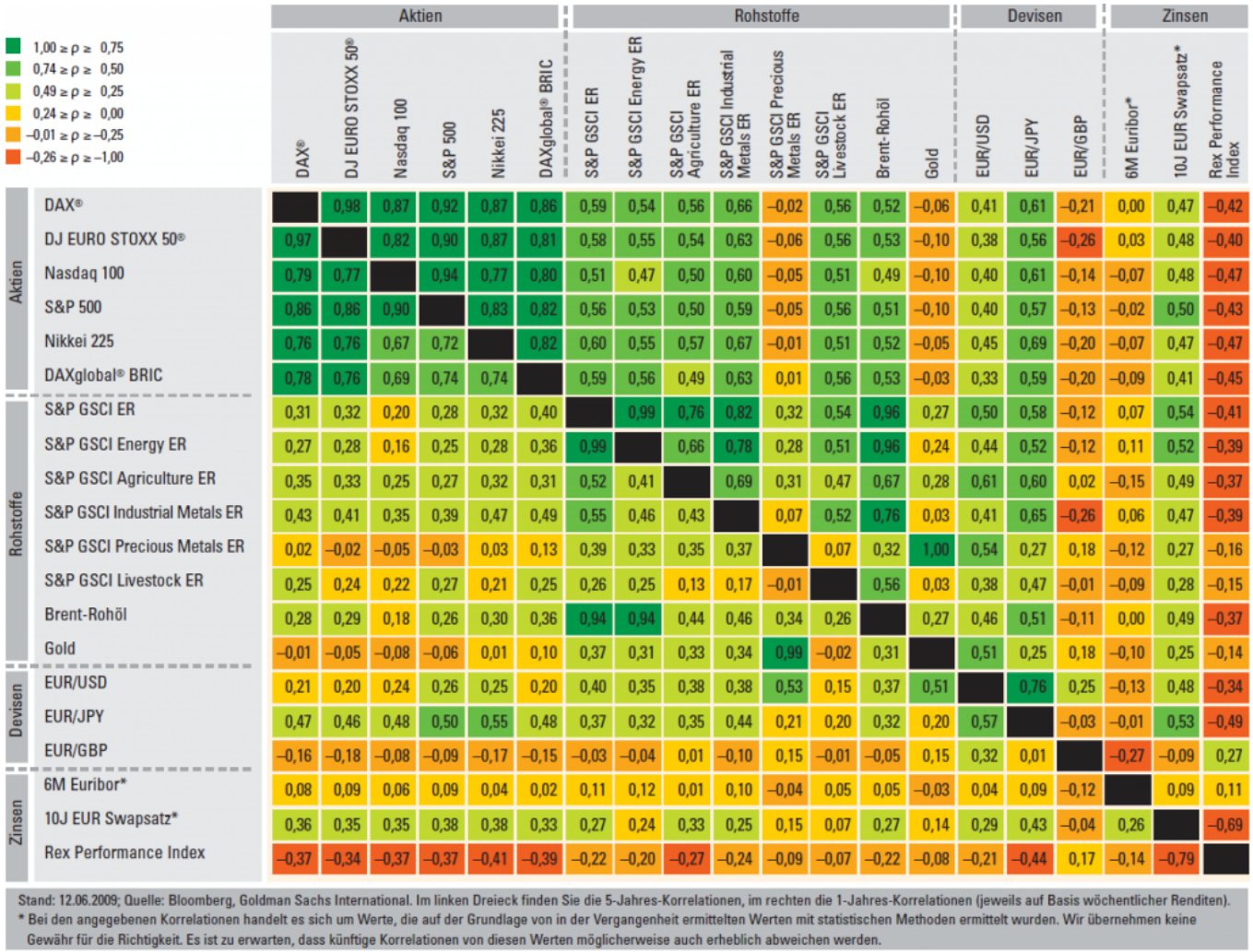
$$\Rightarrow \begin{array}{lll} \tilde{r} = r & \Leftrightarrow & a_X, a_Y > 0 \text{ oder } a_X, a_Y < 0 \\ \tilde{r} = -r & \Leftrightarrow & a_X > 0, a_Y < 0 \text{ oder } a_X < 0, a_Y > 0. \end{array}$$

Korrelationsmatrix

Bei mehr als zwei Merkmalen werden die Korrelationen häufig in Form einer Matrix dargestellt.
Auf der Hauptdiagonalen stehen 1er.
Die Matrix ist symmetrisch.

$$\begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix}$$

Beisipiel: Korrelationen am Finanzmarkt



1,00 ≥ ρ ≥ 0,75
0,74 ≥ ρ ≥ 0,50
0,49 ≥ ρ ≥ 0,25
0,24 ≥ ρ ≥ 0,00
–0,01 ≥ ρ ≥ –0,25
–0,26 ≥ ρ ≥ –1,00

		Aktien						Rohstoffe								Devisen			Zinsen		
		DAX®	DJ EURO STOXX 50®	Nasdaq 100	S&P 500	Nikkei 225	DAXglobal® BRIC	S&P GSCI ER	S&P GSCI Energy ER	S&P GSCI Agriculture ER	S&P GSCI Industrial Metals ER	S&P GSCI Precious Metals ER	S&P GSCI Livestock ER	Brent-Rohöl	Gold	EUR/USD	EUR/JPY	EUR/GBP	6M Euribor*	10J EUR Swapsatz*	Rex Performance Index
Aktien	DAX®		0,98	0,87	0,92	0,87	0,86	0,59	0,54	0,56	0,66	–0,02	0,56	0,52	–0,06	0,41	0,61	–0,21	0,00	0,47	–0,42
	DJ EURO STOXX 50®	0,97		0,82	0,90	0,87	0,81	0,58	0,55	0,54	0,63	–0,06	0,56	0,53	–0,10	0,38	0,56	–0,26	0,03	0,48	–0,40
	Nasdaq 100	0,79	0,77		0,94	0,77	0,80	0,51	0,47	0,50	0,60	–0,05	0,51	0,49	–0,10	0,40	0,61	–0,14	–0,07	0,48	–0,47
	S&P 500	0,86	0,86	0,90		0,83	0,82	0,56	0,53	0,50	0,59	–0,05	0,56	0,51	–0,10	0,40	0,57	–0,13	–0,02	0,50	–0,43
	Nikkei 225	0,76	0,76	0,67	0,72		0,82	0,60	0,55	0,57	0,67	–0,01	0,51	0,52	–0,05	0,45	0,69	–0,20	–0,07	0,47	–0,47
	DAXglobal® BRIC	0,78	0,76	0,69	0,74	0,74		0,59	0,56	0,49	0,63	0,01	0,56	0,53	–0,03	0,33	0,59	–0,20	–0,09	0,41	–0,45
Rohstoffe	S&P GSCI ER	0,31	0,32	0,20	0,28	0,32	0,40		0,99	0,76	0,82	0,32	0,54	0,96	0,27	0,50	0,58	–0,12	0,07	0,54	–0,41
	S&P GSCI Energy ER	0,27	0,28	0,16	0,25	0,28	0,36	0,99		0,66	0,78	0,28	0,51	0,96	0,24	0,44	0,52	–0,12	0,11	0,52	–0,39
	S&P GSCI Agriculture ER	0,35	0,33	0,25	0,27	0,32	0,31	0,52	0,41		0,69	0,31	0,47	0,67	0,28	0,61	0,60	0,02	–0,15	0,49	–0,37
	S&P GSCI Industrial Metals ER	0,43	0,41	0,35	0,39	0,47	0,49	0,55	0,46	0,43		0,07	0,52	0,76	0,03	0,41	0,65	–0,26	0,06	0,47	–0,39
	S&P GSCI Precious Metals ER	0,02	–0,02	–0,05	–0,03	0,03	0,13	0,39	0,33	0,35	0,37		0,07	0,32	1,00	0,54	0,27	0,18	–0,12	0,27	–0,16
	S&P GSCI Livestock ER	0,25	0,24	0,22	0,27	0,21	0,25	0,26	0,25	0,13	0,17	–0,01		0,56	0,03	0,38	0,47	–0,01	–0,09	0,28	–0,15
	Brent-Rohöl	0,28	0,29	0,18	0,26	0,30	0,36	0,94	0,94	0,44	0,46	0,34	0,26		0,27	0,46	0,51	–0,11	0,00	0,49	–0,37
	Gold	–0,01	–0,05	–0,08	–0,06	0,01	0,10	0,37	0,31	0,33	0,34	0,99	–0,02	0,31		0,51	0,25	0,18	–0,10	0,25	–0,14
Devisen	EUR/USD	0,21	0,20	0,24	0,26	0,25	0,20	0,40	0,35	0,38	0,38	0,53	0,15	0,37	0,51		0,76	0,25	–0,13	0,48	–0,34
	EUR/JPY	0,47	0,46	0,48	0,50	0,55	0,48	0,37	0,32	0,35	0,44	0,21	0,20	0,32	0,20	0,57		–0,03	–0,01	0,53	–0,49
	EUR/GBP	–0,16	–0,18	–0,08	–0,09	–0,17	–0,15	–0,03	–0,04	0,01	–0,10	0,15	–0,01	–0,05	0,15	0,32	0,01		–0,27	–0,09	0,27
Zinsen	6M Euribor*	0,08	0,09	0,06	0,09	0,04	0,02	0,11	0,12	0,01	0,10	–0,04	0,05	0,05	–0,03	0,04	0,09	–0,12		0,09	0,11
	10J EUR Swapsatz*	0,36	0,35	0,35	0,38	0,38	0,33	0,27	0,24	0,33	0,25	0,15	0,07	0,27	0,14	0,29	0,43	–0,04	0,26		–0,69
	Rex Performance Index	–0,37	–0,34	–0,37	–0,37	–0,41	–0,39	–0,22	–0,20	–0,27	–0,24	–0,09	–0,07	–0,22	–0,08	–0,21	–0,44	0,17	–0,14	–0,79	

Stand: 12.06.2009; Quelle: Bloomberg, Goldman Sachs International. Im linken Dreieck finden Sie die 5-Jahres-Korrelationen, im rechten die 1-Jahres-Korrelationen (jeweils auf Basis wöchentlicher Renditen).
* Bei den angegebenen Korrelationen handelt es sich um Werte, die auf der Grundlage von in der Vergangenheit ermittelten Werten mit statistischen Methoden ermittelt wurden. Wir übernehmen keine Gewähr für die Richtigkeit. Es ist zu erwarten, dass künftige Korrelationen von diesen Werten möglicherweise auch erheblich abweichen werden.

Spearmans Korrelationskoeffizient = Rang-Korrelationskoeffizient

X, Y (mindestens) ordinal

Idee: Gehe von Werten $x_i, \quad i = 1, \ldots, n$ und $y_i, \quad i = 1, \ldots, n$ über zu ihren Rängen.

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

$$rg(x_{(i)}) = i,$$

analog für $y_{(1)}, \ldots, y_{(n)}$.

Beispiel

x_i	2.3	7.1	1.0	2.1
$rg(x_i)$	3	4	1	2

bei Bindungen (ties):

x_i	2.3	7.1	1.0	2.1	2.3
	3.5	5	1	2	3.5

$\Rightarrow$ Durchschnittsrang $\frac{3+4}{2} = 3.5$ vergeben.
Also: Daten der Größe nach durchsortieren
$\Rightarrow$ Ranglisten $rg(x_i), rg(y_i), i = 1, \ldots, n$ vergeben (bei ties: Durchschnittsränge)
Idee: Berechne den Korrelationskoeffizienten nach Bravais-Pearson für die Ränge statt für die Ursprungsdaten.

Definition: Spearmans Korrelationskoeffizient

Der *Korrelationskoeffizient nach Spearman* ist definiert durch

$$r_{SP} = \frac{\sum(rg(x_i) - \bar{rg}_X)(rg(y_i) - \bar{rg}_Y)}{\sqrt{\sum(rg(x_i) - \bar{rg}_X)^2 \sum(rg(y_i) - \bar{rg}_Y)^2}}.$$

Wertebereich: $-1 \leq r_{SP} \leq 1$

Interpretation

$r_{SP} > 0$ gleichsinniger monotoner Zusammenhang,

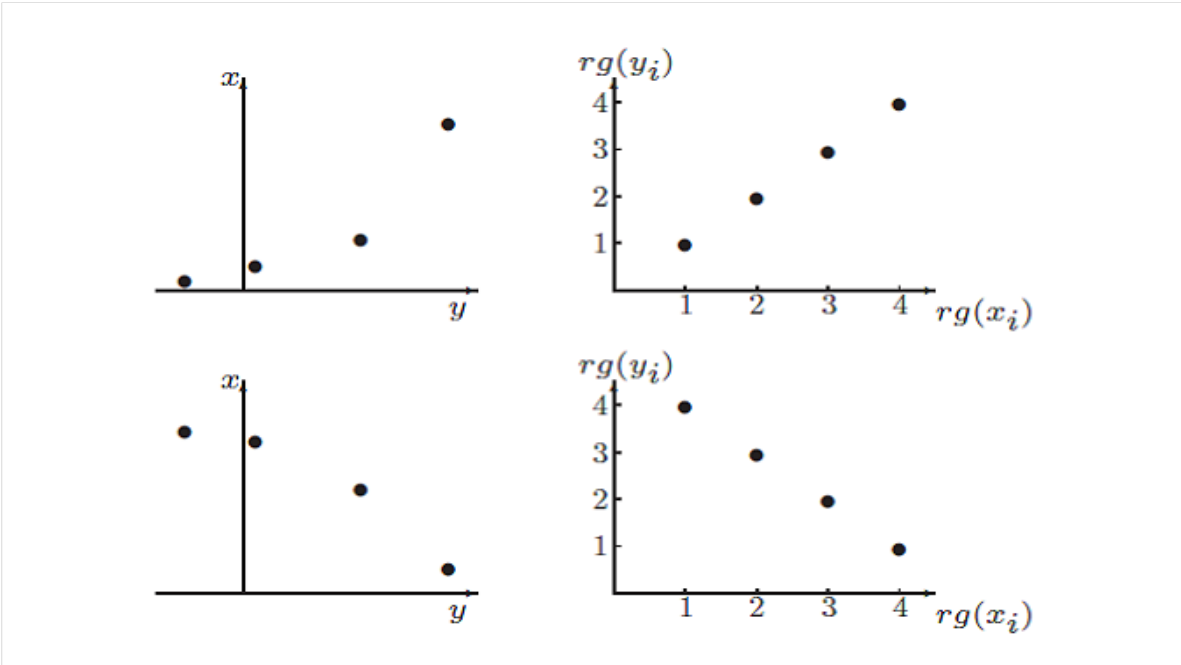
Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein

$r_{SP} < 0$ gegensinniger monotoner Zusammenhang,

Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß

$r_{SP} \approx 0$ kein monotoner Zusammenhang

Beispiele



Extremfälle für Spearmans Korrelationskoeffizienten, $r_{SP} = 1$ (oben) und $r_{SP} = -1$ (unten)

Spearmans Korrelationskoeffizient misst monotone (auch nichtlineare) Zusammenhänge!

Bemerkungen:

- Rechentechnische Vereinfachungen:

$$\bar{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2,$$

$$\bar{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2.$$

Rechentechnisch günstige Version von r_{SP}:

Daten: $(x_i, y_i),\ i = 1, \ldots, n, \quad x_i \neq x_j,\ y_i \neq y_j$ für alle i, j
Rangdifferenzen: $d_i = rg(x_i) - rg(y_i)$

$$r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n}$$

Voraussetzung: keine Bindungen

Monotone Transformationen

$\tilde{X} = g(X)$ $\quad g$ streng monoton,
$\tilde{Y} = h(Y)$ $\quad h$ streng monoton

$\Rightarrow \quad r_{SP}(\tilde{X}, \tilde{Y}) = r_{SP}(X, Y)$,
wenn g *und* h monoton wachsend
bzw. g *und* h monoton fallend sind,

$r_{SP}(\tilde{X}, \tilde{Y}) = -r_{SP}(X, Y)$,
wenn g monoton wachsend *und* h
monoton fallend bzw. g monoton
fallend *und* h monoton wachsend sind.

Kendalls Tau

Betrachte Paare von Beobachtungen (x_i, y_i) und (x_j, y_j)

Ein Paar heißt:

konkordant, falls $x_i < x_j$ und $y_i < y_j$
oder $x_i > x_j$ und $y_i > y_j$

diskordant, falls $x_i < x_j$ und $y_i > y_j$
oder $x_i > x_j$ und $y_i < y_j$

N_C: Anzahl der konkordanten Paare
N_D: Anzahl der diskordanten Paare

$$\tau_a = \frac{N_C - N_D}{n(n-1)/2}$$ Kendalls Tau

7 Regression

Einleitung

Motivation

In vielen Anwendungen ist es bedeutsam zu wissen, welchen *Einfluss* ein quantitatives Merkmal X auf ein weiteres Merkmal Y hat, z.B.

- Einkommen (X) und Kreditwunsch (Y) eines Bankkunden
- Einsatz von Werbung in € (X) und Umsatz in € (Y) einer Handelskette
- Geschwindigkeit (X) und Bremsweg (Y) eines Pkw

In diesem Abschnitt werden Methoden zur Analyse dieses Einflusses behandelt und wie dies in einem *Modell* formuliert werden kann.

Einfache lineare Regression

- Linearer Zusammenhang zwischen zwei metrischen Größen wird als Gerade visualisiert
- Finde Gerade $Y = \alpha + \beta \cdot X$

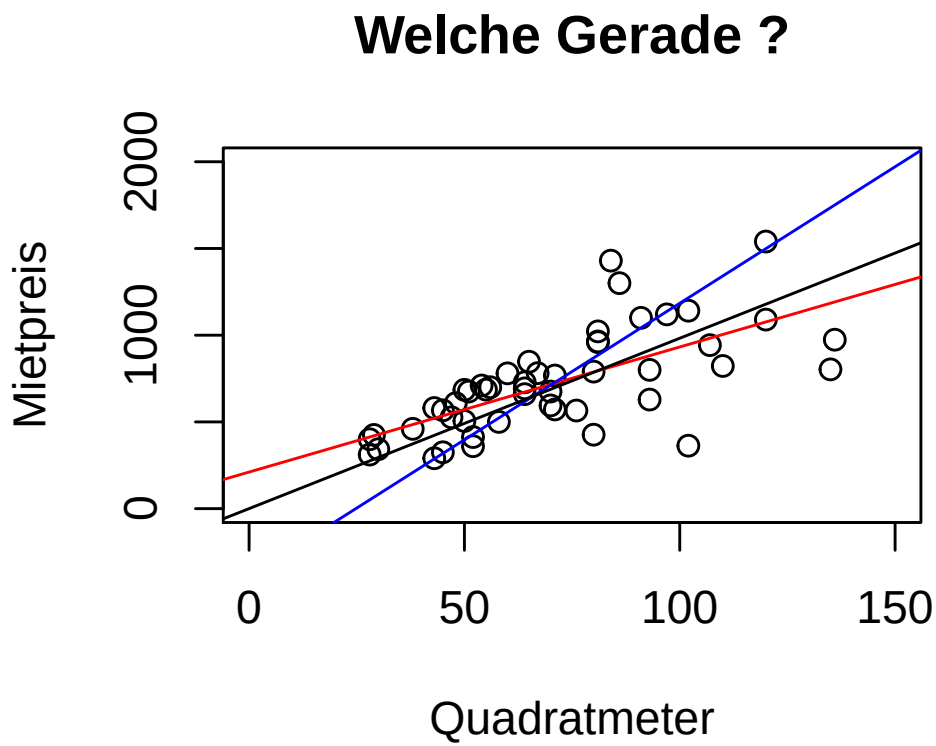
Einfache lineare Regression

- Linearer Zusammenhang zwischen zwei metrischen Größen wird als Gerade visualisiert
- Finde Gerade $Y = \alpha + \beta \cdot X$
- β : Steigung der Geraden, d.h. erhöht sich X um eine Einheit, so erhöht sich Y um β Einheiten.

Einfache lineare Regression

- Linearer Zusammenhang zwischen zwei metrischen Größen wird als Gerade visualisiert
- Finde Gerade $Y = \alpha + \beta \cdot X$
- β : Steigung der Geraden, d.h. erhöht sich X um eine Einheit, so erhöht sich Y um β Einheiten.
- α : Achsenabschnitt, d.h. Wert von Y für $X = 0$

Welche Gerade ?



Bestimmung der Regressionsgerade

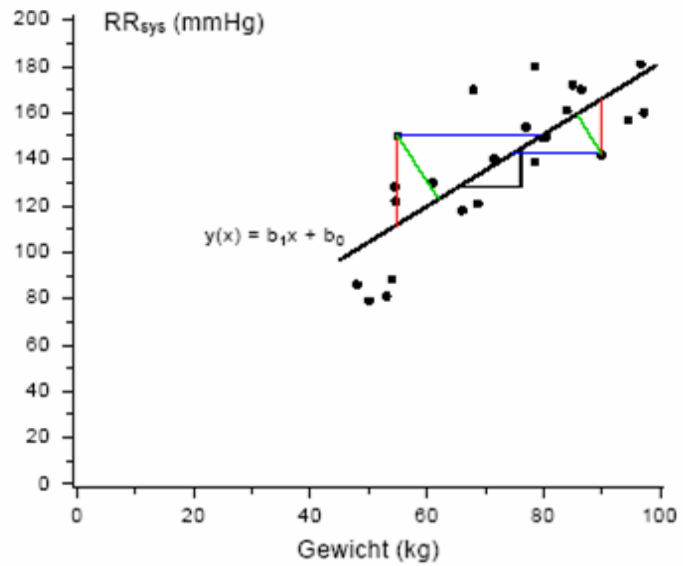
Welche Gerade ist die **Beste** ?

- Sie sollte etwa in der „Mitte“ der Punktwolke liegen
- Abweichungen der Wertepaare (x_i, y_i) (Punkte) von der Geraden sollten möglichst klein (minimal) sein

Bestimmung der Regressionsgerade

Welche Gerade ist die **Beste** ?

- Sie sollte etwa in der „Mitte“ der Punktwolke liegen
- Abweichungen der Wertepaare (x_i, y_i) (Punkte) von der Geraden sollten möglichst klein (minimal) sein



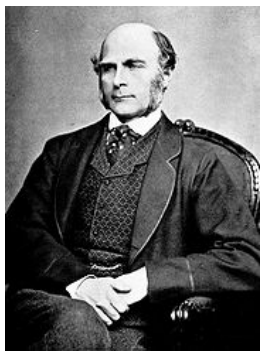
Methode der kleinsten Quadrate

- Y ist Zielgröße und X Einflussgröße
- Y soll mit Hilfe von X erklärt oder prognostiziert werden
- Lineares Modell $Y = \alpha + \beta X + \varepsilon$
- Minimierung der Abstände in Y-Richtung
- Wähle $\hat{\alpha}$ und $\hat{\beta}$ so, dass $\sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i) \right)^2$ minimal wird

Geschichte

Idee der KQ-Schätzung von Gauss (1795) im Alter von 18 Jahren

Veröffentlichung von Legendre
Idee der Regression von Galton (1886)



Lineare Einfachregression und Kleinste-Quadrate-Schätzer

Seien $(x_1, y_1), \ldots, (x_n, y_n)$ Beobachtungen der Merkmale X und Y, dann heißt

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

lineare Einfachregression, wobei α den Achsenabschnitt, β die Steigung und ε den Fehler bezeichnet.

Die Kleinste-Quadrate-Schätzer für $\hat{\alpha}$ und $\hat{\beta}$ sind gegeben durch

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{S_{xy}}{S_x^2}.$$

Die Residuen berechnen sich durch

$$\varepsilon_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n,$$

mit $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

Kleinste-Quadrate-Schätzer

Definition Residuum

Jedem Beobachtungspunkt $P_i = (x_i; y_i)$ wird ein angepasster Punkt $\hat{P}_i = (x_i; \hat{y}_i)$ zugeordnet, der auf der Geraden liegt und es daher gilt:

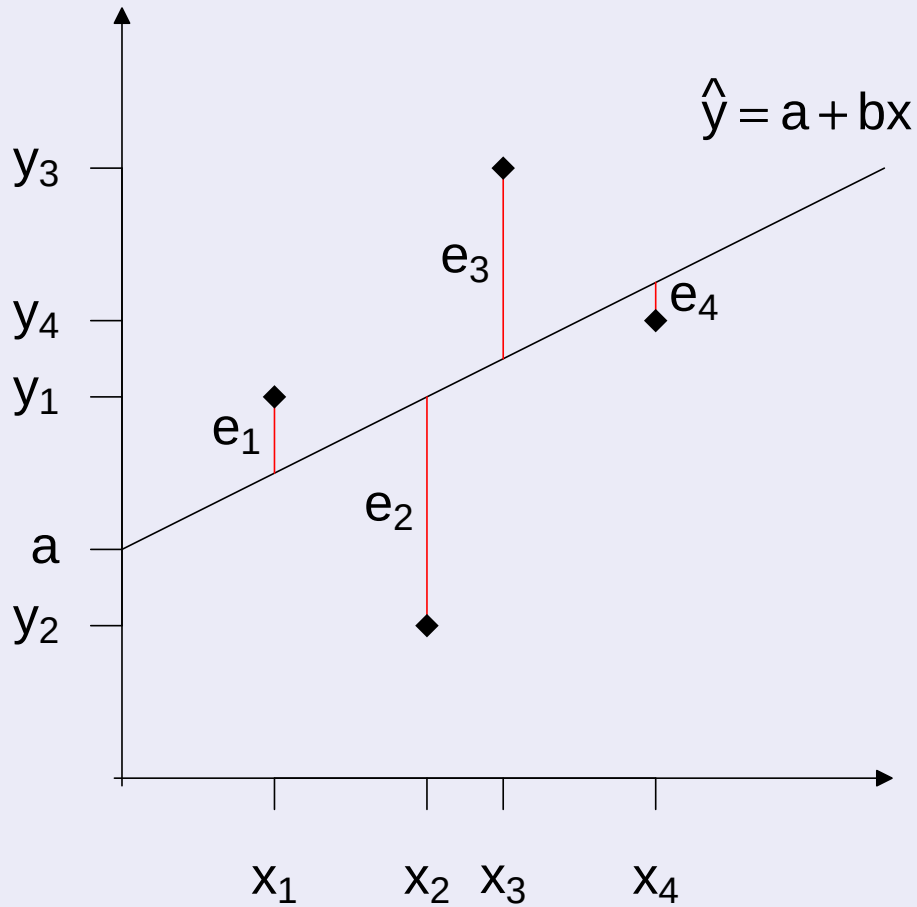
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Die Differenz (in y-Richtung) aus dem Beobachtungspunkt P_i und dem geschätzten Punkt $\hat{P}_i$ ergibt das *Residuum* oder *Fehlerglied*:

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

Kleinste-Quadrate-Schätzer

Residuen graphisch veranschaulicht



Herleitung: Kleinste-Quadrate-Schätzer

Vorgehensweise der Schätzung

1. 1. partiellen Ableitungen der Funktion $S(\alpha, \beta)$ bestimmen
2. Nullstellen der 1. Ableitungen finden
3. 2. partiellen Ableitungen bestimmen (Hesse-Matrix)
4. Ergebnisse aus *Punkt 2* in Hesse-Matrix einsetzten
5. prüfen, ob Hesse-Matrix positiv definit ist (alle Eigenwerte positiv)

Eigenschaften der Regressionsgeraden

sinnvoller Wertebereich

Die Regressionsgerade $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ lässt sich nur im Wertebereich $[x_{(1)}; x_{(n)}]$ der x-Werte sinnvoll interpretieren.

Lageparameter „arithmetisches Mittel“

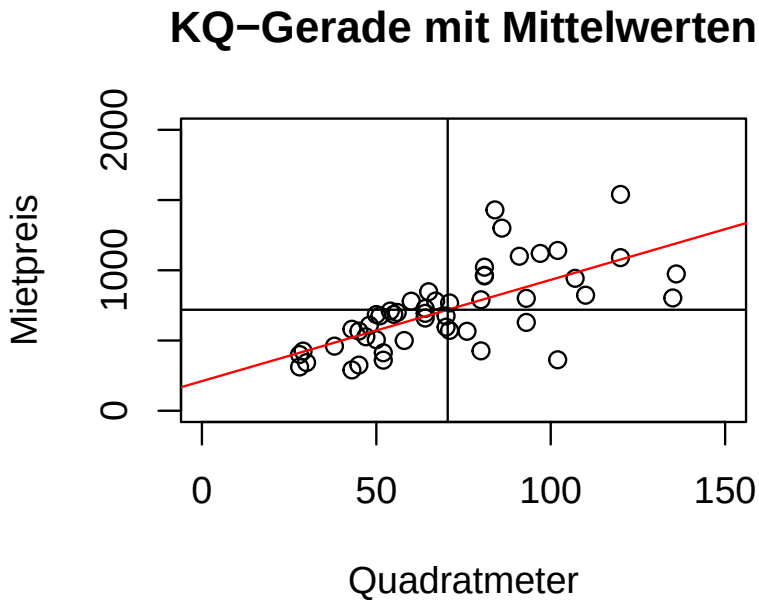
Der Punkt $(\bar{x}; \bar{y})$, physikalisch betrachtet der Schwerpunkt der bivariaten Daten $(x_i; y_i)$, liegt auf der Regressionsgerade.

Fehlerausgleich

Die Summe der negativen Residuen (absolut genommen) gleicht der Summe der positiven Residuen.
Die durch die Regression angepassten Werte $\hat{y}_i$ haben das gleiche arithmetische Mittel wie die Originaldaten y_i:

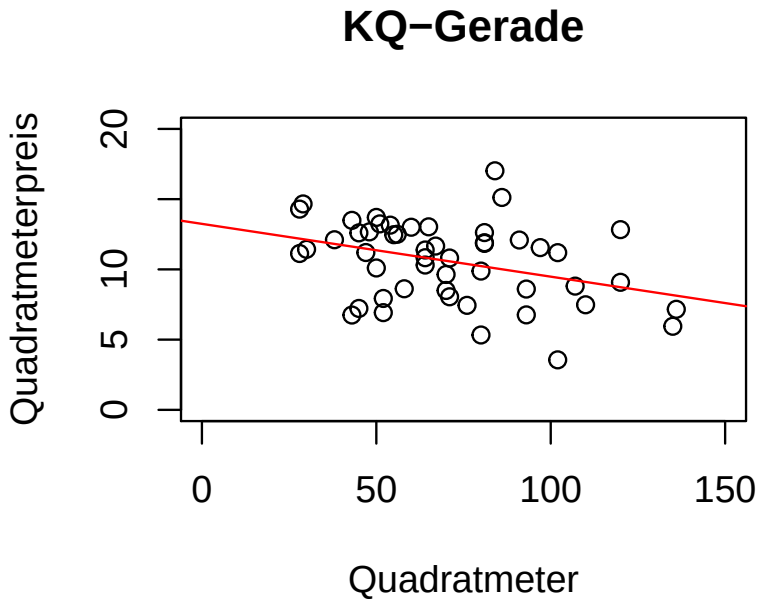
$$\bar{\hat{y}} = \bar{y}$$

Beispiel: Mietspiegel



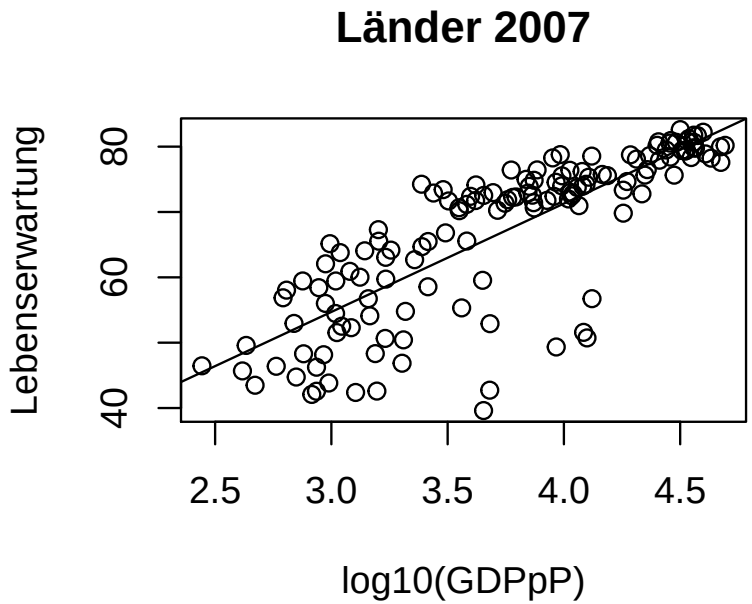
Schätzung der Koeffizienten: $\hat{\alpha} = 210.8$ $\hat{\beta} = 7.2$
Interpretation:
Mit einer Steigerung der Wohnfläche um eine Einheit steigt die Miete im Durchschnitt um 7.2 Euro. Achsenabschnitt: 210.8 ?

Beispiel: Mietspiegel



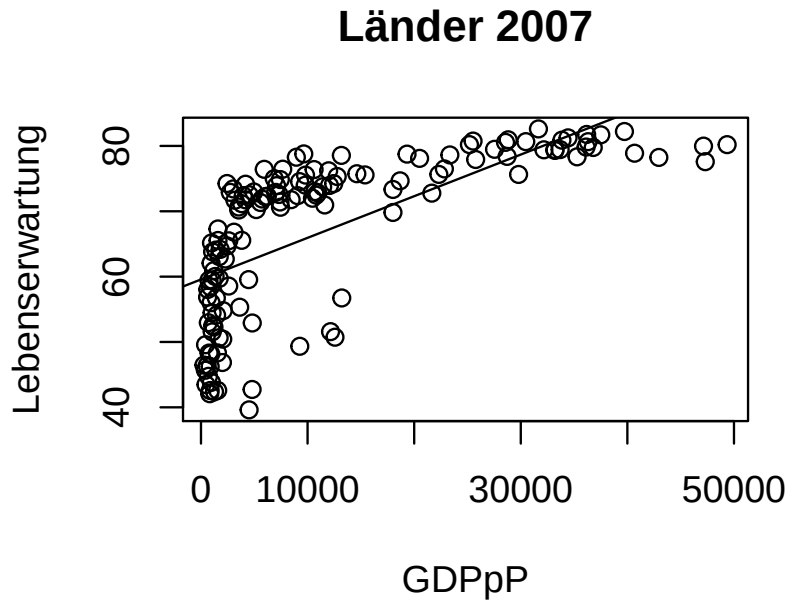
Schätzung der Koeffizienten: $\hat{\alpha} = 13.24$ $\hat{\beta} =$ - 0.038
Interpretation:
Mit einer Steigerung der Wohnfläche um eine Einheit fällt die Miete pro Quadratmeter im Durchschnitt um 0.038 Euro.
Achsenabschnitt: 13.24 ?

Beispiel: Lebenserwartung und GDP



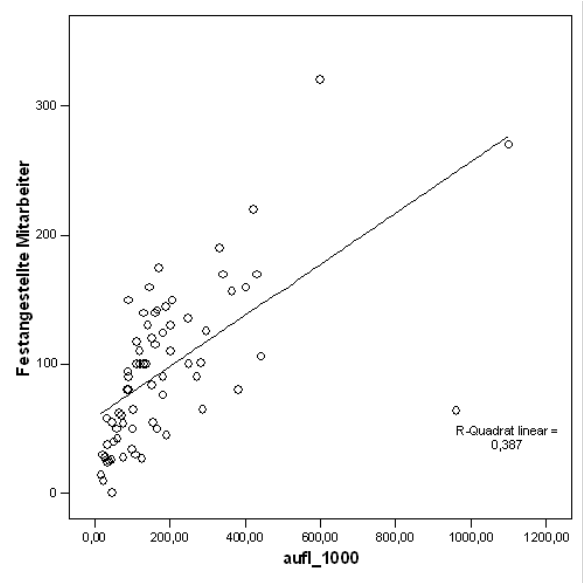
Schätzung der Koeffizienten: $\hat{\alpha} = 4.95$ $\hat{\beta} = 16.5$
Mit einer Steigerung des log (GDP) um eine Einheit (Steigerung um den Faktor 10) steigt die Lebenserwartung im Durchschnitt um 16.5 Jahre. **Besser**: Ist in einem Land das GDP pro Kopf um den Faktor 10 höher als im Land B, so ist dort die durchschnittliche Lebenserwartung um 16.5 Jahre größer.

Beispiel: Lebenserwartung und GDP



Schätzung der Koeffizienten: $\hat{\alpha} = 59$ $\hat{\beta} = 0.00064$
Interpretation:
Mit einer Steigerung des GDP pro Kopf um eine Einheit (Dollar) steigt die Lebenserwartung im Durchschnitt um 0.00064 Jahre
Vorsicht!!!

Beispiel: Mitarbeiter/Auflage bei Tageszeitungen



Interpretation: Auflagensteigerung schafft Arbeitsplätze
Mit einer Auflagensteigerung von 1000 ist durchschnittlich die Einstellung von 0.199 Mitarbeitern verbunden.

Koeffizienten[a]

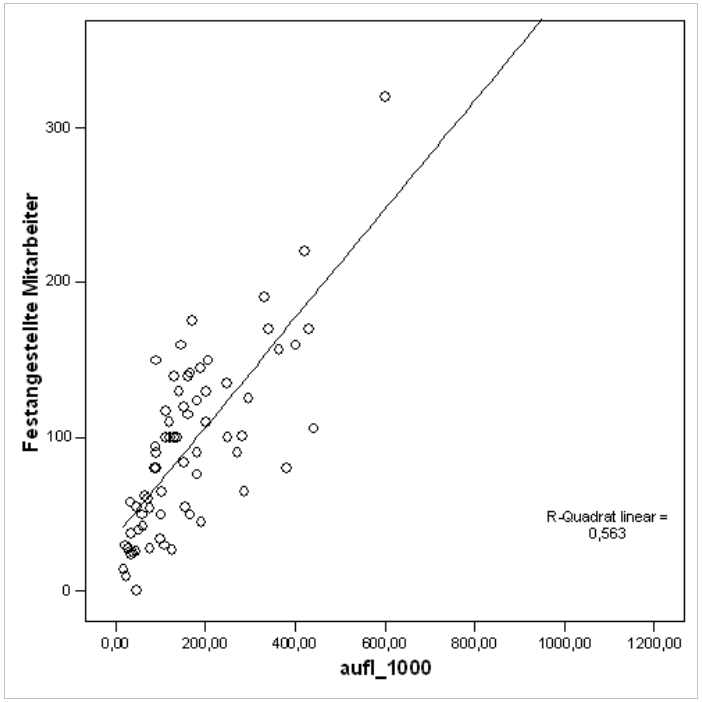
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Korrelationen		
		B	Standardfehler	Beta			Nullter Ordnung	Partiell	Teil
1	(Konstante)	58,193	8,043		7,235	,000			
	aufl_1000	,199	,030	,622	6,549	,000	,622	,622	,622

a. Abhängige Variable: Festangestellte Mitarbeiter

Ausreißer

Häufig wird ein erkennbarer Zusammenhang durch einzelne, von der großen Masse der Daten wesentlich entfernt liegende Werte gestört.
Diese sogenannten *Ausreißer* müssen gesondert eingeschätzt und gegebenenfalls - bei sachlicher oder statistischer Rechtfertigung - aus dem Datensatz entfernt werden.

Regression ohne 2 Extremwerte



Beachte:
Jetzt werden 0.352 Mitarbeiter bei einer Auflagensteigerung von 1000 eingestellt.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Korrelationen		
		B	Standardfehler	Beta			Nullter Ordnung	Partiell	Teil
1	(Konstante)	36,078	7,740		4,661	,000			
	aufl_1000	,352	,038	,750	9,220	,000	,750	,750	,750

a. Abhängige Variable: Festangestellte Mitarbeiter

Standardabweichung des Störterms

Die geschätzte Abweichung der y-Werte von der Geraden ergibt sich zu:

$$\begin{aligned} s_\varepsilon &= \sqrt{\frac{1}{n-2}\sum \varepsilon_i^2} \\ \varepsilon_i &= y_i - \hat{y}_i \end{aligned}$$

Wichtiges intuitives Maß zur Modellanpassung

Streuungs- und Quadratsummenzerlegung

Ziel: Erklärung der Streuung von Y durch X:

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

Streuung von Y	=	Erklärte Streuung	+	Rest
SST	=	SSM	+	SSE

Streuungs- und Quadratsummenzerlegung

Ziel: Erklärung der Streuung von Y durch X:

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

Streuung von Y	=	Erklärte Streuung	+	Rest
SST	=	SSM	+	SSE
Quadratsumme Gesamt (Total)	=	Quadratsumme Regression (Model)	=	Quadratsumme Residuen (Error)

Das Bestimmtheitsmaß R^2

Anteil der durch die Regression (d.h. durch X) erklärten Varianz

$$\begin{aligned} R^2 &= \frac{SSM}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

Es gilt: Bestimmtheitsmaß = Quadrat der Korrelation zwischen X und Y

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2$$

Nachweis von $R^2 = r_{XY}^2$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i) = \hat{\alpha}+\hat{\beta}\bar{x} = (\bar{y}-\hat{\beta}\bar{x})+\hat{\beta}\bar{x} = \bar{y}$$

Daraus folgt:

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2 = \sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}+\hat{\beta}\bar{x})^2 = \hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

somit für R^2:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \\ &= \frac{s_{XY}^2\cdot s_X^2}{(s_X^2)^2\cdot s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y}\right)^2 = r_{XY}^2 \end{aligned}$$

Interpretation von R^2

R^2 sollte bei linearen Regressionen immer angegeben werden.
Interpretation und Eigenschaften

- Zentrales Maß zur Güte der Modellanpassung
- Erklärter Anteil der Varianz
- Liegt zwischen 0 und 1
- Allgemeine Regeln zur Einschätzung problematisch
- R^2 hängt sowohl von den Abweichungen von der Regressiongeraden als auch von der Streuung der X-Werte ab.
- Wichtige Alternative und Ergänzung: Angabe der Standardabweichung der Residuen s_ε

Umkehrregression

Vertauscht man die Rollen von X und Y, so erhält man die Umkehrregression.

Daten $(X_i, Y_i), \quad i = 1, \ldots, n$

Regression:	$Y = \alpha + \beta X$	$\beta = \frac{s_{XY}}{s_X^2}$
Umkehrregression:	$X = \gamma + \delta Y$	$\delta = \frac{s_{XY}}{s_Y^2}$

Im XY-Koordinatensystem hat die Gerade der Umkehrregression die Darstellung

$$Y = -\frac{\gamma}{\delta} + \frac{1}{\delta} X$$

Umkehrregression

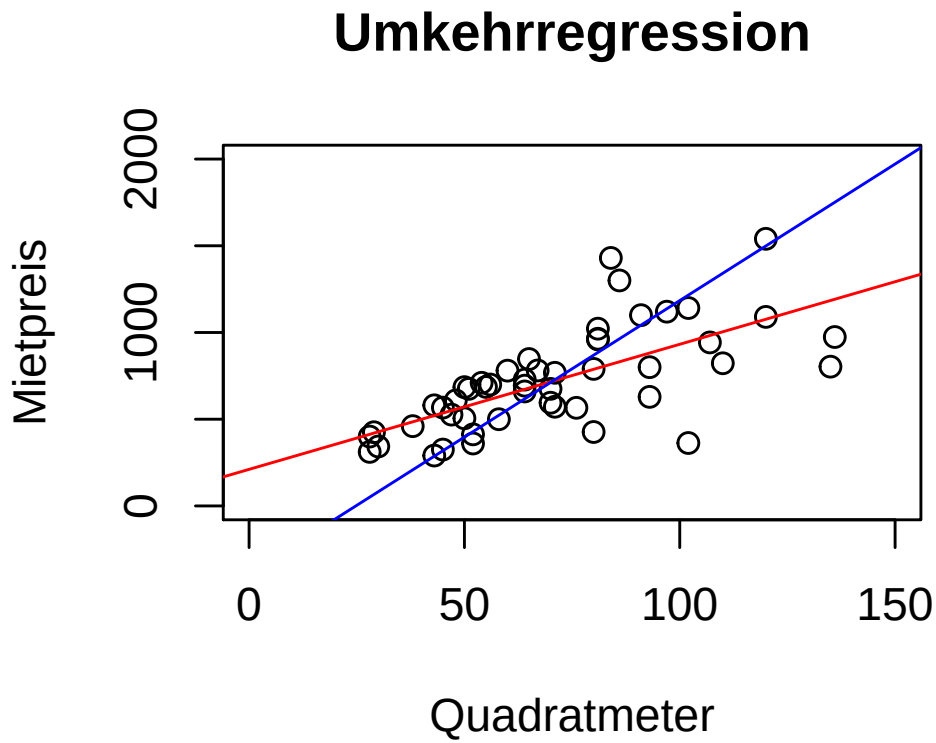
Es gilt:

$$\beta \cdot \delta = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r^2 \leq 1$$

$\Rightarrow |\beta| \leq \frac{1}{|\delta|}$
Gerade der Umkehrregression steiler

und
$\Rightarrow \beta \cdot \delta \geq 0$
β und δ haben gleiches Vorzeichen

Beispiel: Umkehrregression Mietspiegel

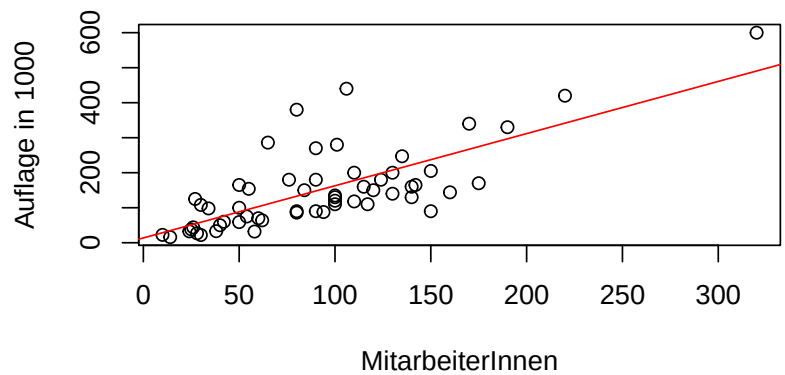


Beachte: Gerade der Umkehrregression steiler (blaue Gerade). Schnittpunkt im Schwerpunkt (x: Mittelwert von Wohnfläche, y: Mittelwert von Monatsmiete)

Beispiel: Auflage und Zahl der Mitarbeiter

$$\text{Auflage in 1000} = \alpha + \beta \cdot \text{Zahl der Mitarbeiter}$$

„Mitarbeiter produzieren Auflage“



Ergebnisse: $R^2 = 0.54$ $\alpha = 13.8$ $\beta = 1.5$ $s_\varepsilon = 80$
Interpretation: ?

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung β steht im Mittelpunkt der Interpretation

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung β steht im Mittelpunkt der Interpretation
- R^2-Wert gibt den Varianz-Erklärungsanteil wieder

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung β steht im Mittelpunkt der Interpretation
- R^2-Wert gibt den Varianz-Erklärungsanteil wieder
- R^2 ist Quadrat der Korrelation

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung β steht im Mittelpunkt der Interpretation
- R^2-Wert gibt den Varianz-Erklärungsanteil wieder
- R^2 ist Quadrat der Korrelation
- s_ε gibt durchschnittliche Abweichung der Werte von der Regressionsgeraden an

Zusammenfassung

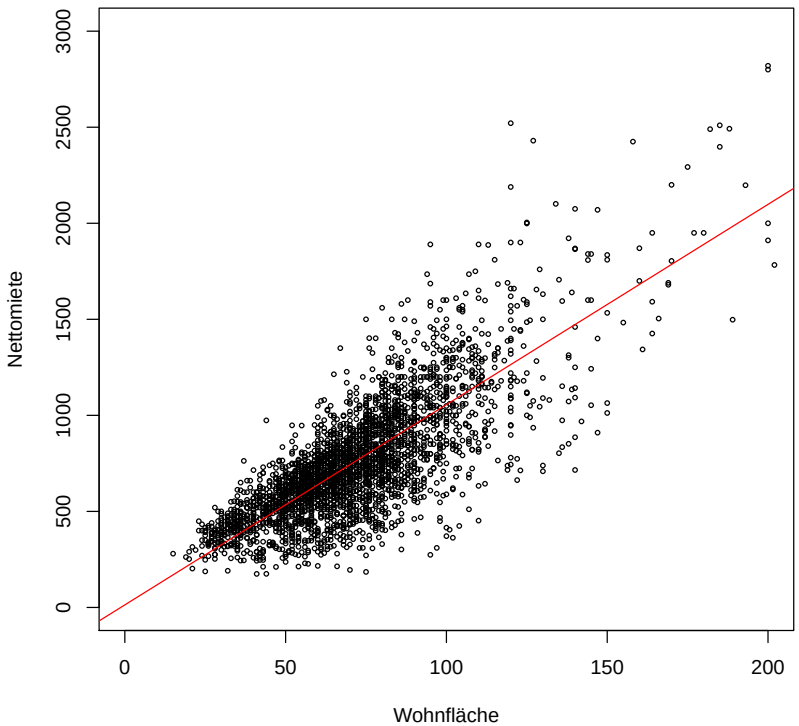
Bei der Auswertung durch eine lineare Regression sollten immer angegeben werden

- Regressionskoeffizienten α und β
- Bestimmtheitsmaß R^2
- Standardabweichung der Residuen s_ε
- Scatter-Plot mit Regressionsgeraden zur Kontrolle des Modells

8 Komplexe Zusammenhänge

Ergänzung: Interpretation R^2

Mietspiegeldaten (vollständig)

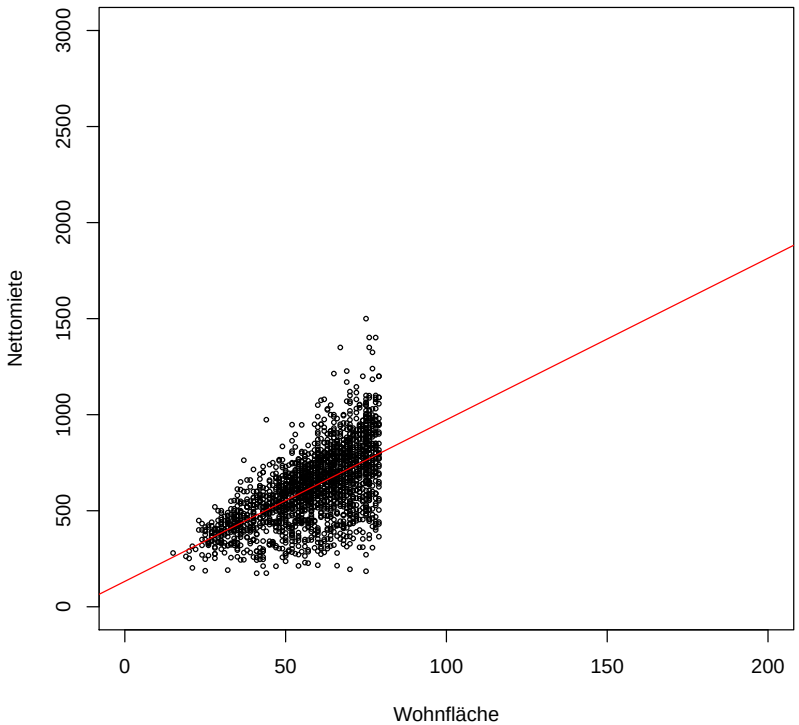


Kenngrößen

α	132.2501
β	8.4116
R^2	0.46
σ_E	215.7

Einschränkung auf kleine Wohnungen

Mietspiegeldaten. $Wfl < 80\ m^2$



Kenngrößen

α	13.3
β	10.4
R^2	0.36
σ_E	152

Anpassung besser aber erklärte Varianz geringer, da Ausgangsvarianz geringer

Partielle Korrelation

Ziel:
Bestimmung der Korrelation zweier Merkmale unter Berücksichtigung eines dritten Merkmals

Beispiel 1:
Korrelation der Zahl der freien und festen Mitarbeiter in Zeitungen
These: Je mehr freie Mitarbeiter desto weniger feste Mitarbeiter
Daten : positive Korrelation ???

Frage: Kommt die Positive Korrelation durch die Größe der Zeitung ? Problem der Drittvariablen

Trends in ökonomischen Daten

Beispiel 2: Lebenserwartung und Person GDP in Deutschland
Beide Größen haben einen Trend. Ist damit der Zusammenhang erklärbar ?

Strategie zum Umgang mit Drittvariablen

Es interessiert der Zusammenhang zwischen X und Y unter Berücksichtigung der Drittvariablen Z
Strategie : Wir bereinigen X und Y um den Einfluss von Z mit Hilfe linearer Regression

1. Berechne lineare Regression von X auf Z
2. Die Residuen REX dieser Regression entsprechen den um den Einfluss von Z bereinigten Werten von X
3. Berechne lineare Regression von Y auf Z
4. Die Residuen REY dieser Regression entsprechen den um den Einfluss von Z bereinigten Werten von Y
5. Die Korrelation von REX und REY ist dann die bereinigte (partielle) Korrelation von X und Y

Partieller Korrelationskoeffizient (Definition)

Es sei:

$$\begin{aligned} x &= \hat{\alpha} + \hat{\beta} Z + rex \\ y &= \hat{\gamma} + \hat{\delta} Z + rey \end{aligned}$$

Dann heißt die Maßzahl

$$r_{XY|Z} = r_{rexrey}$$

partieller Korrelationskoeffizient zwischen X und Y unter Z.

Es gilt:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2} \sqrt{1 - r_{YZ}^2}}$$

Beispiel: Korrelation der Anzahl freier Mitarbeiter mit der Anzahl fest angestellter Mitarbeiter

Korrelationen

		Festangestellte Mitarbeiter	Freie Mitarbeiter
Festangestellte Mitarbeiter	Korrelation nach Pearson	1	,490**
	Signifikanz (2-seitig)		,000
	N	68	57
Freie Mitarbeiter	Korrelation nach Pearson	,490**	1
	Signifikanz (2-seitig)	,000	
	N	57	57

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

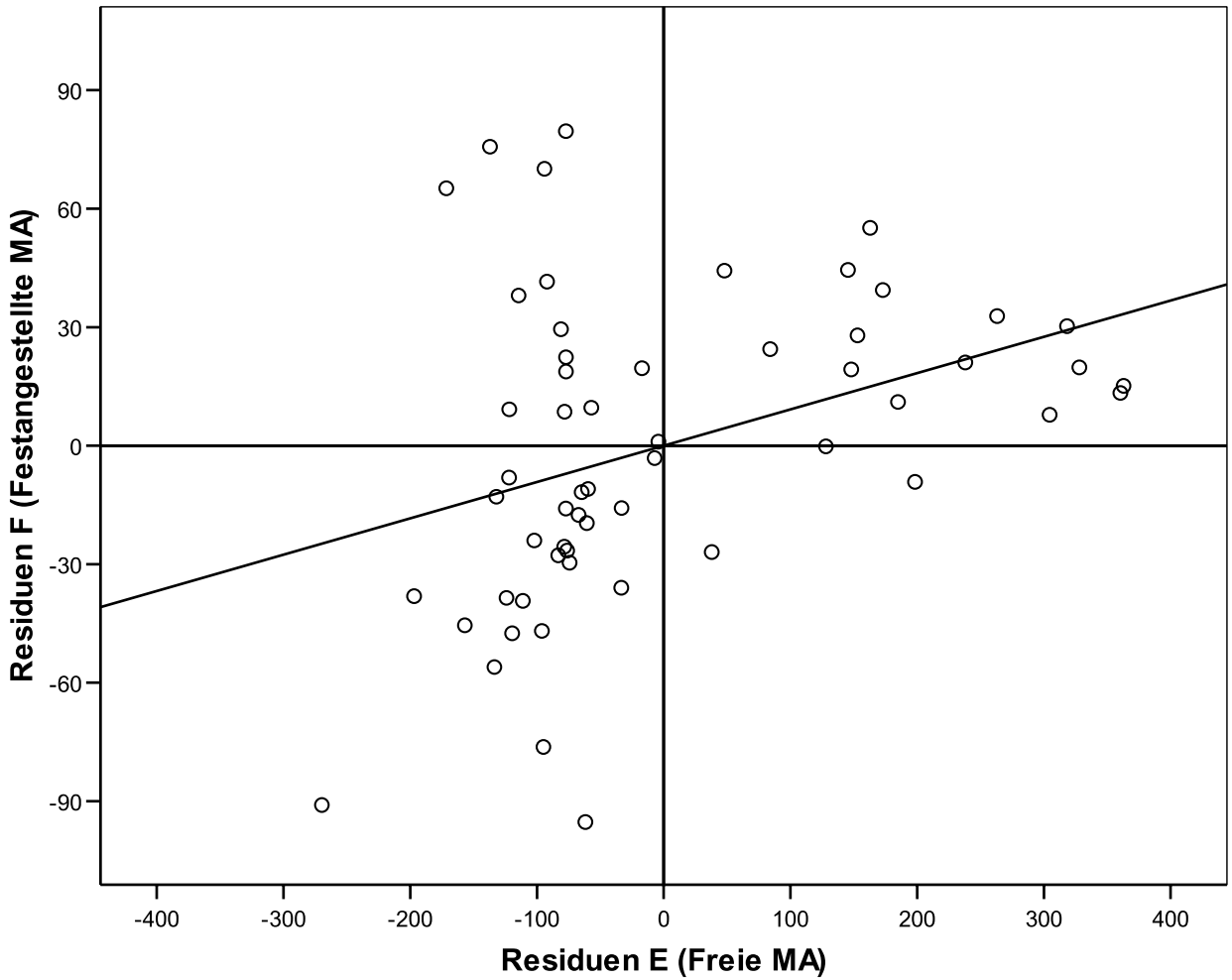
Einfache Korrelation

Korrelationen

Kontrollvariablen			Festangestellte Mitarbeiter	Freie Mitarbeiter
aufl_1000	Festangestellte Mitarbeiter	Korrelation	1,000	,366
		Signifikanz (zweiseitig)	.	,006
		Freiheitsgrade	0	54
	Freie Mitarbeiter	Korrelation	,366	1,000
		Signifikanz (zweiseitig)	,006	.
		Freiheitsgrade	54	0

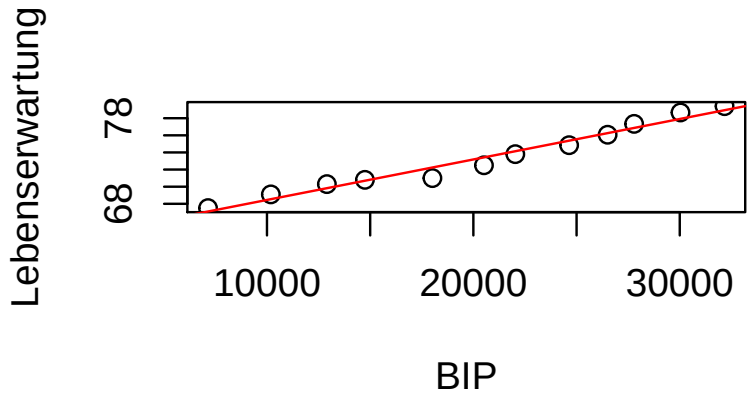
Nach Auflage bereinigte Korrelation

Freie und fest angestellte Mitarbeiter in der Zeitungsstudie (bereinigt nach der Größe der Zeitung)



Korrelation von Zeitreihen

Beispiel Lebenserwartung und GDP

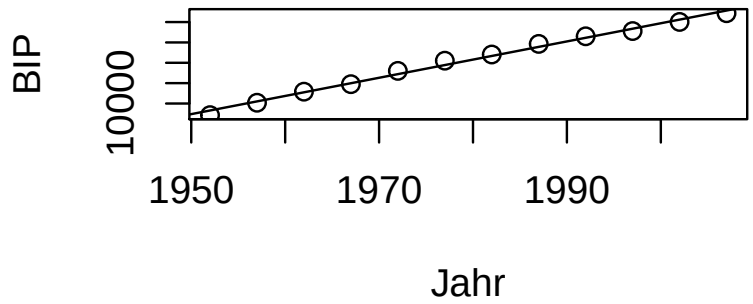


Kenngrößen		
	α	nicht relevant
	β	0.00047
	R^2	0.97
	σ_E	0.67

Sinnvolle Aussagen möglich ?

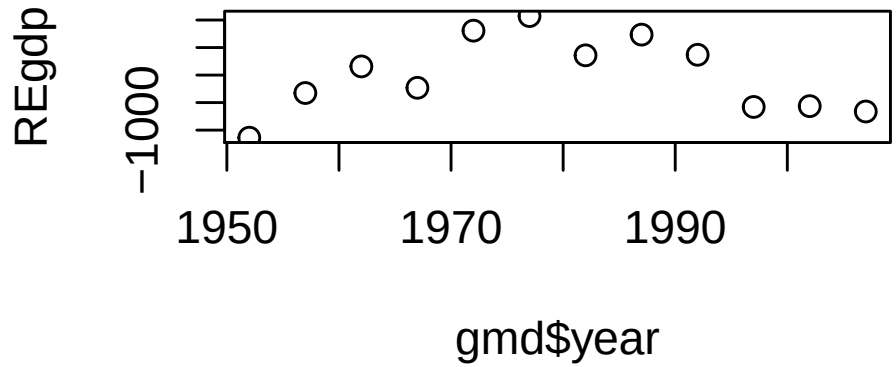
Trendbereinigung

Regressionsmodell: $BIP = \alpha + \beta * t$, t: Jahr (z.B. 1950 = 0)



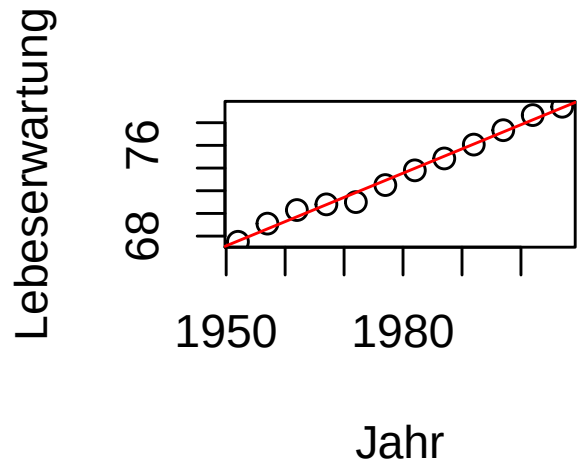
Kenngrößen	α	nicht relevant
	β	446
	R^2	0.99
	σ_E	719

Bereinigte Werte



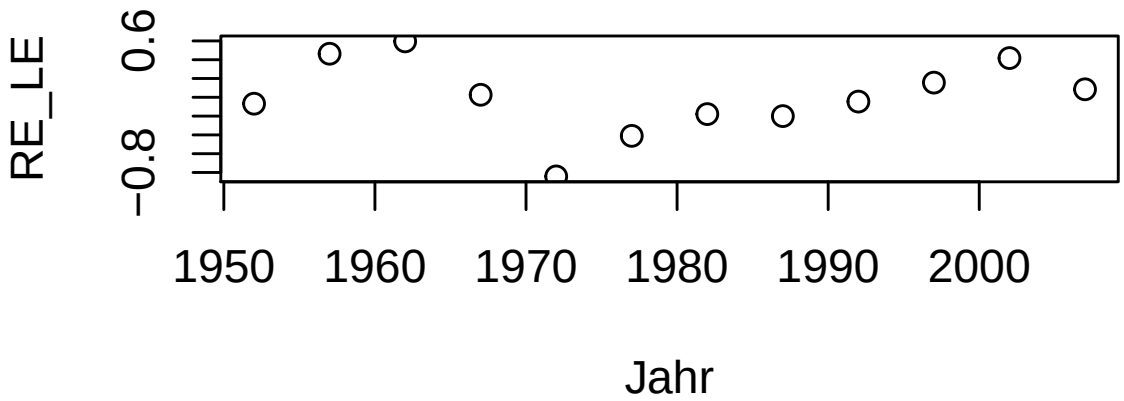
Trendbereinigung Lebenserwartung

Regressionsmodell: $LE = \alpha + \beta * t$, t: Jahr (z.B. 1950 = 0)

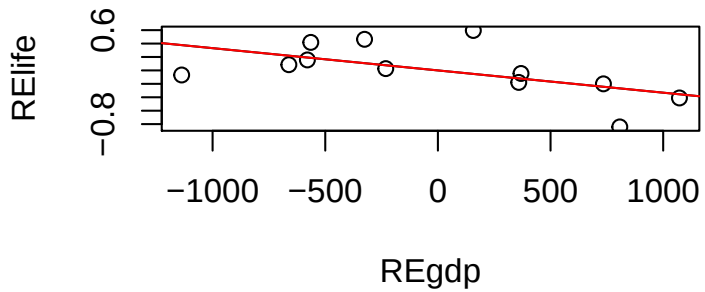


Kenngrößen	α	nicht relevant
	β	0.21
	R^2	0.99
	σ_E	0.41

Bereinigte Werte



Zusammenhang bereinigte Werte



Kenngrößen

α	0
β	- 0.00037
R^2	0.32
σ_E	0.34

Ausblick: Multiples Regressionsmodell

Gegeben sind ein Zielmerkmal Y und die Einflussgrößen X_k

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_p \cdot x_p + \varepsilon$$

Das Modell kann aus den entsprechenden Daten mit Hilfe der KQ-Methode geschätzt werden. Analog zum linearen Modell ist das Bestimmtheitsmaß R^2 ein zentrales Kriterium für die Modellanpassung.

Die Parameter β_k haben folgende Interpretation:
Steigt das Merkmal X_k um eine Einheit und werden die anderen Einflussgrößen festgehalten, so steigt Y im Durchschnitt um β_k Einheiten.

Beispiel: Festangestellte und freie Mitarbeiter

FAM: Anzahl festangestellter Mitarbeiter
FM: Anzahl freier Mitarbeiter
AT: Auflage in Tausend

$$
\begin{aligned}
FAM &= \alpha + \beta_1 \cdot FM + \beta_2 \cdot AT + \varepsilon_1 \\
FAM &= 31 + 0.092 \cdot FM + 0.32 \cdot AT + \varepsilon_1 \\
FAM &= 67 + 0.17 \cdot FM + \varepsilon_2
\end{aligned}
$$

Der Zusammenhang zwischen *FAM* und *FM* wird bei Berücksichtigung von *AT* geringer.

Zusammenfassung multiples Regressionsmodell

Das multiple Regressionsmodell ist nützlich, um Zusammenhänge zwischen Merkmalen zu analysieren.
Es ermöglicht:

- Quantifizierung des Zusammenhangs
- Berücksichtigung von Störgrößen
- Auswahl von relevanten Einflussgrößen

9 Regression und Mittelwertsvergleiche

Regression für nominale Einflussgrößen

Motivation

Bisher wurden bei der linearen Regression die Merkmale Y und X als quantitativ stetig vorausgesetzt.
Im folgenden Abschnitt soll aufgezeigt werden, wie eine lineare Regression bei einem Regressor X mit *nominalem Skalenniveau* modelliert und ausgewertet wird.

Beispiele

Häufig will man Einflüsse von folgenden Variablen analysieren:

Geschlecht männlich, weiblich

Familienstand ledig, verheiratet, geschieden, verwitwet

Staatsangehörigkeit Deutschland, Österreich, Schweiz, ...

Grundproblem

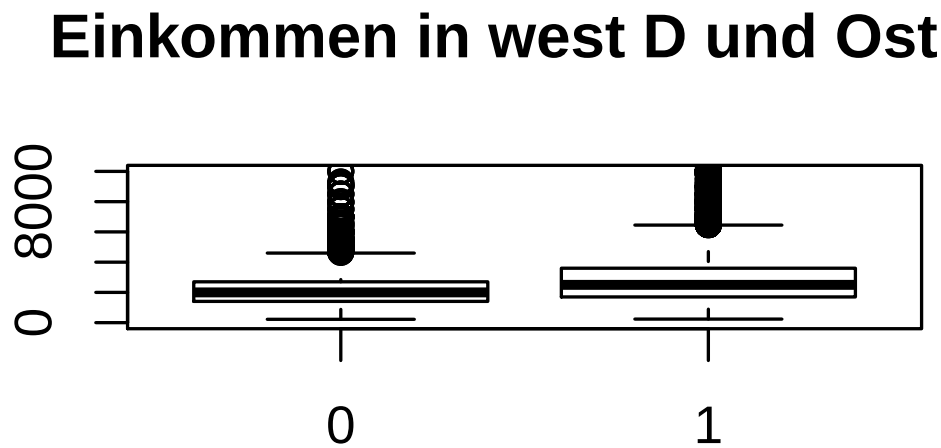
Standard Regression

Die kodierten Merkmalsausprägungen (z.B. 'ledig' $= 1$, 'verheiratet' $= 2$, 'geschieden' $= 3$) können nicht wie reelle Zahlen in die Berechnung der der Parameterschätzungen $\hat{\alpha}$ und $\hat{\beta}$ einbezogen werden, da

- nicht notwendiger Weise eine *Ordnung* zugrunde liegt und
- *Abstände* nicht definiert sind.

Einfacher Spezialfall: Binäre Einflussgröße

Beispiel: Einkommen Deutschland Vergleich Ost / West
SOEP-Daten 2007



Mittelwert West: 2956 Mittelwert Ost: 2245 Euro

Darstellung durch Regression

Mittelwert West: 2956 Mittelwert Ost: 2245 Euro

$$Y = \alpha + \beta X$$

Y: Einkommen
X=1 für West X=0 für Ost
KQ-Schätzung:

$$\begin{aligned} \hat{\alpha} &= 2245 \\ \hat{\beta} &= 711 \\ R^2 &= 0.028 \\ \sigma_E &= 1804 \end{aligned}$$

Mittelwert von X =0 :$\hat{\alpha}$
Mittelwertsunterschied: $\hat{\beta}$

Quadratsummenzerlegung

Residuenquadratsumme

$$\begin{aligned} SSE &= (y_i - \hat{y}_i)^2 = \sum_{ost}(y_i - (\hat{\alpha} + 0))^2 + \sum_{west}(y_i - (\hat{\alpha} + \beta))^2 \\ &= \sum_{ost}(y_i - \bar{y}_{ost})^2 + \sum_{west}(y_i - \bar{y}_{west})^2 \end{aligned}$$

Dies entspricht der Quadratsumme innerhalb der Gruppen (SSwithin)

$$\begin{aligned} SSM &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{ost}(\hat{\alpha} - \bar{y}) + \sum_{west}(\hat{\alpha} + \hat{\beta} - \bar{y})^2 \\ &= \sum_{ost}(\bar{y}_{ost} - \bar{y}) + \sum_{west}(\bar{y}_{west} - \bar{y})^2 \end{aligned}$$

Dies entspricht der Quadratsumme innerhalb der Gruppen
Interpretation von R^2: 2.8 Prozent der Streuung des Einkommens wird durch west/ost erklärt

Regression mit dichotomen (0-1) -Variablen

- Regression mit KQ–Schätzung möglich
- Einfache Regression entspricht Mittelwertsbildung
- Regressionskoeffizient entspricht Unterschied der Gruppenmittelwerte
- R^2 als Verhältnis der Streuung zwischen den Gruppen und der Gesamtstreuung

Korrelation zwischen dichotomen und stetigen Merkmalen

Punktbiseriale Korrelation

Der Korrelations-Koeffizient zwischen einem dichotomen und einem metrischem Merkmal ist sinnvoll berechenbar und lässt sich wie folgt darstellen :

$X \in \{0, 1\}$ Y metrisch

$$r_{XY} = \frac{\bar{Y}_1 - \bar{Y}_0}{S_Y} \cdot \sqrt{\frac{n_0 n_1}{N^2}}$$

$\bar{Y}_0$ Mittelwert bei $X = 0$,
$\bar{Y}_1$ Mittelwert bei $X = 1$

Entspricht normiertem Abstand der Gruppenmittelwerte.

Mehr als 2 Gruppen

Lösungsansatz

Hier ist eine direkte Lösung nicht sinnvoll.
Grundidee:

- aus einem nominalen Regressor mit k Merkmalsausprägungen
- $k - 1$ neue Regressoren (Dummys) gebildet werden.
- Eine Merkmalsausprägung des ursprünglichen Regressors wird zur *Referenzkategorie*.

Nominale Regressoren

Dummykodierung

Nach Wahl der Referenzkategorie $j \in \{1, ..., k\}$ ergeben sich die Dummys $X_i, i = 1, ..., k$ und $i \neq j$ mit folgenden Werten:

$$x_i = \begin{cases} 1 & \text{falls Kategorie } i \text{ vorliegt,} \\ 0 & \text{sonst.} \end{cases}$$

Nominale Regressoren

Beispiel

Gegeben seien folgende Daten:

lfd Nr.	Alter	Studienfach
1	19	BWL
2	22	Sonstige
3	20	VWL
⋮	⋮	⋮

Mit der Kodierung BWL = 1, VWL = 2, Sonstige = 3 erhalten wir bei Wahl der Referenzkategorie = 3 (Sonstige) zwei Dummys X_1 (für BWL) und X_2 (für VWL) gemäß folgendem Schema:

Ausprägung von X	Wert von X_1	Wert von X_2
1 BWL	1	0
2 VWL	0	1
3 Sonstige	0	0

Nominale Regressoren

Beispiel Fortsetzung

Aus der ursprünglichen Erhebung

lfd Nr.	Alter	Studienfach
1	19	BWL
2	22	Sonstige
3	20	VWL
⋮	⋮	⋮

ergibt sich somit der für die Auswertung geeignete Datensatz:

Dummykodierung

lfd Nr.	y	x_1	x_2
1	19	1	0
2	22	0	0
3	20	0	1
⋮	⋮	⋮	⋮

Nominale Regressoren

Modellierung

Nach der Kodierung kann nun ein Regressionsmodell aufgestellt werden:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Die Parameter $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$ lassen sich wie bei der Regression zweier stetiger Merkmale schätzen.

Nominale Regressoren

Berechnung

Um die angepassten Werte $\hat{y}$ für die jeweilige Merkmalsausprägung zu erhalten, werden die Dummyvariablen X_1 und X_2 entsprechend der gewählten Kodierung gesetzt (hier die Werte vom Beispiel):

Ausprägung	Dummykodierung
BWL	$\hat{y} = \hat{\alpha} + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0$ $= \hat{\alpha} + \hat{\beta}_1$
VWL	$\hat{y} = \hat{\alpha} + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1$ $= \hat{\alpha} + \hat{\beta}_2$
Sonstige	$\hat{y} = \hat{\alpha} + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0$ $= \hat{\alpha}$

Nominale Regressoren

Interpretation der Ergebnisse

- $\hat{\alpha}$ ist der Mittelwert der *Referenzkategorie*
- $\hat{\beta}_1, \hat{\beta}_2$ bilden die *Abweichungen* der Mittelwerte der übrigen Kategorien zur Referenzkategorie ab

Streuungszerlegung

Dann gilt:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{j=1}^{r} n_j \bar{x}_j \\ s^2 &= \frac{1}{n}\sum_{j=1}^{r} n_j s_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2 \end{aligned}$$

$$\text{Gesamtstreuung} = \begin{array}{c}\text{Streuung}\\ \text{innerhalb}\\ \text{der Gruppen}\end{array} + \begin{array}{c}\text{Streuung}\\ \text{zwischen}\\ \text{den Gruppen}\end{array}$$

Quadratsummenzerlegung: Varianzanalyse

Durch das nominale Merkmal werden die Daten in Gruppen aufgeteilt. Umindizierung: y_{ij} ist die i-te Beobachtung in der Gruppe j. Residuenquadratsumme

$$SSE = \sum_j \sum_i (y_{ij} - \hat{y}_{ij})^2 = \sum_j \sum_i (y_{ij} - \bar{y}_{.j})^2$$

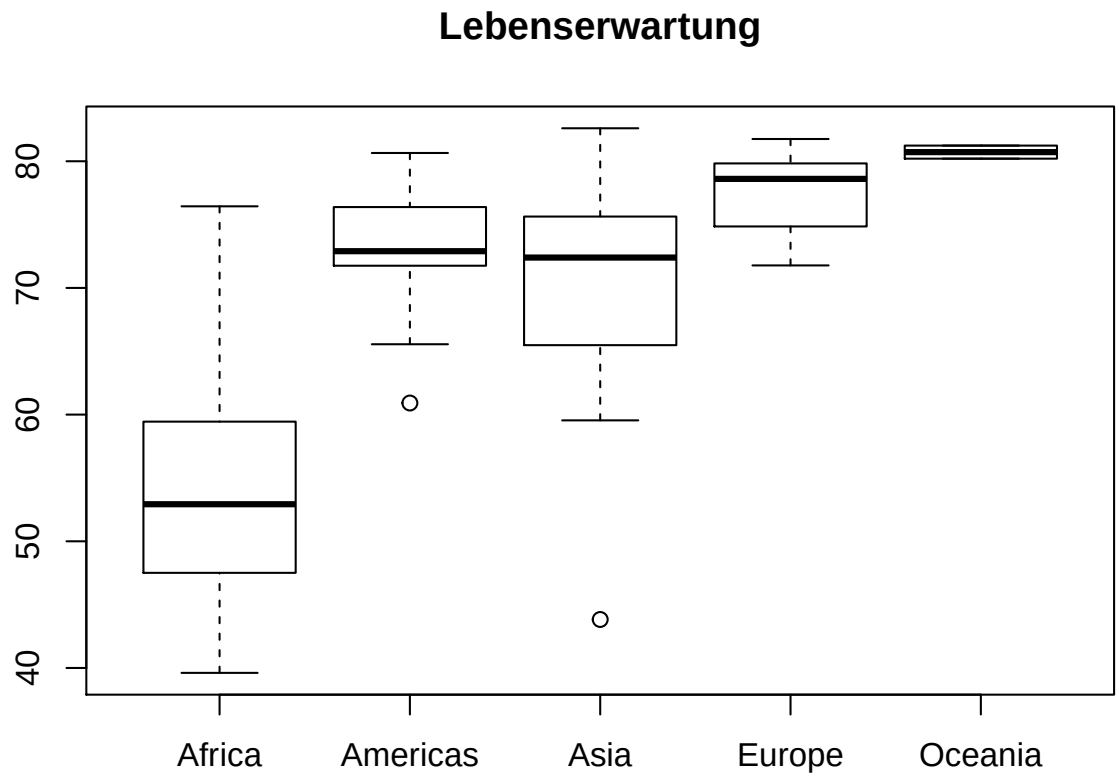
Dies entspricht der Quadratsumme innerhalb der Gruppen (SSwithin)

$$SSM = \sum_j \sum_i (\hat{y}_{ij} - \bar{y})^2 = \sum \sum (\bar{y}_{.j} - \bar{y})$$

Dies entspricht der Quadratsumme zwischen den Gruppen
Interpretation von R^2: Prozent der Streuung durch die nominale Variable

Beispiel: Lebenswerwartung und Kontinent

5 Gruppen von Ländern



Mittelwerte: Africa 54.81, Amerika 73.61 Asien: 70.73 Europa 77.65 Ozeanien 80.72

Regression

Mittelwerte: Africa 54.81, Amerika 73.61 Asien: 70.73 Europa 77.65 Ozeanien 80.72

$$Y = \alpha + \beta_1 X_1 + \cdots \beta_4 X_4$$

Referenz Afrika, X_1 Dummy für Amerika, X_2 Dummy für Asien usw.
Ergebnisse Regression :

$$\begin{aligned} \hat{\alpha} &= 54.81 \\ \hat{\beta}_1 &= 18.80 \\ \hat{\beta}_2 &= 15.92 \\ \hat{\beta}_3 &= 22.84 \\ \hat{\beta}_4 &= 25.91 \\ R^2 &= 0.64 \end{aligned}$$

Zusammenfassung: Regression mit nominalen Einflussgrößen

- nominales Merkmal kann als Gruppierungsvariable gesehen werden
- Einfluss nominaler Größen entspricht Vergleich der Gruppenmittelwerte
- Manchmal wird auch von Varianzanlyse (ANOVA) gesprochen
- Regressionsschätzung durch Dummy-Kodierung
- Quadratsummenzerlegung in Streuung innerhalb und zwischen den Gruppen
- Erweiterung zu multiplem Modell möglich

10 Verhältniszahlen und Indizes

Verhältniszahlen

Definition Verhältniszahlen

Verhältniszahlen entstehen durch *Quotientenbildung* aus

- zwei Maßzahlen
- den Ausprägungen zweier extensiver Merkmale (d.h. Merkmale, bei denen Summenbildung sinnvoll ist)

Verhältniszahlen werden unterteilt in

- Gliederungszahlen
- Beziehungszahlen
- einfache Indexzahlen / Messzahlen

Gliederungs- & Beziehungszahlen

Definition Gliederungszahlen

Gliederungszahlen beziehen eine *Teilmenge* auf eine *übergeordnete Gesamtmenge*.
Die Gliederungszahlen können als Quoten oder als Quote $\times$ 100 in Prozent angegeben werden.

Definition Beziehungszahlen

Beziehungszahlen bilden den Quotienten aus zwei Maßzahlen oder Größen, die verschieden gemessen werden (also nicht Teilmengen von Gesamtmengen), aber in sachlich sinnvoller Beziehung zueinander stehen.
Es wird unterschieden in

Verursachungszahlen Bewegungsmassen bezogen auf Bestandsmassen

Entsprechungszahlen kein Bezug auf einen Bestand möglich

Gliederungs- & Beziehungszahlen

Beispiele Gliederungszahlen

$$\text{Erwerbsquote} = \frac{\text{Zahl der Erwerbspersonen}}{\text{Umfang der Bevölkerung}}$$

$$\text{Arbeitslosenquote} = \frac{\text{Zahl der Arbeitslosen}}{\text{Zahl der Erwerbspersonen}}$$

$$\text{Ausschussquote} = \frac{\text{Zahl der Ausschussteile}}{\text{Gesamtzahl der produzierten Teile}}$$

Beispiele Verursachungszahlen

$$\text{(rohe) Geburtenziffer} = \frac{\text{Lebendgeborene}}{\text{Bevölkerung}}$$

$$\text{(rohe) Sterbeziffer} = \frac{\text{Verstorbene}}{\text{Bevölkerung}}$$

Beispiele Entsprechungszahlen

$$\text{Bevölkerungsdichte} = \frac{\text{Einwohnerzahl}}{\text{Fläche in km}^2}$$

$$\text{Produktivität} = \frac{\text{Nettoproduktion}}{\text{Arbeitseinsatz}}$$

Indexzahlen

Definition einfache Indexzahlen

Die einfachen Indexzahlen beschreiben den Zusammenhang zwischen Ergebnissen für eine Maßzahl, gemessen zu verschiedenen Zeitpunkten der Entwicklung einer Grundgesamtheit.

Es liegt also eine Zeitreihe von Maßzahlen vor:

- x_0, Wert der Basiszahl in der Basisperiode
- x_t, Wert derselben Maßzahl in der Berichtsperiode

Die zugehörige Indexzahl berechnet sich dann aus dem Quotienten der Maßzahl der Berichtsperiode zur Maßzahl der Basisperiode:

$$I_{0t} = \frac{x_t}{x_0}$$

Indexzahlen

Beispiele für einfache Indexzahlen

$$\text{Preismesszahlen} = \frac{p_t}{p_0} = P_{0t} \quad \text{(Preisindex)} \tag{8.1}$$

oder

$$\text{Mengenmesszahlen} = \frac{q_t}{q_0} = Q_{0t} \quad \text{(Mengenindex)}\,. \tag{8.2}$$

Dabei ist p der Preis eines bestimmten Produkts und q die produzierte oder verkaufte Menge (quantity) dieses Produkts jeweils zur Basisperiode 0 bzw. zur Berichtsperiode t. Damit wird eine Zeitreihe von Messungen (Preise, Mengen) durch Bezug auf eine Basisperiode in gewisser Weise standardisiert oder bereinigt.

Indexzahlen

Angabe in Prozent

Indizes können nach Multiplikation mit 100 in Prozent angegeben werden:

$$I_{0t} = \frac{x_t}{x_0} \cdot 100\,\% .$$

Indexzahlen

Veränderung des Basisjahres

Bei längeren Zeitreihen kann es zu Strukturbrüchen kommen, die eine Umbasierung, d.h., die die Festlegung eines neuen Basiszeitpunktes erforderlich machen. Wählt man die neue Basisperiode k, so gilt:

$$I_{kt} = \frac{x_t}{x_k} = \frac{x_t \cdot x_0}{x_0 \cdot x_k} = \frac{\frac{x_t}{x_0}}{\frac{x_k}{x_0}} = \frac{I_{0t}}{I_{0k}}$$

Damit müssen bei Umbasierung einer Indexreihe, die vor dem neuen Basisjahr gemessen wurde, die vorangegangenen Daten $x_i (i = 1, ..., k-1)$ nicht bekannt sein. Es reicht aus, die Indexreihe $I_{01}, ..., I_{0k}$ zu kennen.

Verkettungsregel

$$I_{0t} = I_{0k} \cdot I_{kt}$$

Indexzahlen

Beispiel

t	q_t	$Q_{1985,t}$
1988	85	1.0625
1989	90	1.1250
1990	95	1.1875
1991	95	1.1875
1992	100	1.2500
1993	110	1.3750

Neues Basisjahr 1990:

$$I_{1990,1993} = \frac{110}{95} = 1.1579\,.$$

Die Verkettungsregel liefert z. B.

$$I_{1985,1993} = I_{1985,1990} \cdot I_{1990,1993} = 1.1875 \cdot 1.1579 = 1.3750\,.$$

Preisindizes

Einleitung

Im Unterschied zu den bisherigen Messzahlen werden in den folgenden Abschnitten sogenannte *zusammengesetzte Indexzahlen* betrachtet, die gleichartige Indexreihen für n verschiedene Güter verknüpfen.

Preisindizes

Definitionen

Seien n verschiedene Güter ausgewählt. Dann bezeichnet

$\mathrm{p}_0' = (p_0(1), ..., p_0(n))$ den Vektor der *Preise* dieser Güter in der *Basisperiode*

$\mathrm{p}_t' = (p_t(1), ..., p_t(n))$ den Vektor der *Preise* dieser Güter in der *Berichtsperiode*

$\mathrm{q}_0' = (q_0(1), ..., q_0(n))$ den Vektor der *Mengen* dieser Güter in der *Basisperiode*

$\mathrm{q}_t' = (q_t(1), ..., q_t(n))$ den Vektor der *Mengen* dieser Güter in der *Berichtsperiode*

Preisindizes

arithmetisches Mittel der Preismesszahlen

$$P_{0t} = \frac{1}{n} \sum_{i=1}^{n} I_{0t}^{p}(i)$$

- einfachste Möglichkeit
- unterschiedliche Gewichtung der Güter geht verloren

Preisindizes

gewichtetes arithmetisches Mittel der Preismesszahlen

$$P_{0t} = \frac{\frac{p_t(1)}{p_0(1)} w(1) + \cdots + \frac{p_t(n)}{p_0(n)} w(n)}{w(1) + \cdots + w(n)}$$

Durch Transformation der Gewichte

$$\tilde{w}(i) = \frac{w(i)}{\sum_k w(k)}, \sum_{i=1}^{n} \tilde{w}(i) = 1$$

ergibt sich die alternative Formel

$$P_{0t} = I_{0t}^{p}(1)\tilde{w}(1) + \cdots + I_{0t}^{p}(n)\tilde{w}(n)$$

Preisindex nach Laspeyres

Definition

Die Gewichtung der berücksichtigten Güter ist deren *Ausgabensumme* (Menge×Preis) jeweils aus der Basisperiode:

$$w(i) = \mathrm{p}_0(i)\mathrm{q}_0(i)$$

Damit gilt für den Laspeyres-Preisindex:

$$\begin{aligned} P_{0t}^{\mathrm{L}} &= \frac{\sum_{i=1}^{n} \mathrm{p}_t(i)\mathrm{q}_0(i)}{\sum_{i=1}^{n} \mathrm{p}_0(i)\mathrm{q}_0(i)} \\ &= \frac{\mathrm{p}_t'\mathrm{q}_0}{\mathrm{p}_0'\mathrm{q}_0} \end{aligned}$$

Also der Quotient aus dem *Wert des Warenkorbs der Basisperiode zu aktuellen Preisen* und *Wert des Warenkorbs der Basisperiode zu Basispreisen*.

Preisindex nach Laspeyres

Aussage

Der Preisindex nach Laspeyres gibt an, wie sich das Preisniveau geändert hat, wenn der Warenkorb der Basisperiode zum Vergleich herangezogen wird.

Vorteile

- Der Preisindex einer neu erhobenen Berichtsperiode ist sofort vergleichbar mit früher ermittelten Indizes.
- Leicht ermittelbar, da der Inhalt des Warenkorbs von früheren Untersuchungen bekannt ist.

Nachteile

- Die Zusammenstellung des Warenkorbs veraltet mit der Zeit. Darum muss dieser in regelmäßigen Abständen aktualisiert werden, um repräsentativ zu bleiben.

Preisindex nach Paasche

Definition

Die Gewichtung der berücksichtigten Güter ist deren *Ausgabensumme*, bestehend aus der Menge aus der Berichtsperiode und dem Preis aus der Basisperiode:

$$w(i) = \mathrm{p}_0(i)\mathrm{q}_t(i)$$

Damit gilt für den Paasche-Preisindex:

$$\begin{aligned} P_{0t}^{\mathrm{P}} &= \frac{\sum_{i=1}^{n} \mathrm{p}_t(i)\mathrm{q}_t(i)}{\sum_{i=1}^{n} \mathrm{p}_0(i)\mathrm{q}_t(i)} \\ &= \frac{\mathrm{p}_t'\mathrm{q}_t}{\mathrm{p}_0'\mathrm{q}_t} \end{aligned}$$

Also der Quotient aus dem *Wert des Warenkorbs der Berichtsperiode zu aktuellen Preisen* und *Wert des Warenkorbs der Berichtsperiode zu Basispreisen*.

Preisindex nach Paasche

Aussage

Der Preisindex nach Paasche gibt an, wie sich das Preisniveau geändert hat, wenn der Warenkorb der Berichtsperiode zum Vergleich herangezogen wird.

Vorteile

- Die Zusammenstellung des Warenkorbs ist stets aktuell.

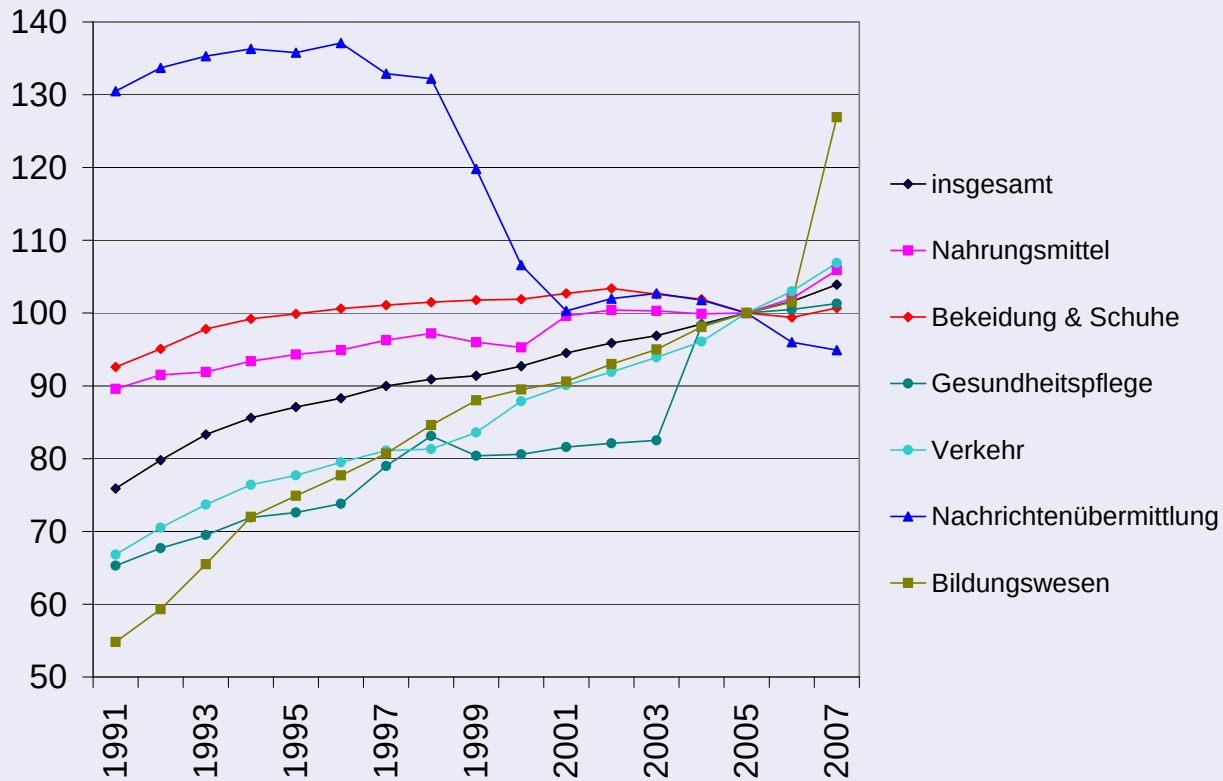
Nachteile

- Der Preisindex einer neu erhobenen Berichtsperiode lässt sich mit früheren Perioden nur vergleichen, wenn diese auf den neuen Warenkorb umgestellt werden.
- Für jedes neue Berichtsjahr muss ein neuer Warenkorb zusammengestellt werden.

Preisindex

Beispiel

Verbraucherpreisindex von Deutschland 1991 bis 2007
Bezugsjahr: 2005



Quelle: Statistisches Bundesamt, Stand: 10.06.2008

Preisindizes

Beispiel

Gut	Preise		Mengen	
i	$p_0(i)$	$p_t(i)$	$q_0(i)$	$q_t(i)$
1	4	6	5	4
2	6	8	10	15
3	10	12	8	16

$$
\begin{aligned}
\mathbf{p}_0' &= (4, 6, 10) \quad \text{(Basispreise)} \\
\mathbf{p}_t' &= (6, 8, 12) \quad \text{(aktuelle Preise)} \\
\mathbf{q}_0' &= (5, 10, 8) \quad \text{(Basiswarenkorb)} \\
\mathbf{q}_t' &= (4, 15, 16) \quad \text{(aktueller Warenkorb)}
\end{aligned}
$$

Preisindizes

Beispiel

Preisindex nach Laspeyres:

$$\begin{aligned} P_{0t}^{\mathsf{L}} &= \frac{\mathbf{p}_t'\mathbf{q}_0}{\mathbf{p}_0'\mathbf{q}_0} = \frac{6 \cdot 5 + 8 \cdot 10 + 12 \cdot 8}{4 \cdot 5 + 6 \cdot 10 + 10 \cdot 8} \\ &= \frac{206}{160} = 1.2875\,. \end{aligned}$$

Preisindex nach Paasche:

$$\begin{aligned} P_{0t}^{\mathsf{P}} &= \frac{\mathbf{p}_t'\mathbf{q}_t}{\mathbf{p}_0'\mathbf{q}_t} = \frac{6 \cdot 4 + 8 \cdot 15 + 12 \cdot 16}{4 \cdot 4 + 6 \cdot 15 + 10 \cdot 16} \\ &= \frac{336}{266} = 1.2632\,. \end{aligned}$$

Mengenindizes

Einleitung

Vertauscht man die Rolle von Preisen und Mengen in den beiden Preisindizes, so erhält man Mengenindizes, die die Änderung des Warenkorbs über die Zeit angeben, bewertet mit den Preisen einer bestimmten Periode.

Mengenindex nach Laspeyres

Definition

Der Mengenindex nach Laspeyres verwendet die Preise der Basisperiode und ist definiert als

$$Q_{0t}^{\mathrm{L}} = \frac{\mathrm{p}_0' \mathrm{q}_t}{\mathrm{p}_0' \mathrm{q}_0}$$

Aussage

Q_{0t}^{L} gibt das Verhältnis an, in dem sich der Wert des Warenkorbs von der Basis- zur Berichtsperiode – bewertet mit Preisen der Basisperiode – durch Veränderung der Mengen geändert hat.

Mengenindex nach Paasche

Definition

Der Mengenindex nach Paasche verwendet die Preise der Berichtsperiode und ist definiert als

$$Q_{01}^{\mathrm{P}} = \frac{\mathrm{p}_t' \mathrm{q}_t}{\mathrm{p}_t' \mathrm{q}_0}$$

Aussage

Q_{0t}^{P} gibt die Veränderung des Wertes des Warenkorbs an, wobei zur Bewertung die Preise der Berichtsperiode verwendet werden.

Umsatzindex

Definition

Der Umsatzindex ergibt sich aus dem Produkt der *Preise und Mengen der Berichtsperiode* geteilt durch das Produkt der *Preise und Mengen der Basisperiode*.

$$W_{0t} = \frac{\mathrm{p}_t' \mathrm{q}_t}{\mathrm{p}_0' \mathrm{q}_0}$$

Aussage

W_{0t} gibt die Veränderung des Wertes des Warenkorbs der Berichtsperiode im Verhältnis zum Wertes des Warenkorbs der Basisperiode an.

Verknüpfung von Indizes

(Laspeyres-Preisindex) $\times$ (Paasche-Mengenindex)

$$P_{0t}^{\mathrm{L}} \cdot Q_{0t}^{\mathrm{P}} = \frac{(\mathrm{p}_t'\mathrm{q}_0)}{\mathrm{p}_0'\mathrm{q}_0} \cdot \frac{\mathrm{p}_t'\mathrm{q}_t}{(\mathrm{p}_t'\mathrm{q}_0)} = \frac{\mathrm{p}_t'\mathrm{q}_t}{\mathrm{p}_0'\mathrm{q}_0} = W_{0t}$$

(Paasche-Preisindex) $\times$ (Laspeyres-Mengenindex)

$$P_{0t}^{\mathrm{P}} \cdot Q_{0t}^{\mathrm{L}} = \frac{(\mathrm{p}_t'\mathrm{q}_t)}{\mathrm{p}_0'\mathrm{q}_t} \cdot \frac{\mathrm{p}_0'\mathrm{q}_t}{(\mathrm{p}_0'\mathrm{q}_0)} = \frac{\mathrm{p}_t'\mathrm{q}_t}{\mathrm{p}_0'\mathrm{q}_0} = W_{0t}$$

Matrizensymbolik

Gestaltung

Statt des allgemeinen Index t wird die Berichtsperiode durch die 1 dargestellt:

Preisindex nach Laspeyres	$P_{01}^{L} = \frac{p_1' q_0}{p_0' q_0}$	$\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right.$
Preisindex nach Paasche	$P_{01}^{P} = \frac{p_1' q_1}{p_0' q_1}$	$\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right.$
Mengenindex nach Laspeyres	$Q_{01}^{L} = \frac{p_0' q_1}{p_0' q_0}$	$\left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right.$
Mengenindex nach Paasche	$Q_{01}^{P} = \frac{p_1' q_1}{p_1' q_0}$	$\left\{ \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right.$
Umsatzindex	$W_{01} = \frac{p_1' q_1}{p_0' q_0}$	$\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right.$

Spezielle Probleme

Erweiterung des Warenkorbs (Preisindex nach Laspeyres)

t': Zeitpunkt der Einführung der neuen Ware (Nummer $(n+1)$).
Man berechnet zuerst den Preisindex nach Laspeyres:

$$P_{0t'}^{\mathsf{L}} = \frac{\mathbf{p}_{t'}'\mathbf{q}_0}{\mathbf{p}_0'\mathbf{q}_0}\,.$$

Danach berechnet man den Index für $(t', t'+1)$:

$$P_{t',t'+1}^{\mathsf{L}}(\text{erweitert}) = \frac{\mathbf{p}_{t'+1}'\mathbf{q}_0 + p_{t'+1}(n+1)q_{t'}(n+1)}{\mathbf{p}_{t'}'\mathbf{q}_0 + p_{t'}(n+1)q_{t'}(n+1)}\,.$$

Da $p_0(n+1)$ und $q_0(n+1)$ nicht existieren, wird die Formel von Laspeyres für 0 als Basisperiode dahingehend abgewandelt, dass man $p_{t'}(n+1)$ und $q_{t'}(n+1)$ verwendet.
Der verkettete Index lautet schließlich

$$P_{0,t'+1}^{\mathsf{L}}(\text{verkettet}) = P_{0,t'}^{\mathsf{L}} P_{t',t'+1}^{\mathsf{L}}(\text{erweitert})\,.$$

Erweiterung des Warenkorbs

Beipiel

Periode	Damenkostüme		Herrenanzüge		Trainingsanzüge	
t	p_t	q_t	p_t	q_t	p_t	q_t
0	300	10	40	20	–	–
1	400	15	50	25	–	–
2	500	17	60	25	300	10
3	400	18	50	30	400	20

Kleiner Warenkorb (Damenkostüme und Herrenanzüge):

$$P_{02}^{\mathsf{L}} = \frac{\mathbf{p}_2'\mathbf{q}_0}{\mathbf{p}_0'\mathbf{q}_0} = \frac{500 \cdot 10 + 60 \cdot 20}{300 \cdot 10 + 40 \cdot 20}$$

$$= \frac{6\,200}{3\,800} = 1.6316\,.$$

Erweiterung des Warenkorbs

Beipiel

Für den Übergang von Periode 2 auf 3 berechnen wir:

$$\begin{aligned} P_{23}^{\mathsf{L}}(\text{erweitert}) &= \frac{\mathbf{p}_3'\mathbf{q}_0 + p_3(3)q_2(3)}{\mathbf{p}_2'\mathbf{q}_0 + p_2(3)q_2(3)} \\ &= \frac{(400 \cdot 10 + 50 \cdot 20) + 400 \cdot 10}{(500 \cdot 10 + 60 \cdot 20) + 300 \cdot 10} \\ &= \frac{5\,000 + 4\,000}{6\,200 + 3\,000} = \frac{9\,000}{9\,200} = 0.9783\,. \end{aligned}$$

Damit gilt schließlich

$$P_{03}^{\mathsf{L}}(\text{verkettet}) = 1.6316 \cdot 0.9783 = 1.5962\,.$$

Substitution einer Ware

Beispiel

		Perioden				
	$q_0(i)$ $\times 10\,000$	0	1	2	3	4
		Preise $p_t(i)$				
Radios	1	400	420	430	440	450
S.W.-TV	2	2 000	1 900	1 800	–	–
Farb-TV	–	–	–	3 000	3 500	4 200

Wir verwenden die Preissteigerungen für Farbfernsehgeräte, um die Preise der alten Ware Schwarzweiß-Fernsehgeräte fortzuschreiben.

$$\tilde{p}_3(\text{S.W.-TV}) = 1\,800 \cdot \frac{3\,500}{3\,000} = 2\,100$$

$$\tilde{p}_4(\text{S.W.-TV}) = 1\,800 \cdot \frac{4\,200}{3\,000} = 2\,520\,.$$

Substitution einer Ware

Beispiel

Damit können wir mit dem alten Warenkorb weiterrechnen. Wir erhalten dann die verketteten Reihen

	$q_0(i)$ $\times 10\,000$	0	1	2	3	4
Radios	1	400	420	430	440	450
TV	2	2 000	1 900	1 800	2 100	2 520
Wert ($\times$ 10 000)		4 400	4 220	4 030	4 640	5 490
P_{0t}^{L}		1.000	0.959	0.916	1.055	1.248

Subindizes

Beispiel

Ein Warenkorb bestehe aus zwei Subkörben, Korb I und Korb II. Die zugehörigen Warenmengen sind $\mathbf{q}_I' = (q_1, \ldots, q_m)$ und $\mathbf{q}_{II}' = (q_{m+1}, \ldots, q_n)$. Die Laspeyres-Preisindizes für die beiden Subkörbe lauten

$$\begin{aligned} P_{0t}^{\mathsf{L}}(I) &= \frac{\sum_{i=1}^{m} p_t(i) q_0(i)}{\sum_{i=1}^{m} p_0(i) q_0(i)}, \\ P_{0t}^{\mathsf{L}}(II) &= \frac{\sum_{i=m+1}^{n} p_t(i) q_0(i)}{\sum_{i=m+1}^{n} p_0(i) q_0(i)}. \end{aligned}$$

Der Gesamtumsatz zur Basisperiode ist

$$U = \sum_{i=1}^{n} p_0(i) q_0(i).$$

Subindizes

Beispiel

Damit sind die Umsatzanteile bezogen auf die Basisperiode

$$\begin{aligned} w^{I} &= \frac{\sum_{i=1}^{m} p_0(i) q_0(i)}{U}, \\ w^{II} &= 1 - w^{I} = \frac{\sum_{i=m+1}^{n} p_0(i) q_0(i)}{U}. \end{aligned}$$

Der Gesamtindex ist dann

$$P_{0t}^{\mathsf{L}} = w^{I} P_{0t}^{\mathsf{L}}(I) + w^{II} P_{0t}^{\mathsf{L}}(II)\,.$$

Subindizes

Numerisches Beispiel

	Korb I				Korb II	
	Damenkostüme		Herrenanzüge		Trainingsanzüge	
t	p_t	q_t	p_t	q_t	p_t	q_t
0	400	1	500	1	300	1
1	420	1	550	2	320	1
2	450	2	600	3	340	2
3	500	2	650	4	360	2

Der Gesamtumsatz im Basisjahr ist die Summe der Umsätze von Korb I und Korb II. Er beträgt

$$U = (400 \cdot 1 + 500 \cdot 1) + (300 \cdot 1) = 1200\,.$$

Die Umsatzanteile zum Basisjahr sind damit

$$w^I = (400 \cdot 1 + 500 \cdot 1)/1\,200 = 0.75$$

und

$$w^{II} = (300 \cdot 1)/1\,200 = 0.25\ .$$

Subindizes

Numerisches Beispiel

Die Teilindizes sind:

	Korb I			Korb II	
t	$\sum p_t(i)q_0(i)$		$P_{0t}^{\mathrm{L}}(I)$	$p_t(3)q_0(3)$	$P_{0t}^{\mathrm{L}}(II)$
0	$400 \cdot 1 + 500 \cdot 1 =$	900	1.000	300	1.000
1	$420 \cdot 1 + 550 \cdot 1 =$	970	1.078	320	1.067
2	$450 \cdot 1 + 600 \cdot 1 =$	1 050	1.167	340	1.133
3	$500 \cdot 1 + 650 \cdot 1 =$	1 150	1.278	360	1.200

Die Gesamtindizes für die einzelnen Zeitpunkte sind damit:

$$
\begin{aligned}
P_{00}^{\mathrm{L}} &= 1.000 \cdot 0.75 + 1.000 \cdot 0.25 = 1.0000 \\
P_{01}^{\mathrm{L}} &= 1.078 \cdot 0.75 + 1.067 \cdot 0.25 = 1.0753 \\
P_{02}^{\mathrm{L}} &= 1.167 \cdot 0.75 + 1.133 \cdot 0.25 = 1.1585 \\
P_{03}^{\mathrm{L}} &= 1.278 \cdot 0.75 + 1.200 \cdot 0.25 = 1.2585
\end{aligned}
$$

11 Zeitreihen

Zeitreihen

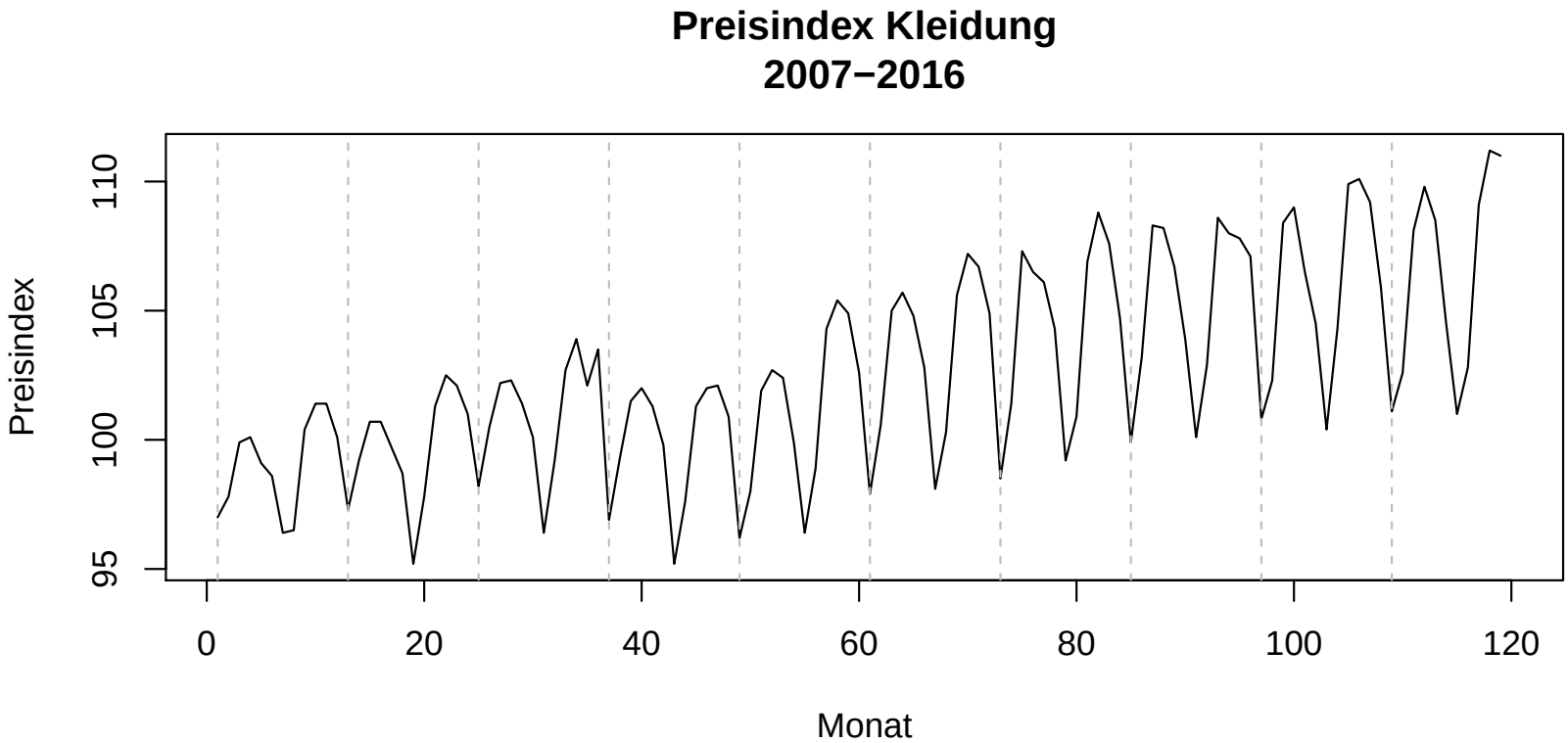
Motivation

In vielen Anwendungen wird ein Merkmal über die Zeit wiederholt beobachtet. Die zeitliche Entwicklung kann dann als *Kurvendiagramm* dargestellt werden. Auf der vertikalen Achse werden die Merkmalsausprägungen abgetragen, die horizontale Achse ist die Zeitachse. Übliche Zeitachsen sind:

- Tage, z.B. Aktienindex, Temperatur, Niederschlagsmenge
- Monate
- Quartale
- Jahre

In diesem Abschnitt werden Methoden zur beschreibenden Analyse vorgestellt. Wir gehen von *äquidistanten* Zeitreihen aus.

Preisindex Kleidung



Zerlegung von Zeitreihen

Komponentenmodell

Die Beobachtungen y_t, $t = 1, \ldots, T$ werden als Summe verschiedener Einzelkomponenten aufgefasst:

- Grundbestandteil ist die *glatte Komponente* g_t, die die langfristige Entwicklung (*Trend*) beschreibt.
- Saisonale Schwankungen (Quartale, Monate) werden durch die saisonale Komponente s_t wiedergegeben.
- Die Differenz zwischen der beobachteten Reihe y_t und dem durch g_t und s_t modellierten Anteil wird in der *irregulären Komponente* oder *Restkomponente* erfasst, die im Mittel 0 sein soll.

Additives und multiplikatives Komponentenmodell

Additives und multiplikatives Modell

Das additive Modell lautet

$$y_t = g_t + s_t + r_t \ , \quad t = 1, \ldots, T \ ,$$

unter der Nebenbedingung $\sum r_t = 0$ (streng genommen: Erwartungswert der r_t soll 0 sein, siehe induktive Statistik). Oftmals ist es besser, einen multiplikativen Ansatz zu verfolgen:

$$\tilde{y}_t = \tilde{g}_t \cdot \tilde{s}_t \cdot \tilde{r}_t \ ,$$

unter der Nebenbedingung $\prod \tilde{r}_t = 1$ (Erwartungswert der $\tilde{r}_t$ soll 1 sein). Durch Logarithmieren erhält man

$$y_t = \log(\tilde{y}_t), \ g_t = \log(\tilde{g}_t), \ s_t = \log(\tilde{s}_t), \ r_t = \log(\tilde{r}_t) \ ,$$

und man kann damit das multiplikative Modell wiederum in additiver Form schreiben.

Schätzung des Komponentenmodells

- Es werden zunächst keinerlei Angaben gemacht, wie die einzelnen Komponenten modelliert und geschätzt werden sollen.
- Identifikation: Oftmals hat man noch eine zyklische Komponente (z.B. 7-jähriger Konjunkturzyklus). Es ist dann oftmals schwierig, die einzelnen Komponenten zu berechnen
- Es gibt verschiedene Strategien, die Komponenten zu schätzen

Gleitende Durchschnitte: Motivation

Ziele

- Glättung der Zeitreihe: die Zeitreihe, die man nach Anwendung der gleitenden Durchschnitte erhält, hat geringere Variabilität. Damit läßt sich ein eventuell vorhandener Trend besser erkennen bzw. schätzen
- Im Komponentenmodell: Schätzung der glatten Komponente g_t durch Herausfiltern der saisonalen Schwankungen s_t.

Gleitende Durchschnitte: Berechnung

Definition: gleitender Durchschnitt ungerader Ordnung

Unter einem gleitenden Durchschnitt der ungeraden Ordnung $2k+1$ ($k = 0, 1, 2, \ldots$) für den Zeitreihenwert y_t verstehen wir das arithmetische Mittel

$$y_t^* = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j} \, .$$

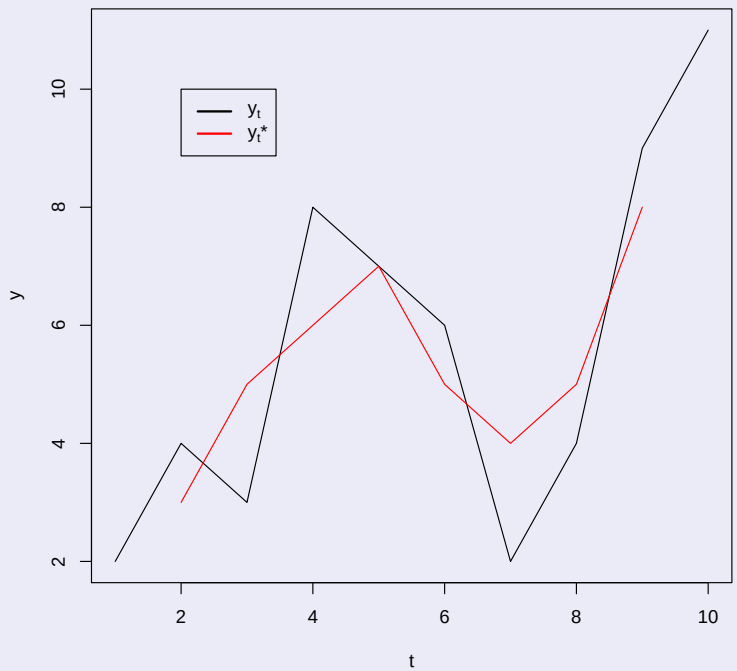
Wir mitteln über die k vor dem Zeitpunkt t liegenden Werte, den Wert y_t selbst und über die k nach dem Zeitpunkt t liegenden Werte. Beispiel mit $t = 7$, $k = 1$ (Ordnung 3):

$$y_7^* = \frac{1}{3}(y_6 + y_7 + y_8)$$

Gleitende Durchschnitte: Beispiel

Zahlenbeispiel bei ungerader Ordnung mit $k = 1$

t	1	2	3	4	5	6	7	8	9	10
y_t	2	4	3	8	7	6	2	4	9	11
y_t^*	-	3	5	6	7	5	4	5	8	-



$$y_6^* = \frac{1}{3}(y_5 + y_6 + y_7) = \frac{1}{3}(7 + 6 + 2) = 5$$

Gleitende Durchschnitte: Berechnung

Definition: gleitender Durchschnitt gerader Ordnung

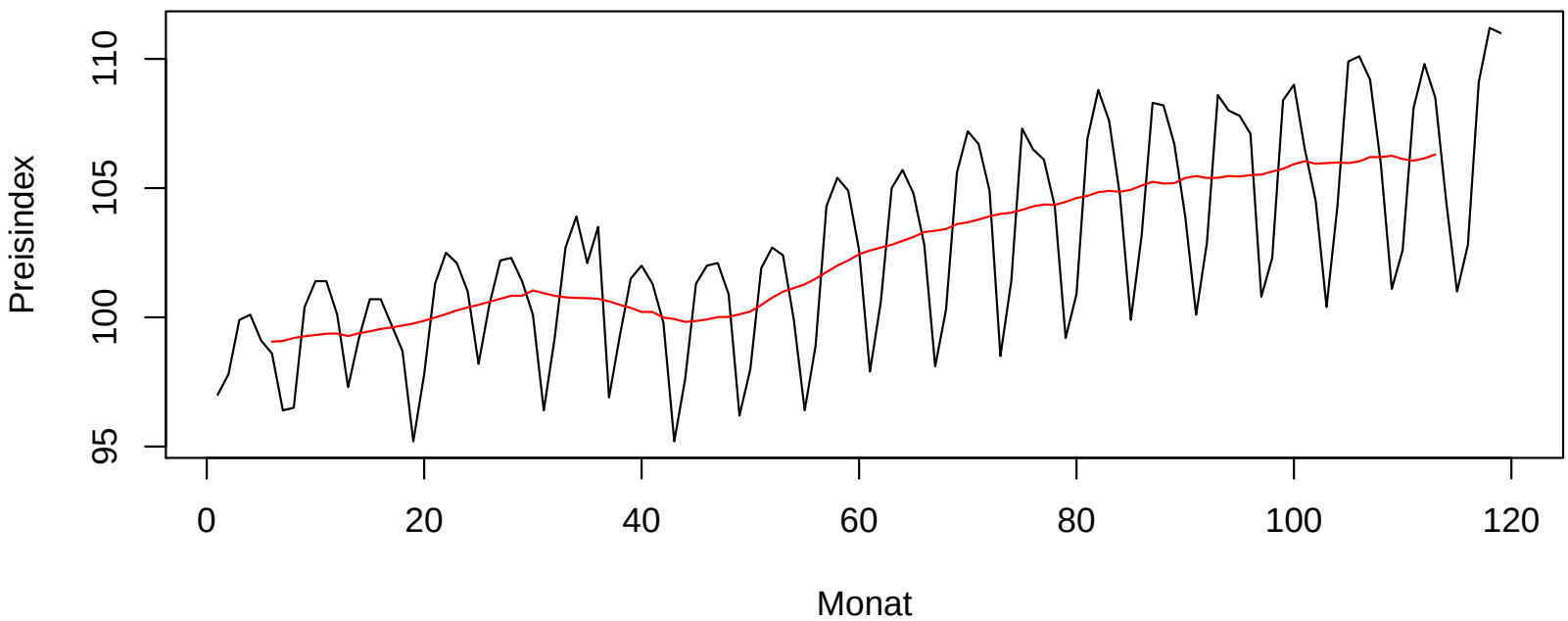
Unter einem gleitenden Durchschnitt der geraden Ordnung $2k$ ($k = 0, 1, 2, \ldots$) für den Zeitreihenwert y_t verstehen wir das arithmetische Mittel

$$y_t^* = \frac{1}{2k}\left(\frac{1}{2}y_{t-k} + \sum_{j=-k+1}^{k-1} y_{t+j} + \frac{1}{2}y_{t+k}\right) .$$

Hier werden die gleichen Beobachtungswerte wie bei ungerader Ordnung berücksichtigt, jedoch gehen die Randwerte nur mit halbem Gewicht ein.

Gleitende Durchschnitte

Beispiel Kleidungsindex bei gerader Ordnung mit $k = 6$



Konstante Saisonfigur

Modell

Man betrachtet das (additive) Modell

$$y_t = g_t + s_t + r_t \ , \quad t = 1, \ldots, T \ .$$

Man spricht von einer konstanten Saisonfigur mit Periode p, falls

$$s_t = s_{t+p} \ .$$

Es soll dann stets gelten:

$$\sum_{j=0}^{p-1} s_{t+j} = 0$$

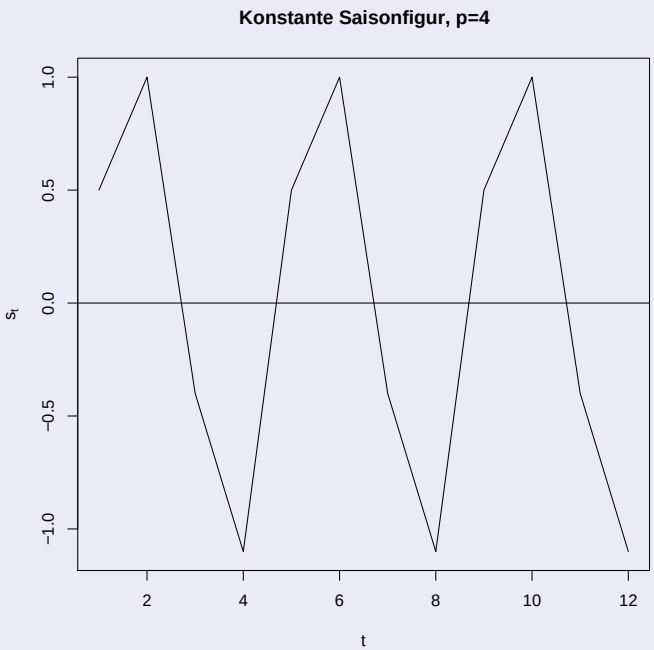
Konstante Saisonfigur

Beispiel $p = 4$

$$s_t = 0.5, \quad s_{t+1} = 1, \quad s_{t+2} = -0.4, \quad s_{t+3} = -1.1$$

Es gilt dann

$$\sum_{j=0}^{p-1} s_{t+j} = 0.5 + 1 + (-0.4) + (-1.1) = 0 \ .$$



Zerlegung in Trend und Saison

Vorgehen

Wir verstehen die saisonale Komponente als sich regelmäßig wiederholende Schwankungen um die glatte Komponente der Zeitreihe. Bilden wir nun gleitende Durchschnitte der Ordnung $2k = l \cdot p$ $(l = 1, 2, \ldots)$, so erhalten wir

$$y_t^* = g_t^* + s_t^* + r_t^* = g_t^* + r_t^* \,.$$

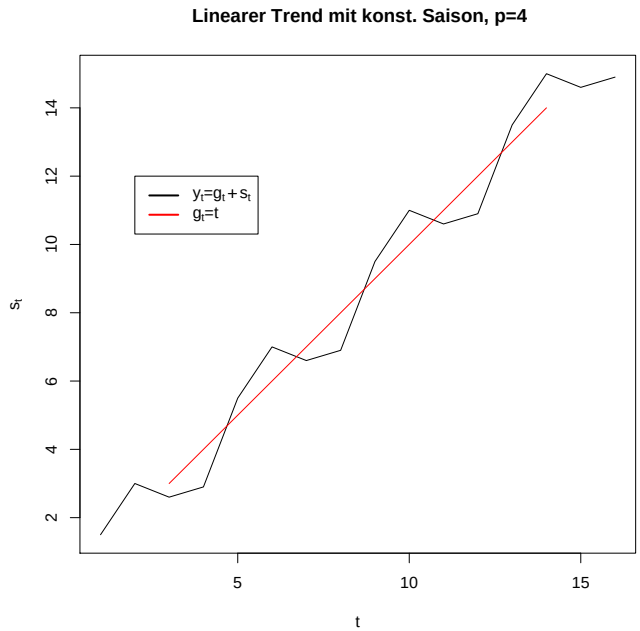
Die saisonale Komponente entfällt durch die Glättung, da wegen $\sum_{j=0}^{p-1} s_{t+j} = 0$ gilt: $s_t^* = 0$. Wir haben dadurch mit y_t^* wieder eine Schätzung für die glatte Komponente g_t erhalten.
$\Rightarrow$ Gleitende Durchschnitte $\rightarrow$ „Filtermethode“

Beispiel für Trend und Saison

Saisonfigur wie oben $(0.5, 1.0, -0.4, -1.1)$, linearer Trend, also

$$y_t = g_t + s_t = \mathbf{t} + s_t \ , t = 1, \ldots, 12$$

Gleitender Durchschnitt mit $k = 2, p = 4$ liefert den linearen Trend



Zerlegung in Trend und Saison

Schätzung der Saison

- Gleitende Durchschnitte mit $2k = l \cdot p$, also y_t^*, bilden. y_t^* ist dann (inklusive Fehler) im Wesentlichen die Trendkomponente $g_t^* + r_t^*$.
- Bilde die Differenz

$$d_t = y_t - y_t^*$$

Zerlegung in Trend und Saison

Schätzung der Saison (Fortsetzung)

- Die d_t entsprechen dann, bis auf die zusätzlichen Fehler, der Saisonkomponente, d.h.

$$d_t \approx d_{t+p} \ .$$

Damit werden für jede „Saison“ Mittelwerte gebildet, also z.B. der Mittelwert über alle Werte des ersten Quartals, der Mittelwert über alle Werte des zweiten Quartals, etc. Man erhält also $\bar{d}_1, \ldots, \bar{d}_p$. Diese Mittelwerte werden nochmals zentriert, damit die Summe aller Saisonkomponenten 0 ist. Diese zentrierten Mittelwerte sind dann die Schätzung der Saisonkomponente:

$$\hat{s}_{t+1} = \bar{d}_1 - \frac{1}{p}\sum_{l=1}^{p} \bar{d}_l \ , \ldots , \hat{s}_{t+p} = \bar{d}_p - \frac{1}{p}\sum_{l=1}^{p} \bar{d}_l \ .$$

Zerlegung in Trend und Saison

Schätzung der Saison (Fortsetzung)

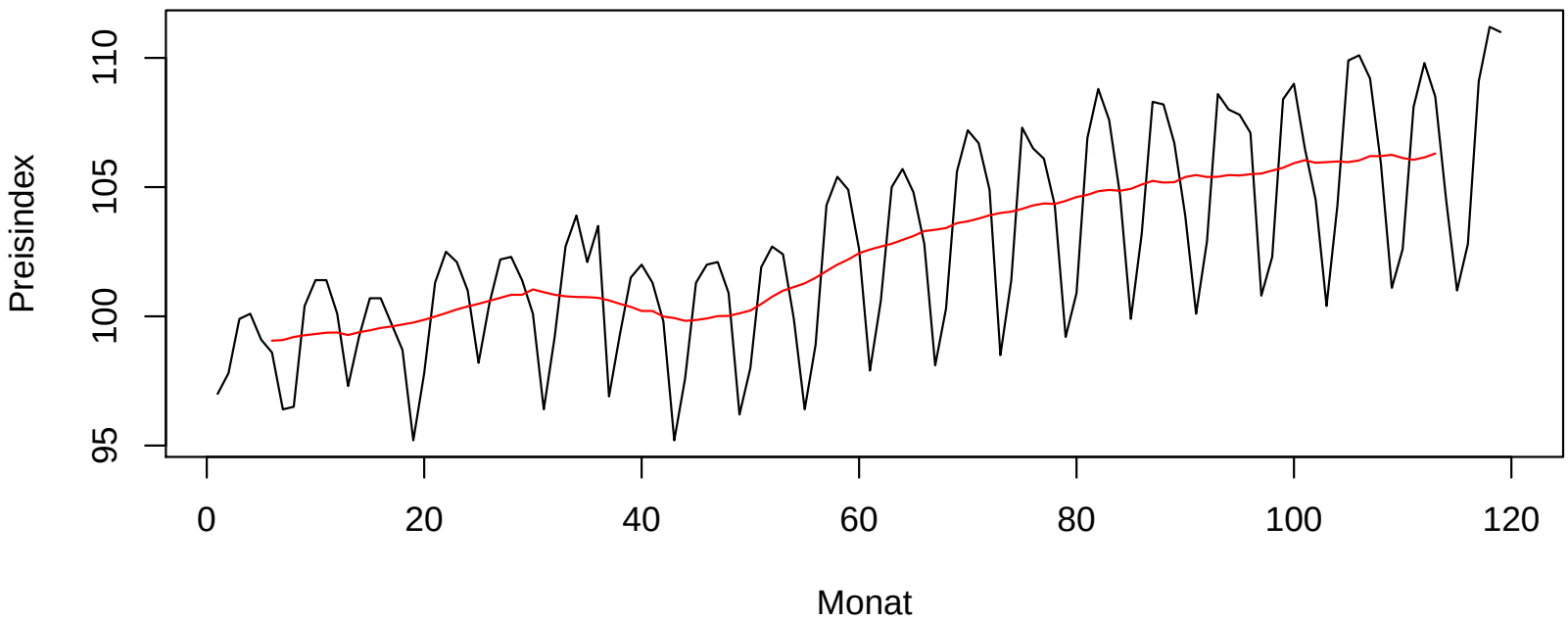
- Die Differenz aus der ursprünglichen Reihe y_t und der geschätzten Saisonkomponenten, also

$$y_t - \hat{s}_t \ ,$$

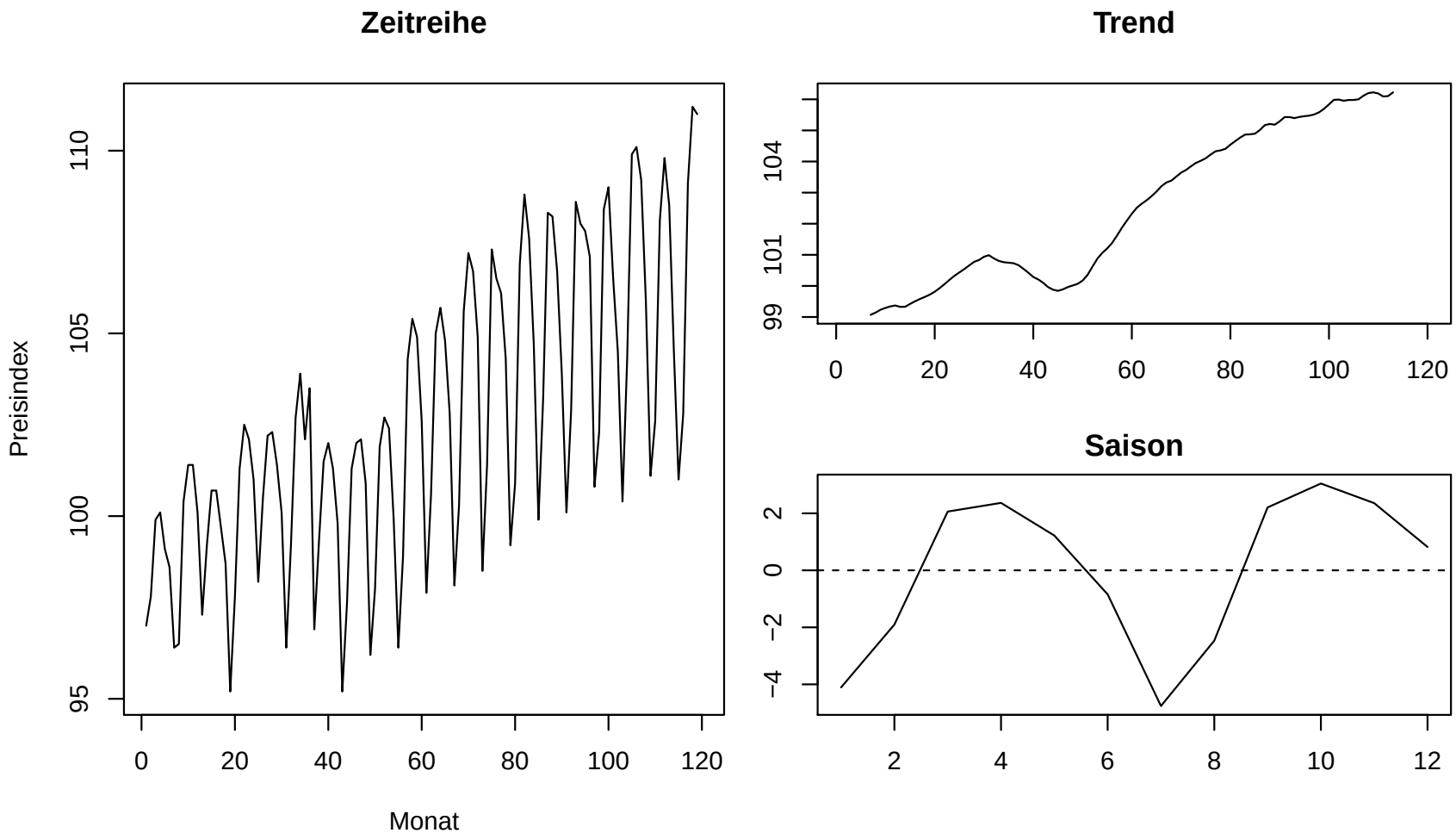
nennt man die saisonbereinigte Reihe.

Beispiel: Preisindex für Kleidung

Zeitreihe mit gleitendem Durchschnitt der Ordnung $12 = p \cdot 1$



Zeitreihenzerlegung



Trendbestimmung mithilfe von Regression

Verwende Trendmodell, z.B.

$g(t) = \beta_0 + \beta_1 t$	linearer Trend
$g(t) = \beta_0 + \beta_1 t + \beta_2 t^2$	quadratischer Trend
$g(t) = \beta_0 \cdot \exp(\beta_1 t)$	exponentielles Wachstum

Schätzung der Trendkomponente durch Regression

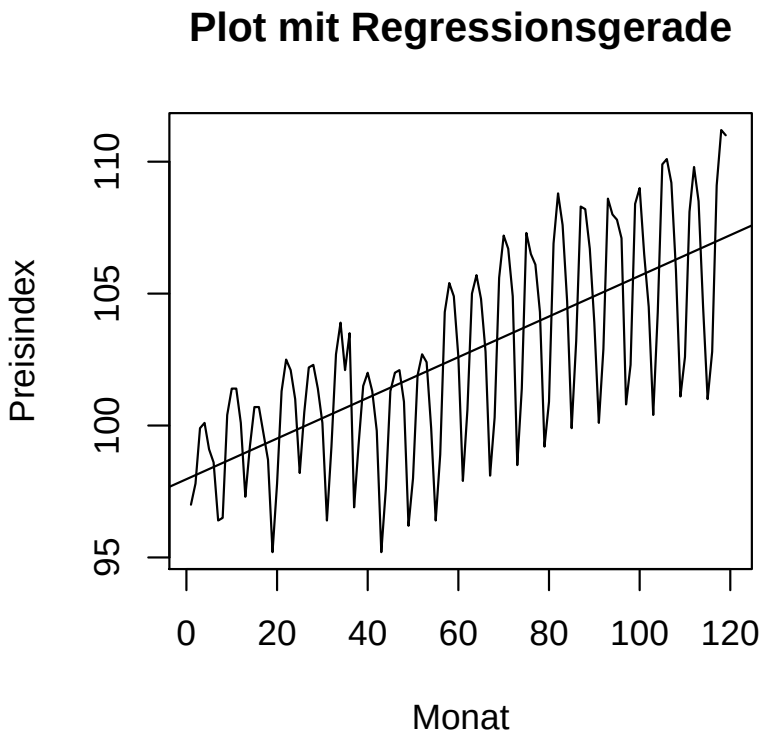
Die KQ-Methode für lineare Regression liefert mit der Regressionsgleichung $y_t = \beta_0 + \beta_1 t + \varepsilon_t$ die KQ-Schätzung $\hat{\beta}_0, \hat{\beta}_1$. Die Trendkomponente ist dann

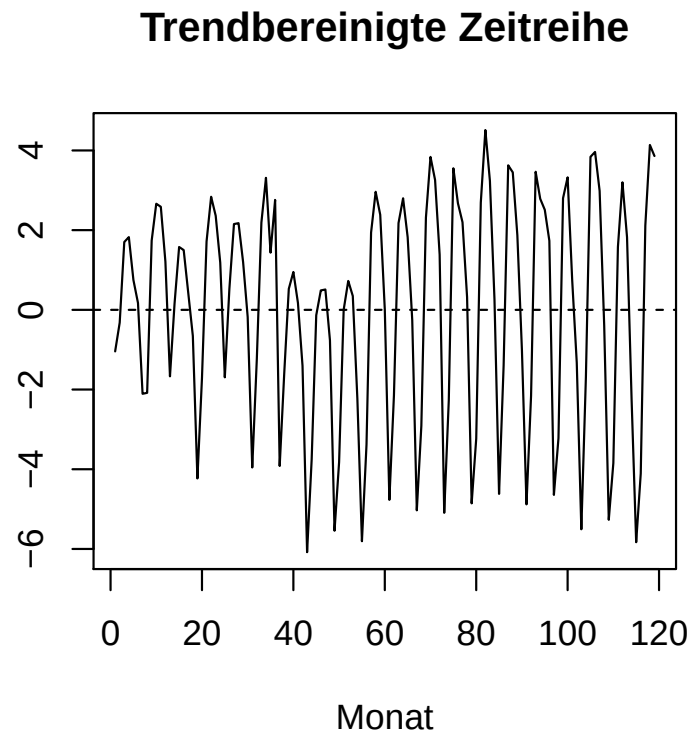
$$g(t) = \hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$$

und die trendbereinigte Zeitreihe

$$\tilde{y}_t = y_t - \hat{y}_t = y_t - g(t).$$

Beispiel: Preise für Bekleidung





Gleichung: $\hat{y}_t = 97.97 + 0.077t$

Bestimmung der Saisonkomponente

Modellierung mit Dummyvariablen für jeden Monat

$$s_j(t) = \begin{cases} 1 & t \text{ gehört zu Monat } j \\ 0 & \text{sonst} \end{cases}, j = 1, \ldots, 12$$

$$s_t = \gamma_1 s_1(t) + \ldots + \gamma_{12} s_{12}(t)$$

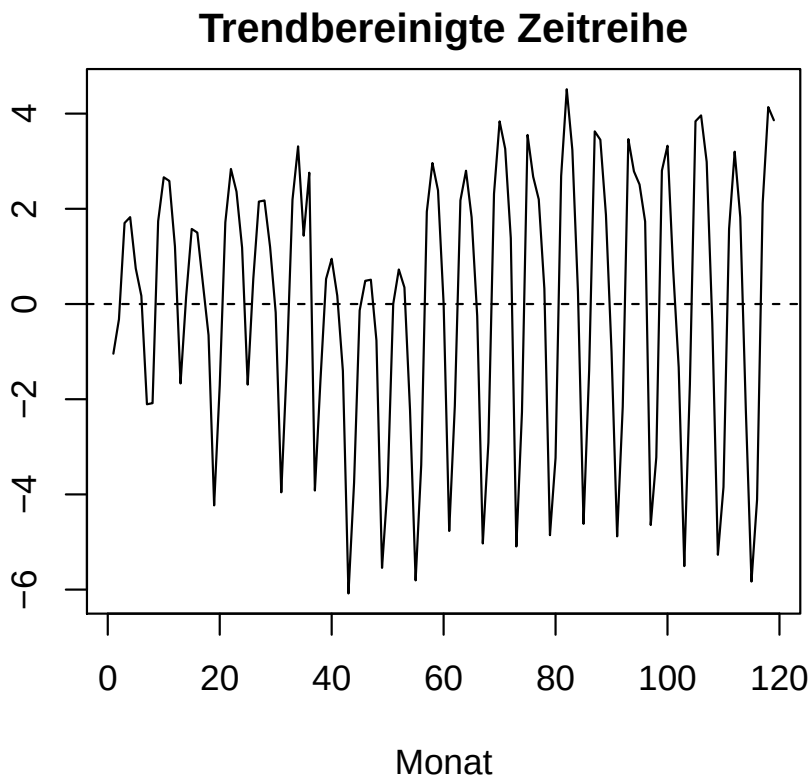
Schätzung der Saisonkomponente mit KQ-Methode

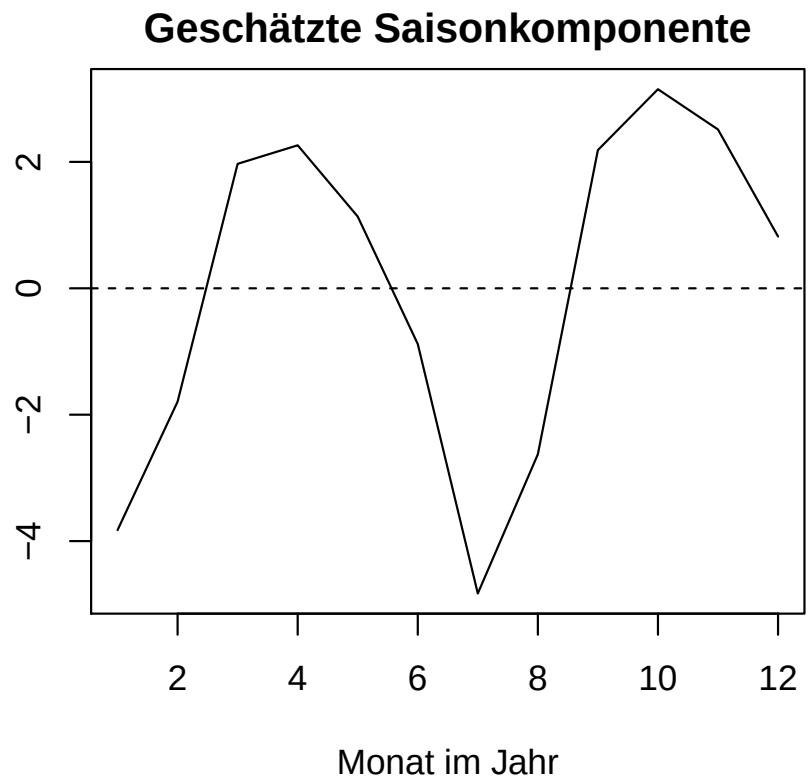
für Zeitreihe ohne Trend bzw. für trendbereinigte Zeitreihe.

$$y_t = \gamma_1 s_1(t) + \ldots + \gamma_{12} s_{12}(t) + \varepsilon_t$$

Verwende Regression zur Schätzung der Parameter $\gamma_1, \ldots, \gamma_{12}$.

Beispiel: Preise für Bekleidung





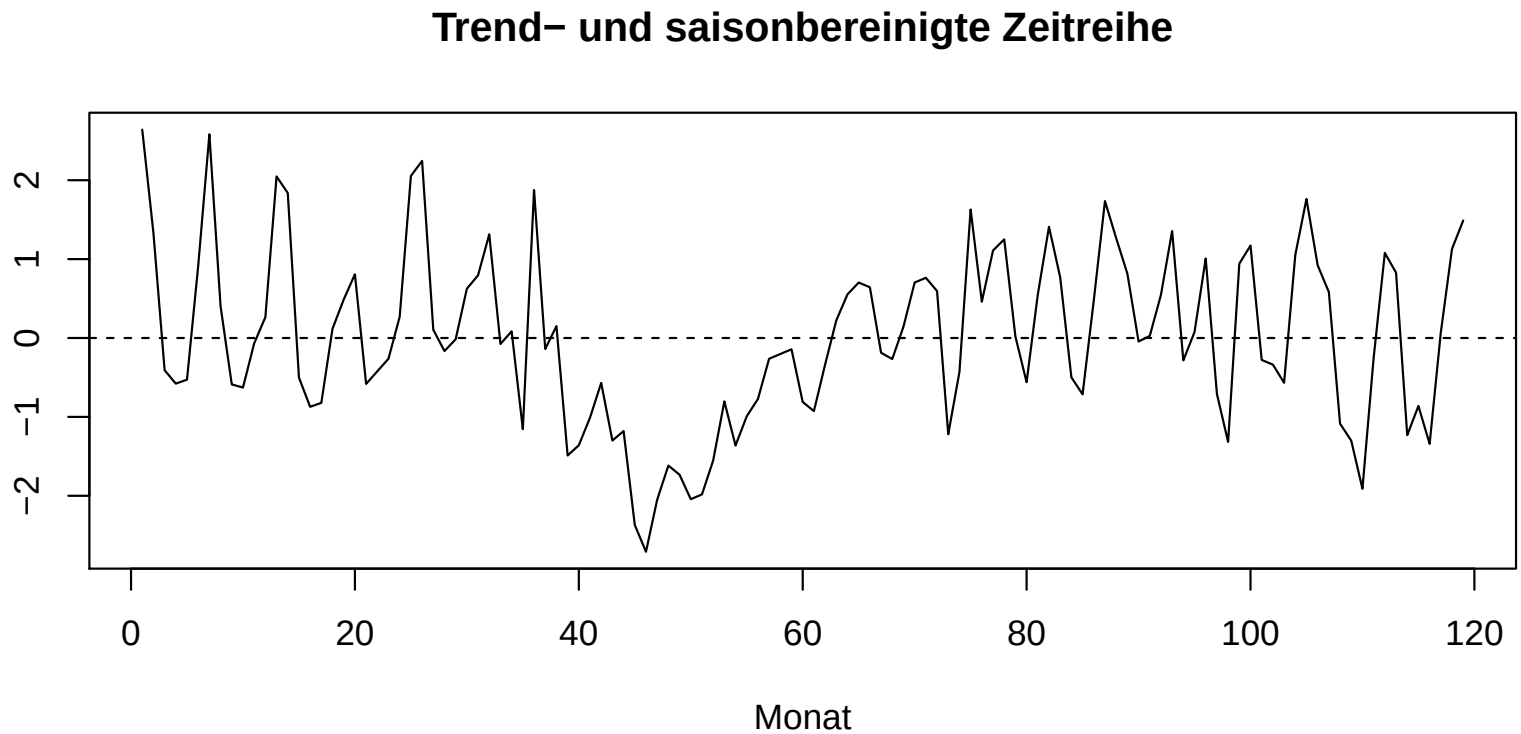
Trend-Saison-Modell

Es ist in der Regel besser, Trend und Saison in einem Modell simultan zu schätzen:

$$y_t = \beta_0 + \beta_1 t + \gamma_1 s_1(t) + \ldots + \gamma_{12} s_{12}(t) + \varepsilon_t$$

mit $\sum_{i=1}^{12} \gamma_i = 0$.

Beispiel: Preise für Bekleidung



Weitere Regressionsansätze

- Verwende kompliziertere Trendmodelle, z.B. polynomialer Trend:

$$g(t) = \beta_0 + \beta_1 t + \ldots + \beta_p t^p$$

Exponentielles Glätten

Rekursive Definition

$$g^e(1) = y_1$$
$$g^e(t) := \beta g^e(t-1) + (1-\beta) y_t$$

einfaches exponentielles Glätten mit Glättungsparameter β.

$$g^e(1) = y_1$$
$$g^e(2) = (1-\beta) y_2 + \beta y_1$$
$$g^e(3) = (1-\beta) y_3 + \beta(1-\beta) y_2 + \beta^2 y_1$$
$$g^e(4) = (1-\beta) y_4 + \beta(1-\beta) y_3 + \beta^2(1-\beta) y_2 + \beta^3 y_1$$

usw.

Alternative Ansätze

Außer den vorgestellten Methoden der Zerlegung gibt es noch eine Vielzahl anderer Verfahren.

Beispiele

1. Lokales lineares Trendmodell: „Gleitende Regression“ statt gleitender Durchschnitt.
2. Census X-11 ARIMA, Census X-12 ARIMA, BV 4.1 (Berliner Verfahren), etc.
3. (S)ARIMA: (**S**easonal) **A**uto**r**egressive **I**ntegrated **M**oving **A**verage

12 Wahrscheinlichkeit

Wahrscheinlichkeitsrechnung für Wirtschaftswissenschaft?

1. Probabilistisches Denken (d.h. das Denken in Wahrscheinlichkeiten) unerlässlich! Strenge Kausalitäten (wenn A dann folgt immer B) findet man bestenfalls vereinzelt in Naturwissenschaften, in den Wirtschaftswissenschaften gilt typischerweise nur: wenn A dann folgt eher B als C.

Wahrscheinlichkeitsrechnung für Wirtschaftswissenschaft?

1. Probabilistisches Denken (d.h. das Denken in Wahrscheinlichkeiten) unerlässlich! Strenge Kausalitäten (wenn A dann folgt immer B) findet man bestenfalls vereinzelt in Naturwissenschaften, in den Wirtschaftswissenschaften gilt typischerweise nur: wenn A dann folgt eher B als C.
2. Wahrscheinlichkeiten und Umgang mit Unsicherheit spielen in der Wirtschaft eine wichtige Rolle. Bei naiver Herangehensweise (ohne Wahrscheinlichkeitsrechnung) kann man sich leicht täuschen. Risikobewertung ist ein zentraler Aspekt bei unternehmerischem Handeln.

Wahrscheinlichkeitsrechnung für Wirtschaftswissenschaft?

1. Probabilistisches Denken (d.h. das Denken in Wahrscheinlichkeiten) unerlässlich! Strenge Kausalitäten (wenn A dann folgt immer B) findet man bestenfalls vereinzelt in Naturwissenschaften, in den Wirtschaftswissenschaften gilt typischerweise nur: wenn A dann folgt eher B als C.
2. Wahrscheinlichkeiten und Umgang mit Unsicherheit spielen in der Wirtschaft eine wichtige Rolle. Bei naiver Herangehensweise (ohne Wahrscheinlichkeitsrechnung) kann man sich leicht täuschen. Risikobewertung ist ein zentraler Aspekt bei unternehmerischem Handeln.
3. Stichprobenverfahren und statistische Modelle spielen in den (empirisch orientierten) Wirtschaftswissenschaften eine zentrale Rolle. Für das Verständnis sind Grundlagenkenntnisse in Wahrscheinlichkeitsrechnung zentral.

Klassische Aspekte und Meilensteine

Wahrscheinlichkeit

- Wahrscheinlichkeit im Glücksspiel, v.a. Würfelspiel: Profanisierung erst im Mittelalter, dort erst als Zufall gedeutet, vorher oft als Gottesurteil etc.
  - Cardano (1501-1576)
  - Gallilei (1546-1642)
  - Briefwechsel zwischen Pascal (1623-1662) und Fermat (1601-1665), erste systematische Wahrscheinlichkeitsrechnung: Lösung für Frage, wie Einsätze gerecht aufzuteilen sind, wenn Spiel unterbrochen wurde
  - Huygens (1629-1695)
- Wahr-schein-lichkeit (Prove-ability $\rightarrow$ probability)

Historische Wurzeln

Mathematisierung von Glücksspiel

- als philosophischer/theologischer Begriff
- der Philosophie des Unsicheren und
- der Mathematik der Glücksspiele

Jacob Bernoulli (1654 - 1705)

Binomialverteilung
Theorem von Bernoulli: durch genügend große Versuchsreihen kann der Unterschied zwischen der relativen Häufigkeit eines Ereignisses und seiner Wahrscheinlichkeit beliebig gering gemacht werden.

Laplace'scher Wahrscheinlichkeitsbegriff

Laplace (1749 - 1827)

- Aufbauend auf Symmetrieüberlegungen
- Wahrscheinlichkeit eines Ereignisses A:

$$P(A) := \frac{\text{Anzahl der für } A \text{ günstigen Fälle}}{\text{Anzahl der (gleich) möglichen Fälle}}$$

Beispiel

Wurf eines fairen Würfels

- Wahrscheinlichkeit des Ereignisses A: Es wird eine gerade Zahl gewürfelt
  möglich: $\{1, 2, 3, 4, 5, 6\}$
  günstig: $\{2, 4, 6\}$

$$\Longrightarrow P(A) = \frac{3}{6} = \frac{1}{2}$$

- Erfolgreiche Anwendung v.a. auf Glücksspiele, in der Physik (stochastische Mechanik) und in der Stichprobentheorie bei einer **einfachen Zufallsauswahl**
- Intuitiv einleuchtend, aber beschränkte Anwendbarkeit

Probleme

Warum reichen Laplace-Wahrscheinlichkeiten nicht?

Essentielle Voraussetzung: alle Fälle müssen gleich möglich (also gleich wahrscheinlich) sein!

Beispiel: Wie wird das Wetter morgen? 3 Möglichkeiten:

$$\{\text{Sonne,Regen,Gemischt}\} \Longrightarrow \text{P(Sonne)} = \frac{1}{3}$$

Wahrscheinlichkeitsbegriffe

Objektivistisch / frequentistische Richtungen / aleatorische Wahrscheinlichkeiten

- Anschluss an die göttliche Ordnung
- Wahrscheinlichkeiten beschreiben tatsächlich vorhandene, zufällige Gesetzmäßigkeiten
- Objektbezogen: Wahrscheinlichkeit ist eine Eigenschaft des untersuchten Objekts (z.B. Würfel), objektiv $\longleftrightarrow$ objektbezogen (wie z.B. spezifisches Gewicht, Länge)
- Häufigkeitsinterpretation bzw. sogar -definition Wahrscheinlichkeit als relative Häufigkeiten in unendlich langen reproduzierbaren Experimenten

Frequentistischer Wahrscheinlichkeitsbegriff I

R. von Mises (1883 - 1953):

„Die Wahrscheinlichkeit eines Ereignisses ist die langfristige relative Häufigkeit seines Auftretens“

Für ein Ereignis A:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

n_A : Anzahl der Erfolge
n : Anzahl der Versuche

Frequentistischer Wahrscheinlichkeitsbegriff II

Probleme bei der Definition

- Einmalige Ereignisse
- Grenzwertdefinition
- Experimentdurchführung

Subjektivistische Richtungen I

- Wahrscheinlichkeit hat ausschließlich mit Unsicherheit, nicht mit Zufälligkeit zu tun
  (Man kann auch über völlig deterministische Aspekte unsicher sein!)
- Wahrscheinlichkeit ist Eigenschaft des untersuchenden Subjekts
  $\Longrightarrow$ verschiedene Subjekte können durchaus zu unterschiedlichen Bewertungen kommen.

Subjektivistische Richtungen II

- Anwendung auch auf Aussagen.
  Bsp: Die Wahrscheinlichkeit, dass die Regierungskoalition die gesamte Legislaturperiode hält, ist...
- behaviouristischer Standpunkt: Wahrscheinlichkeiten äußern sich im Verhalten und können so gemessen werden
  z.B. bei Wetten

Wichtig

Subjektiv sind die Wahrscheinlichkeiten aber nicht die Rechenregeln.

Subjektiver Wahrscheinlichkeitsbegriff I

Laplace, Ramsey, de Finetti:

„Die Wahrscheinlichkeit eines Ereignisses ist der Grad der Überzeugung, mit der ein Beobachter aufgrund eines bestimmten Informationsstandes an das Eintreten eines Ereignisses glaubt“

$P(A)$ ist der Wetteinsatz in Euro, den eine Person höchstens einzugehen bereit ist, falls diese bei Eintreten von A einen Euro gewinnt.

Beispiele:

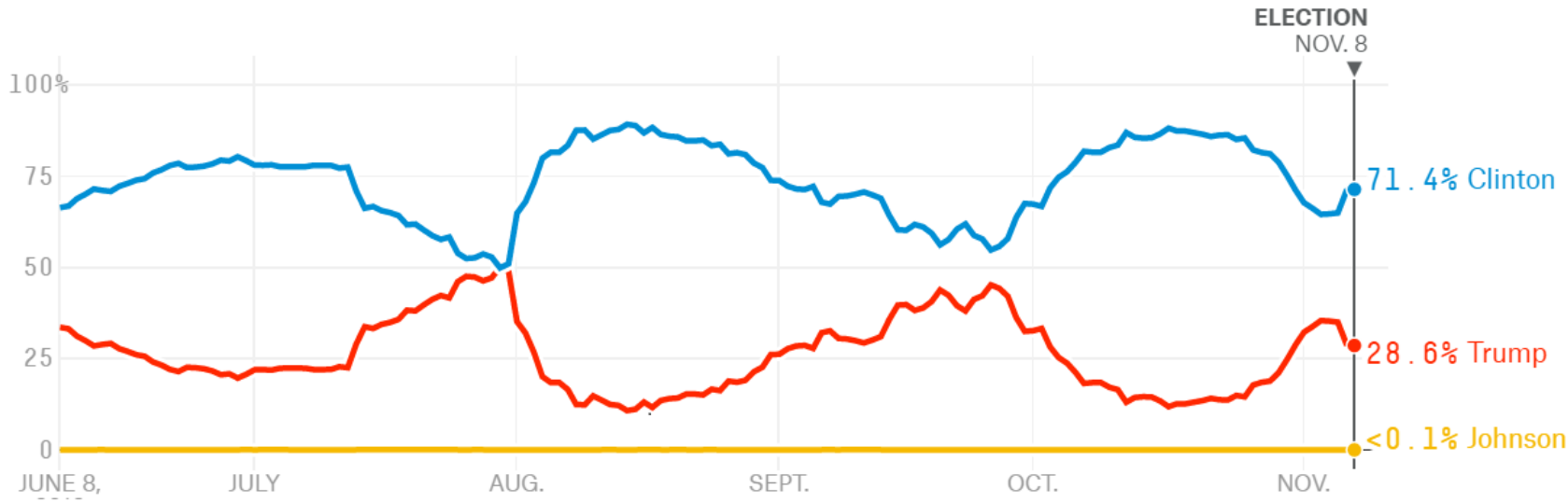
Münzwurf:	Einsatz auf „Zahl“ bis zu 0.5 € sinnvoll
Würfel:	Einsatz auf „5 oder 6“ bis zu 1/3 € sinnvoll

Subjektiver Wahrscheinlichkeitsbegriff II

Probleme

- subjektiv = unwissenschaftlich ?
- Wettdefinition
- Informationsstand

Beispiel: US Wahl



https://projects.fivethirtyeight.com/2016-election-forecast/

Zur Kommunikation von Wahrscheinlichkeiten

Darstellung durch natürliche Häufigkeiten (nach Gigerenzer)

- Superrepräsentative Stichprobe vorstellen, in der sich genau die Häufigkeitsverhältnisse in der Grundgesamtheit wiederfinden, z.B. 10 000 Personen
- Dann $P(A) = 0.1756$ vorstellen als: 1756 Personen haben die Eigenschaft A.

\+ einfachere Kommunikation von Wahrscheinlichkeiten und Risiken, reduziert Fehler beim Rechnen mit Wahrscheinlichkeiten
Experimente mit Ärzten zeigen, dass die Darstellungsform Wahrscheinlichkeiten vs. natürliche Häufigkeiten) einen starken Einfluss auf die Korrektheit von Berechnungen hat.

– Gefahr der Verschleierung von Unsicherheit: die natürlichen Häufigkeiten sind zu *erwartende Durchschnittswerte*, wenn man sehr viele Stichproben hätte.

Beispiel: Beipackzettel

Angabe des Risikos von Nebenwirkungen auf Beipackzetteln

sehr häufig:	mehr als 1 von 10 Behandelten
häufig:	weniger als 1 von 10, aber mehr als 1 von 100 Behandelten
gelegentlich:	weniger als 1 von 100, aber mehr als 1 von 1000 Behandelten
selten	weniger als 1 von 1000, aber mehr als 1 von 10000 Behandelten
sehr selten:	1 Fall oder weniger von 10000 Behandelten, einschließlich Einzelfälle

Welche Nebenwirkungen können bei der Anwendung von *** auftreten?

Gelegentlich wurde über das Auftreten von Mundschleimhautentzündungen, Kopfschmerzen, Ohrengeräuschen berichtet.
Selten können auftreten: Beschwerden im Magen-Darm-Bereich (z.B. Sodbrennen, Übelkeit, Erbrechen oder Durchfall).

Beispiel: Lotto

6 aus 49

- Beim Lotto ist die Wahrscheinlichkeit bei einem Spiel einen 6er zu bekommen:

$$\frac{1}{\binom{49}{6}} = \frac{1}{13983816} = 0.000000072$$

- „Einmal in 14 Millionen Spielen“
- „Einmal in 20.000 Jahren bei wöchentlichem Spielen“
- „Es ist wahrscheinlicher, den Tag der Ziehung nicht mehr zu erleben, als zu gewinnen“
- Simulationsexperiment

Risiko

- Häufig als Wahrscheinlichkeit verwendet
- Manchmal auch als Paar von Wahrscheinlichkeit und Höhe eines Verlustes
- Produkt aus Wahrscheinlichkeit und Schaden
- Entscheidungstheorie unterschiedet verschiedenes Risikoverhalten

Value at risk

- Risikomaß für Wertpapiere
- Der Verlust, der mit einer Wahrscheinlichkeit von $1 - \alpha$ innerhalb eines bestimmten Zeitraums nicht überschritten wird.
- Für verschiedene Portfolios einsetzbar
- Anwendungen auch für Firmen
- Aufsichtsbehörden

Risikodarstellung in der Medizin

Beschreibung von Risiken für die menschliche Gesundheit

- **Absolutes Risiko:**
  Angabe von Krankheitswahrscheinlichkeiten, jeweils getrennt für die Gruppe mit und ohne Risikofaktor
- **Relatives Risiko:**
  Verhältnis der Krankheitswahrscheinlichkeiten mit und ohne Risikofaktor
- **Anzahl der zusätzlich geschädigten Personen**
  (erwarteter Effekt)

Beispiel: Wirkung von Pravastatin

„Menschen mit hohem Cholesterinspiegel können das Risiko eines erstmaligen Herzinfarkts sehr schnell um 22 Prozent vermindern, wenn sie einen häufig angewandten Wirkstoff namens Pravastatin einnehmen“

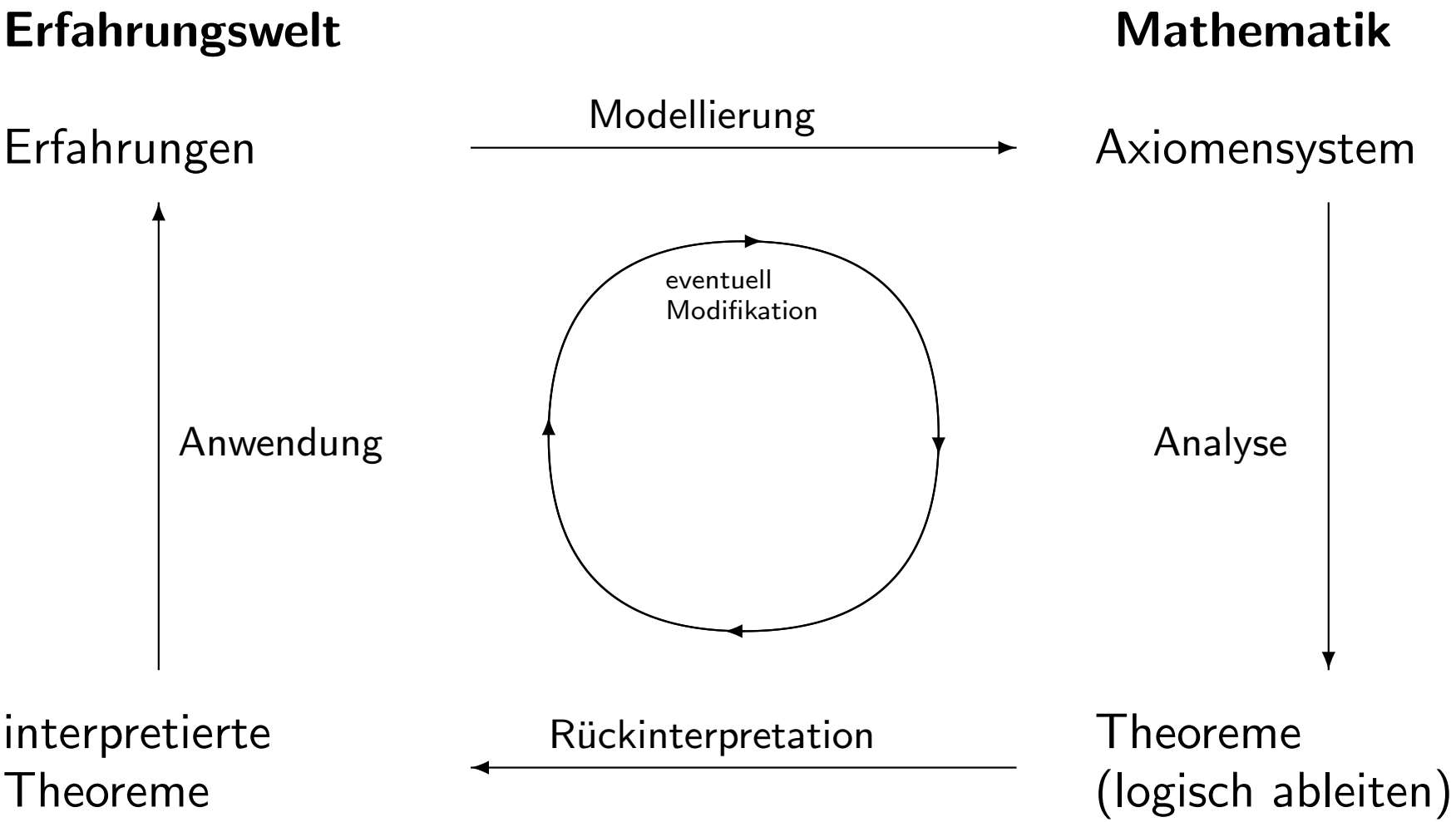
- Reduktion der Todesfälle von 41 auf 32 pro 1000 Patienten mit hohem Cholesterin (32 = 41·(1-0.22)=41·0.78)
  Wahrscheinlichkeit für Todesfall: Reduktion von 4.1% auf 3.2%
  Absolute Risikodifferenz: 0.9%
- Reduktion um 22% (relatives Risiko 0.78) „22% werden gerettet“
- Es müssen 111 Patienten behandelt werden, um ein Menschenleben zu retten.
  Number needed to treat = 1/Absolute Risikodifferenz = 1/0.009 = 111.11

Mathematisch-formaler Wahrscheinlichkeitsbegriff

Axiome

- Axiomatik nach Kolmogoroff
- typische Anwendung der axiomatischen Methode:
  Axiom: Nicht bezweifelte Grundannahme für Kalkül
- Die Kolmogoroffsche Axiomatik ist eine reine Definition, die sich zunächst im luftleeren Raum bewegt. Es wird rein formal festgelegt, was eine Wahrscheinlichkeit sein soll.
- Die Axiomatik ist *verträglich* sowohl mit der *Häufigkeits-* als auch mit der *Wettinterpretation*.
- Die Axiome von Kolmogoroff geben an, wie man mit Wahrscheinlichkeiten rechnet.
- Welche Phänomene man durch Wahrscheinlichkeiten beschreiben darf und wie die Ergebnisse zu interpretieren sind, ist aber damit nicht geklärt.

Die axiomatische Methode



Ausblick

- In der Tat gibt es auch Kritik an dieser Axiomatik: zu streng und überpräzise $\longrightarrow$ aktueller Forschungsgegenstand (*Imprecise Probabilities*, *Intervallwahrscheinlichkeit*); hier nicht näher thematisiert: Kolmogoroff als absolute Wahrheit. Kritik:
  * Modellierung unsicheren (partiell widersprüchlichen, unvollständigen) Expertenwissens
  * Ökonomie: Entscheidungen unter komplexer Unsicherheit widersprechen Prognosen aus der üblichen Wahrscheinlichkeitsrechnung

Zufallsvorgänge

Ein Zufallsvorgang (Zufallsexperiment) führt zu einem von mehreren, sich gegenseitig ausschließenden Ergebnissen. Es ist vor der Durchführung ungewiss, welches Ergebnis eintreten wird.
Was benötigen wir zur Beschreibung eines Zufallsvorganges?

Zwei wesentliche Aspekte:

a) Welche Ergebnisse eines Zufallsvorgangs sind möglich? (Was kann alles passieren?)

b) Mit welcher Wahrscheinlichkeit treten die einzelnen Ergebnisse ein?

Formale Beschreibung I

Ergebnisraum

Festlegen eines *Ergebnisraums* (Grundraum, Stichprobenraum) Ω, der alle möglichen *Ergebnisse* ω enthält.

Beispiele:

- $\Omega = \{1, \ldots, 6\}$ beschreibt die möglichen Ergebnisse eines Würfelexperiments
  Ein mögliches Ergebnis: $\omega = 4$; $\omega = 17$ ist kein mögliches Ergebnis.
- $\Omega = \mathbb{R}_0^+$ beschreibt die möglichen Erwerbseinkommen
  Ein mögliches Ergebnis: $\omega = 17513\,€$
- Ziehung einer Person: $\Omega = \{1, \ldots, N\}$
  Ein mögliches Ergebnis: $\omega = 17$

Formale Beschreibung II

Ereignisse

Ereignisse sind **Teilmengen** von Ω

Beispiele:

- „gerade Zahl“ $= \{2, 4, 6\}$
- „1 oder 2“ $= \{1, 2\}$
- „Einkommen zwischen 1000 und 2000 €“ $= \{\omega | 1000 \leq \omega \leq 2000\}$
- „Person ist weiblich“ $= \{$alle Nummern, die zu Frauen gehören$\}$

Ereignissen sollen Wahrscheinlichkeiten zugeordnet werden.
Wir bezeichnen Ereignisse mit A,B,C,...

Ereignisoperationen

$A \cup B$:	Vereinigung	= „A oder B“
$A \cap B$:	Durchschnitt	= „A und B“
A^C:	Komplement	= „Nicht A“

Beispiele:

Ω	$= \{1,2,3,4,5,6\}$	
A	$= \{2,4,6\}$	„gerade“
B	$= \{4,5,6\}$	„groß“
$A \cup B$	$= \{2,4,5,6\}$	„gerade oder groß“
$A \cap B$	$= \{4,6\}$	„gerade und groß“
A^C	$= \{1,3,5\}$	„ungerade“
B^C	$= \{1,2,3\}$	„klein“

Wahrscheinlichkeit (formale Definition)

Wahrscheinlichkeit

Eine Wahrscheinlichkeitsfunktion ordnet jedem Ereignis seine Wahrscheinlichkeit zu. Eine Wahrscheinlichkeit ist also eine Abbildung von Ereignissen (Elementen der Potenzmenge von Ω) auf reelle Zahlen:

$$\begin{aligned} P : \mathcal{P}(\Omega) &\rightarrow \mathbb{R} \\ A &\mapsto P(A) \end{aligned}$$

Dabei sollen gewisse fundamentale Rechenregeln gelten, z.B.

- 108 kann keine Wahrscheinlichkeit sein, nur Zahlen zwischen 0 und 1.
- $P(\{2, 3\})$ muss mindestens so groß sein wie $P(\{3\})$.

Axiome von Kolmogorov

Die drei Axiome

Eine Funktion P (P steht für Probability), die Ereignissen aus Ω reelle Zahlen zuordnet, heißt *Wahrscheinlichkeit*, wenn gilt

(K1) $P(A) \geq 0$ für alle Ereignisse $A \subset \Omega$.

(K2) $P(\Omega) = 1$.

(K3) Falls $A \cap B = \emptyset$, dann gilt $P(A \cup B) = P(A) + P(B)$

Axiome von Kolmogoroff (1933)

- Die Axiome von Kolmogoroff stellen zunächst eine reine Definition dar, die festlegt, was eine Wahrscheinlichkeit sein soll.
- Es gibt verschiedene Versuche Wahrscheinlichkeiten operational zu definieren (also durch eine Messvorschrift) und verschiedene Interpretationen, die die Axiomatik mit Leben füllen sollen.
- Die Axiome passen zu den beiden bisher diskutierten Wahrscheinlichkeitsbegriffen