Vorlesung: Statistik I für Statistiker, Mathematiker und Informatiker

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

WiSe 2015/2016

Termine und Informationen

Homepage:
https://www.elab.moodle.elearning.lmu.de/course/view.php

Vorlesung:

Prof. Helmut Küchenhoff

Mo. (14 tägig) 12:00 bis 14:00 Uhr	Schellingstr. 3 S 001
Mi. 12:00 bis 14:00 Uhr	Geschw.-Scholl-Pl. 1 E 004

Übung (ca. 14–tägig):
Veronika Deffner

Übung I:	Do. 10:00 bis 12:00 Uhr	Geschwister-Scholl Platz 1 E 004
Übung II:	Do. 12:00 bis 14:00 Uhr	Geschwister-Scholl Platz 1 E 004

Tutorium (ca. 14–tägig):
Andreas Singer

Do. 16:00 bis 18:00 Uhr	Geschwister-Scholl Platz 1 D 209

Hilfsmittel

Programmpaket R

- Frei verfügbar unter www.r-project.org
- Kurse im Rahmen der Veranstaltung „Statistische Software“
- viele Möglichkeiten
- aktuelle Forschung geht in Form von „packages“ ein

Links auf der Homepage zur Vorlesung

Anregungen bitte an Veronika Deffner

Literatur

L.Fahrmeir, R.Künstler, I.Pigeot, G.Tutz:
Statistik - Der Weg zur Datenanalyse
Springer-Verlag, 7. Auflage, 2009

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

Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

Beispiel 1: Bundestagswahl 2013

Prognose 18:00 Infratest Dimap (ARD)

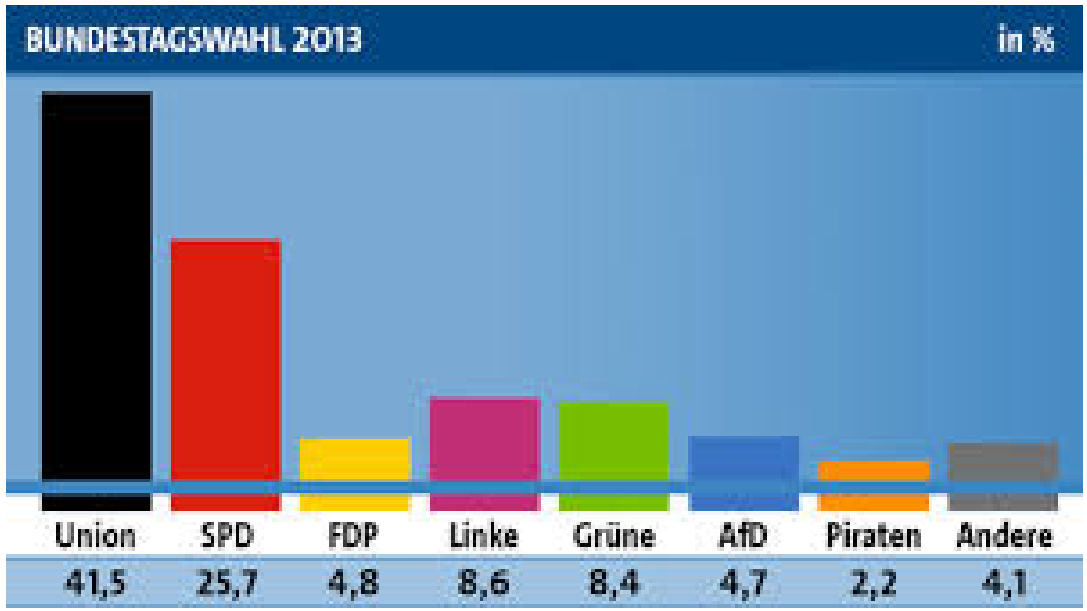
CDU/CSU	SPD	FDP	Linke	Grüne	AFD
42,0	26,0	4,7	8,5	8,0	4,9

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:

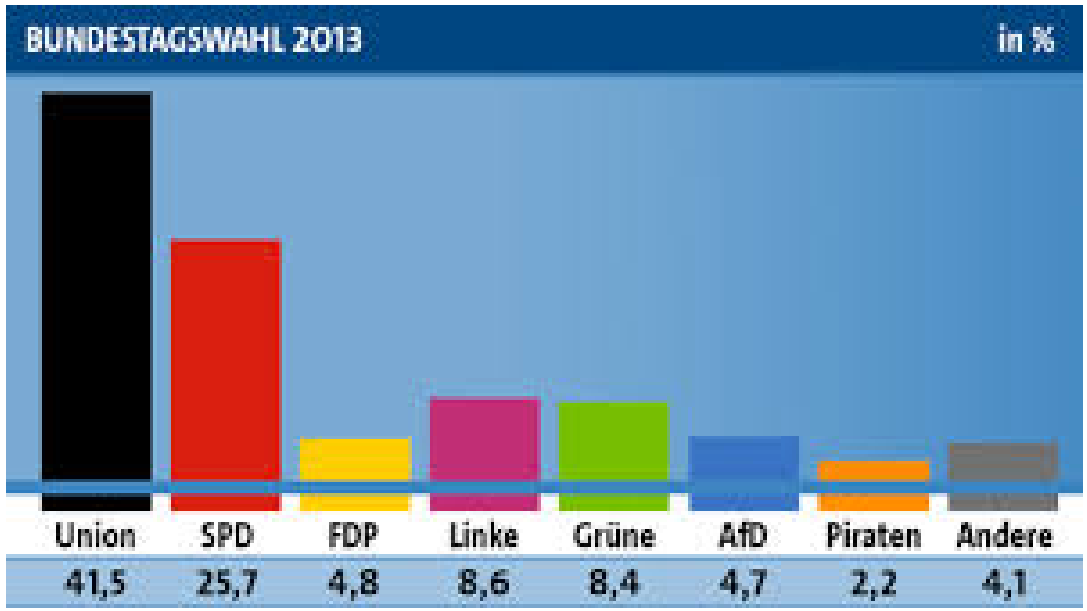


Beispiel 1: Bundestagswahl 2013

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 1: Bundestagswahl 2013

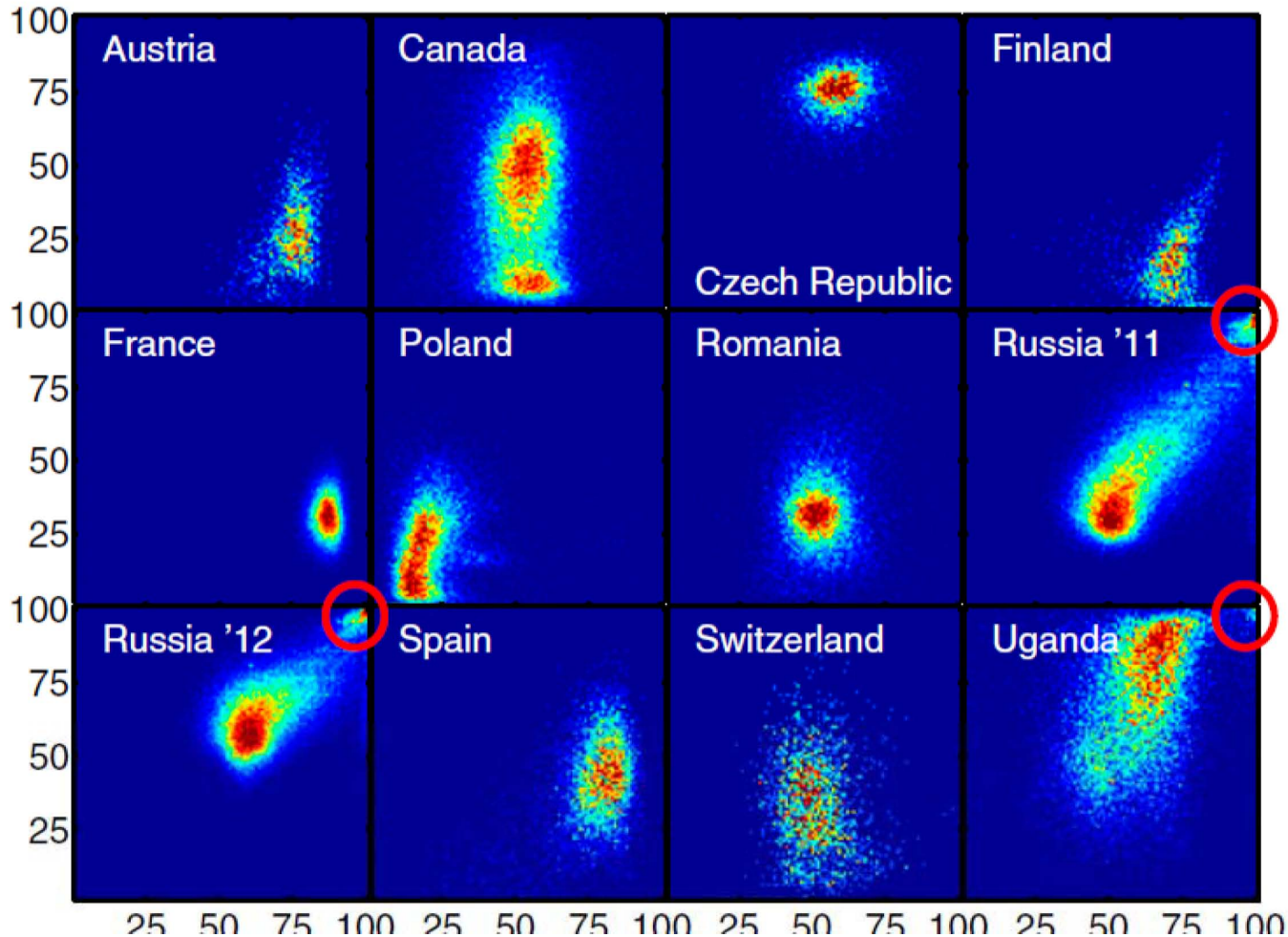
Ziele:

- Schluss von den Befragungsdaten auf das Endergebnis
- Analyse von Wahlverhalten durch weitere Fragen

Beispiel 2: Wahlfälschung

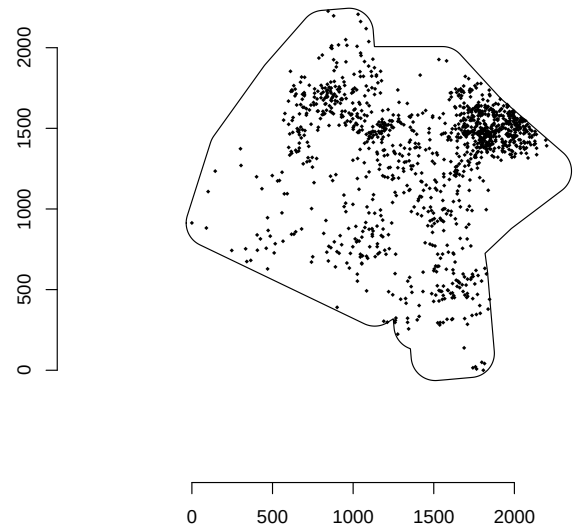
Arbeit von Klimek et al.

Einfache Idee: Untersuche Zusammmenhang zwischen Wahlergebnis (Stimmenanteil des Siegers) gegen die Wahlbeteiligung.



Beispiel 3: Analyse von Daten zu Bombentrichtern

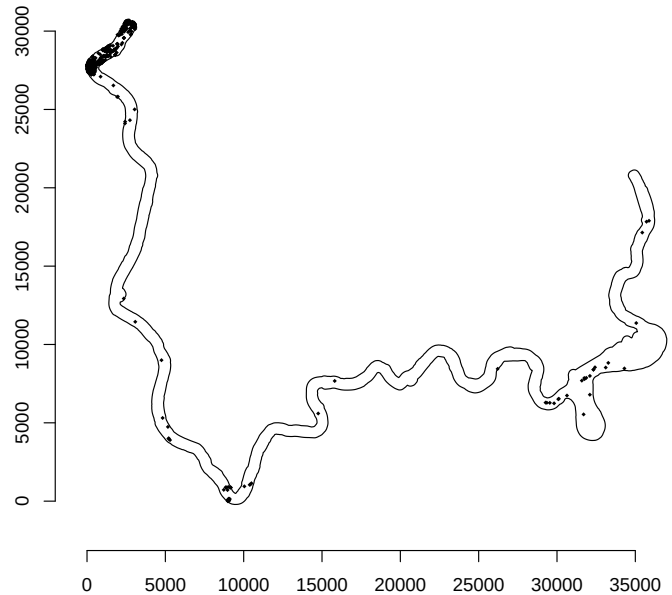
Projekt mit Michael Höhle und Monia Mahling



M. Mahling, M. Hoehle, and H. Küchenhoff. Determining high-risk zones for unexploded World War II bombs by using point process methodology. *Journal of the Royal Statistical Society Series C-Applied Statistics* 62(2):181-199, 2013.

Beispiel 3: Analyse von Daten zu Bombentrichtern

Projekt mit Michael Höhle und Monia Mahling



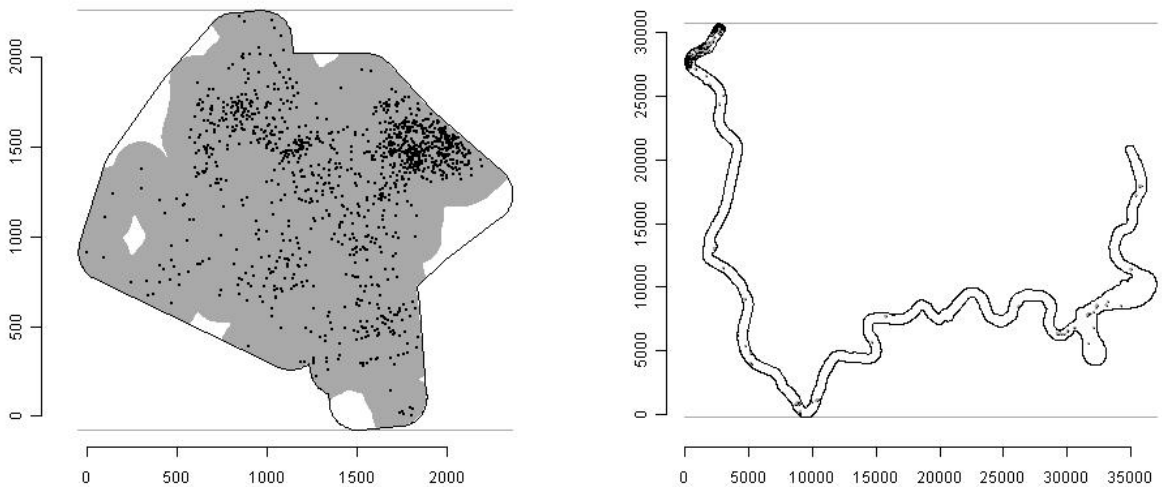
M. Mahling, M. Hoehle, and H. Küchenhoff. Determining high-risk zones for unexploded World War II bombs by using point process methodology. *Journal of the Royal Statistical Society Series C-Applied Statistics* 62(2):181-199, 2013.

Ziele und Methoden

- Räumliche Punktmuster analysieren
- Effiziente Risikoabschätzung
- Algorithmus zur Bestimmung von Sicherheitszonen bei gegebenen Risikoparametern
- Ausweisung von Risikozonen

Lösung

1. Intensitätsschätzung mit Kernmethoden
2. Cut off Wert der Sicherheitszone aus Annahme zum Anteil der Blindgänger



Sicherheitszonen

Beispiel 4: 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.

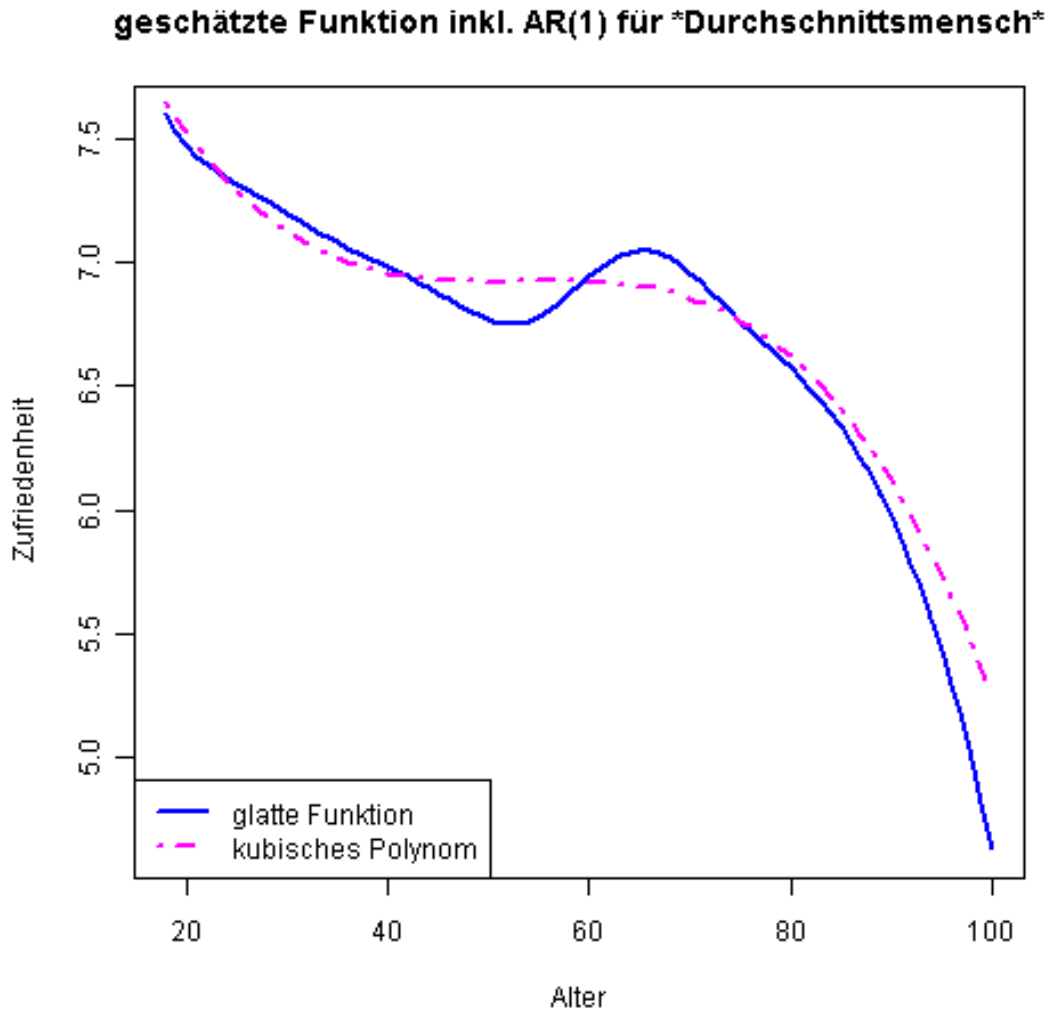
A. Wiencierz, S. Greven, and H. Küchenhoff. Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic Journal of Statistics* 5:1718-1734, 2011.

Datengrundlage

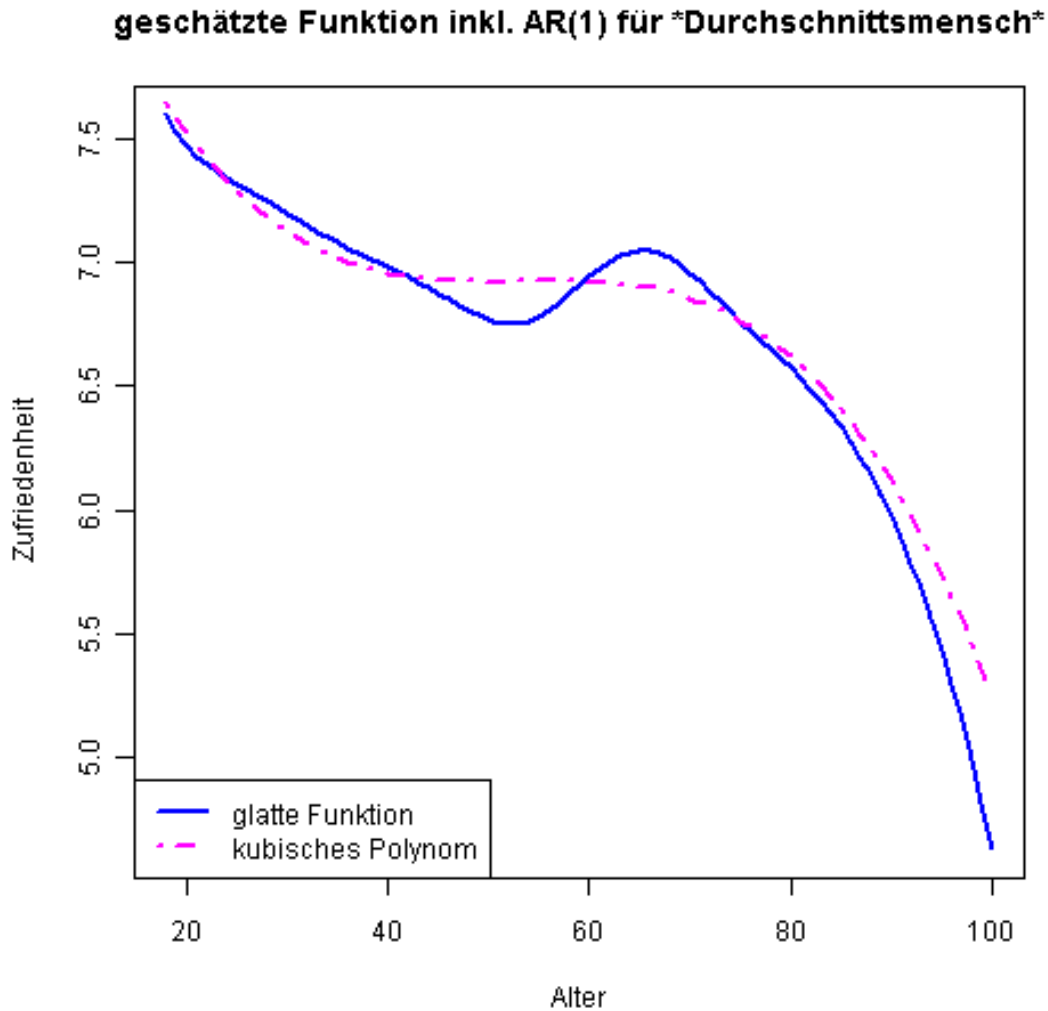
- Daten stammen aus den Haushaltsstichproben A (Westdeutsche) und C (Ostdeutsche) des Sozio-Ökonomischen Panels (SOEP)
- für die ausgewählten Modellvariablen liegen Beobachtungen aus den Jahren 1992, 1994 bis 2006 vor
- durchschnittliche Anzahl von Beobachtungen pro Person: 7.77
- in die Modellberechnungen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

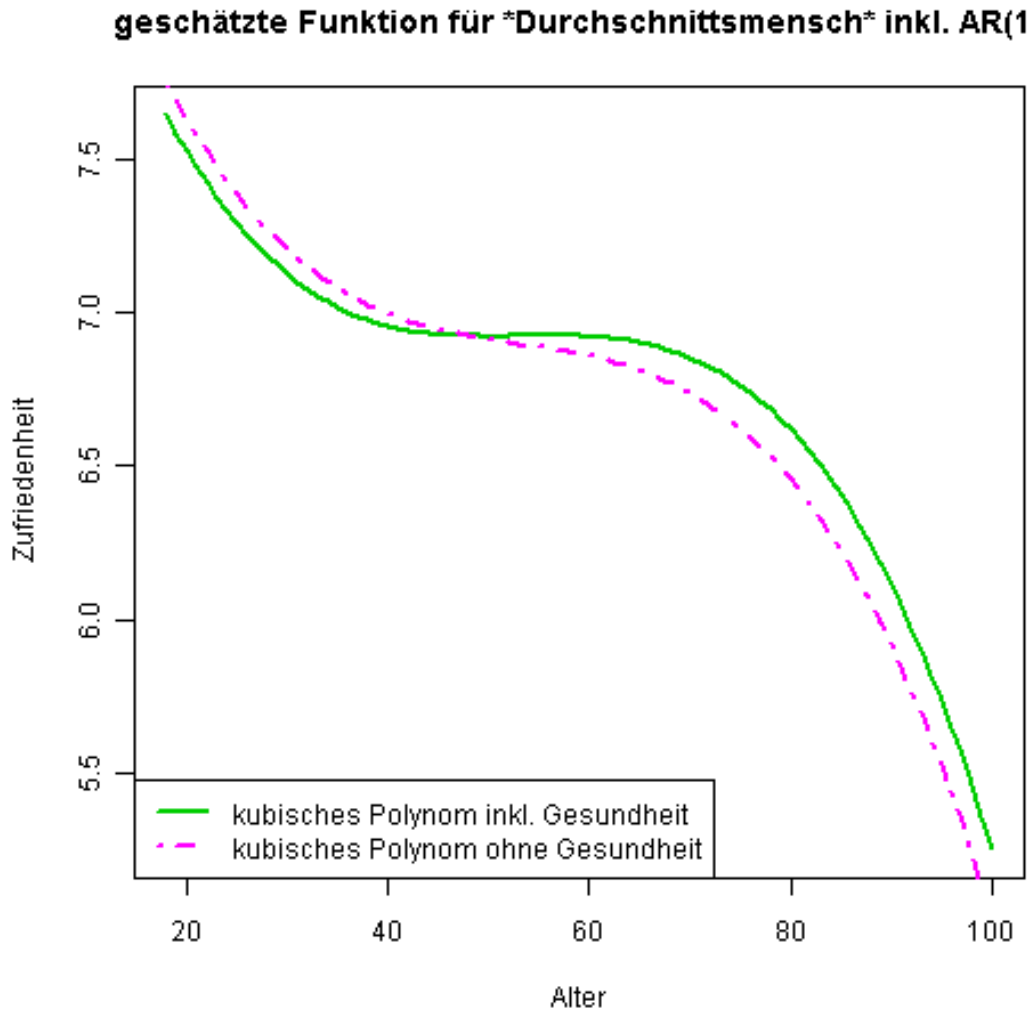


Ergebnis für Alterseffekt

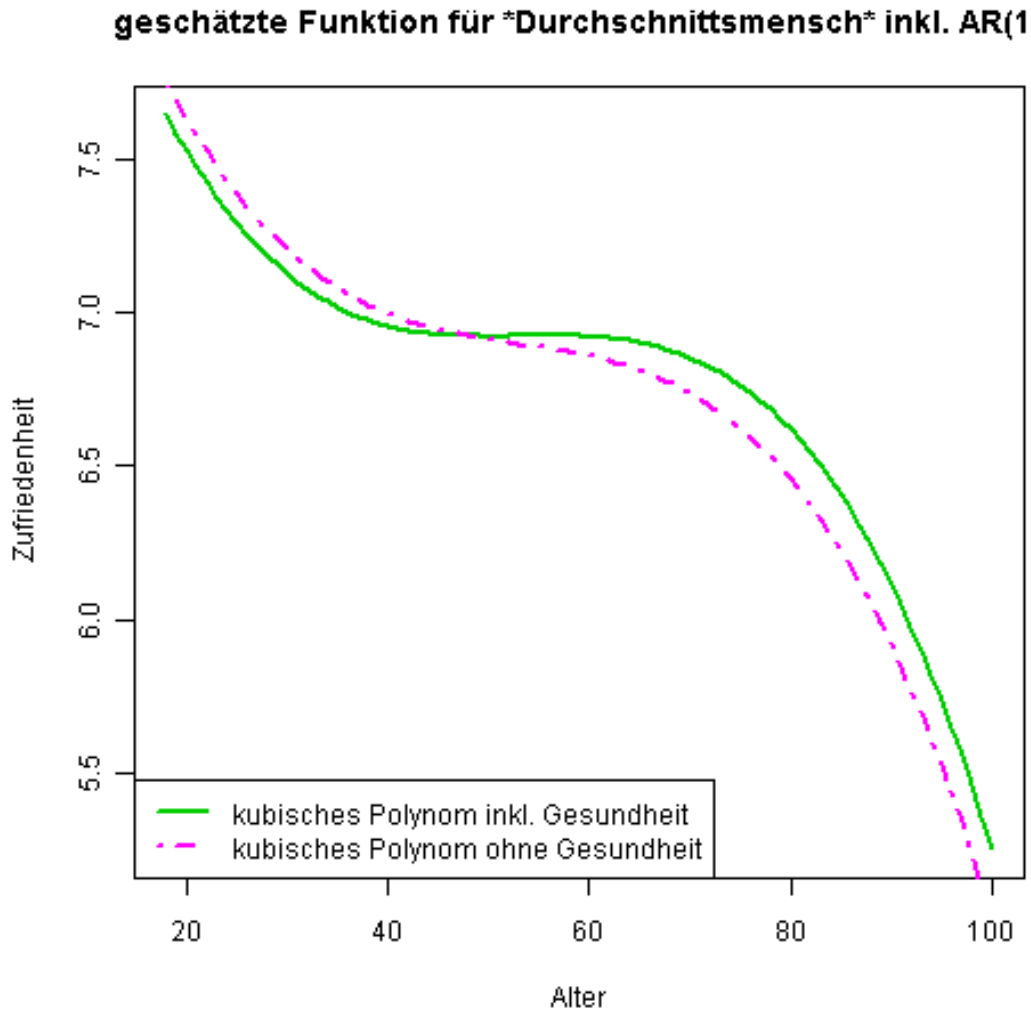


Midlife-Crisis nur bei glatter Funktion erkennbar.

Ergebnisse ohne Gesundheitsvariable



Ergebnisse ohne Gesundheitsvariable



Beachte: Deutlich stärkerer Abfall ohne adjustieren nach Gesundheit

Ziele und Methoden

- Zusammenhänge analysieren
- Komplexe Einflüsse
- flexibles Modell

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

Umweltzone und Feinstaubbelastung

Wirkt die Umweltzone?

Einfacher Ansatz: Vergleiche Mittelwerte vor und nach der Einführung von Umweltzone und Fahrverbot

Umweltzone und Feinstaubbelastung

Wirkt die Umweltzone?

Einfacher Ansatz: Vergleiche Mittelwerte vor und nach der Einführung von Umweltzone und Fahrverbot

Probleme:

- Grundbelastung ohne Autoverkehr kann sich ändern
- Starke Wettereinflüsse
- Schwankungen über Tag und Jahreszeit

Daher: Regressionsmodell mit Referenzstation, Wetter, Tagesverlauf

V. Fensterer, H. Küchenhoff, V. Maier, H.-E. Wichmann, S. Breitner, A. Peters, J. Gu, and J. Cyrys. Evaluation of the impact of low emission zone and heavy traffic ban in Munich (Germany) on the reduction of PM_{10} in ambient air. *International Journal of Environmental Research and Public Health* 11(5):5094-5112, 2014.

Wirkung der Umweltzone

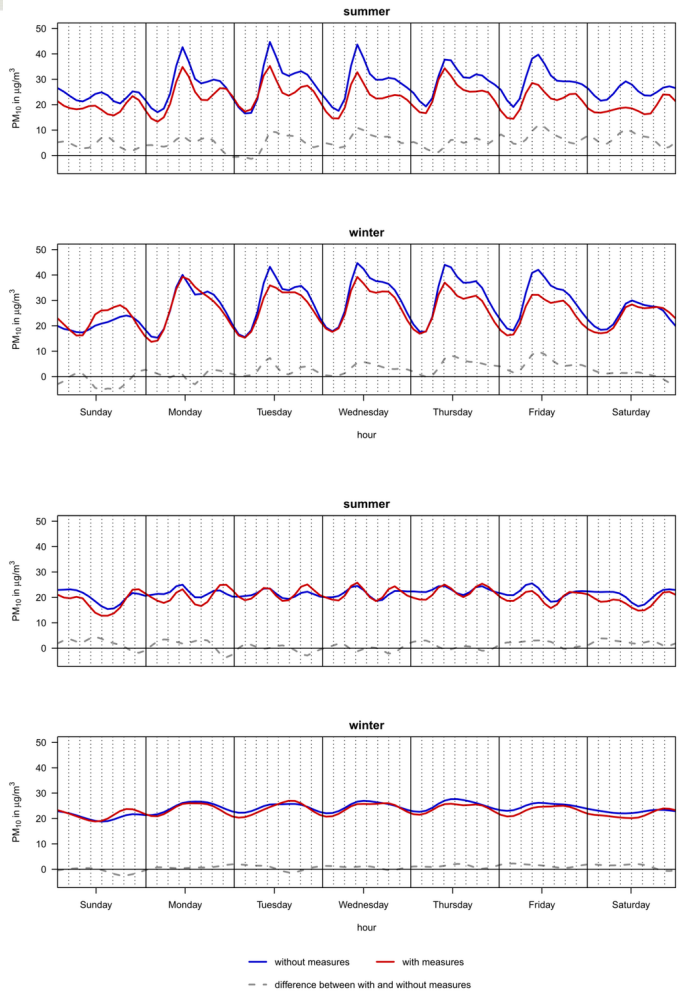


Figure 3: Modeled hourly concentrations of PM10 at Prinzregentenstrasse (first and second chart) and Lothstrasse (third, fourth chart) adjusted for PM10 at the reference station, wind direction and public holidays.

Weitere Beispiele

- Pflege-TÜV
- Wahlforschung
- Epidemiologische Studien

Weitere Beispiele

- Pflege-TÜV
- Wahlforschung
- Epidemiologische Studien
- Marktforschung
  - Einschaltquoten
  - Bewertung und Vergleich von Produkten gleichen Typs aber verschiedener Produzenten durch Verbraucher (Waschmittel, Kaffee, Schokolade, usw.)
- Sportstatistik
- Analyse von Genexpressionsdaten

Weitere Beispiele

- Pflege-TÜV
- Wahlforschung
- Epidemiologische Studien
- Marktforschung
  - Einschaltquoten
  - Bewertung und Vergleich von Produkten gleichen Typs aber verschiedener Produzenten durch Verbraucher (Waschmittel, Kaffee, Schokolade, usw.)
- Sportstatistik
- Analyse von Genexpressionsdaten
- Mustererkennung ("Pattern recognition"): Beispiel: Erkennung von SPAM Mails.

Beispiele Informatik

- Optimierung des Antwortverhaltens von Datenbanken
- Zugriffsstatistiken auf Webserver
- Resourcenplanung von Netzwerken und Servern (der “8-9 Uhr Effekt“, wenn sich alle Mitarbeiter mehr oder weniger gleichzeitig im Netz anmelden und ihre E-Mail abrufen)

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)

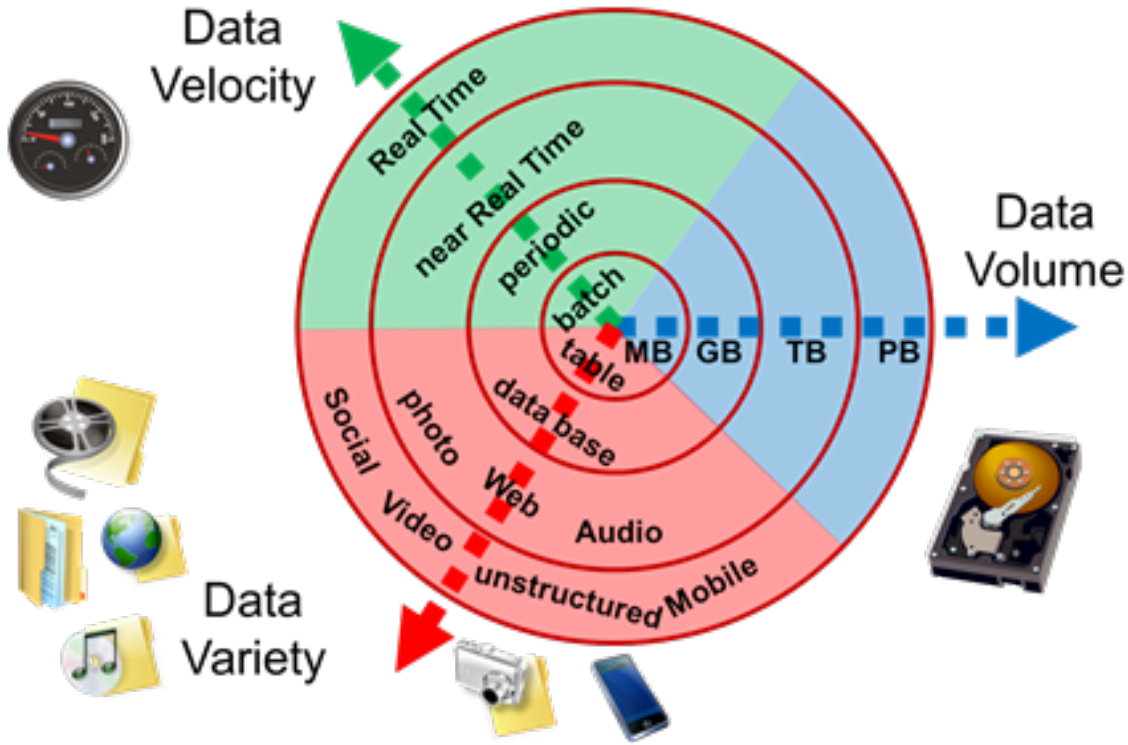
Zitate

- „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)
- Statistik ist für mich das Informationsmittel der Mündigen. Wer mit ihr umgehen kann, ist weniger leicht zu manipulieren. Der Satz „Mit Statistik kann man alles beweisen“ gilt nur für die Bequemen, die keine Lust haben, genau hinzusehen.
  (E. Noelle-Neumann)

BIG DATA

- Analyse und Verarbeitung großer Datenmengen
- Drei Vs
  - Volume
  - Velocity
  - Variety
- Algorithmische Herangehensweise oft ohne Modelle
- Große Herausforderung für die Statistik

BIG DATA


Data
Velocity
Real Time
near Real Time
periodic
batch
table
MB
GB
TB
PB
Data
Volume
data base
photo
Web
Audio
Social
Video
unstructured
Mobile
Data
Variety

Vorlesungsplanung Statistik 1 nach Vorlesungswochen I

1. Einführung, Beispiele, Geschichte
2. Grundlagen der Datenerhebung: Messung, Skalenniveaus
3. Typen von Studien, Auswahlverfahren
4. Univariate deskriptive Statistik 1: Häufigkeiten und graphische Darstellungen, kumulierte Verteilung, Lage- und Streuungsparameter
5. Univariate deskriptive Statistik 2: Boxplots, Schiefe und Wölbung, Kerndichteschätzung
6. Kontingenztafeln, Zusammenhangsmaße für nominalskalierte Merkmale, Plots
7. Rangkorrelationskoeffizient, Zusammenhänge bei quantitativen Merkmalen, Bravais-Pearson-Korrelationskoeffizient

Vorlesungsplanung Statistik 1 nach Vorlesungswochen II

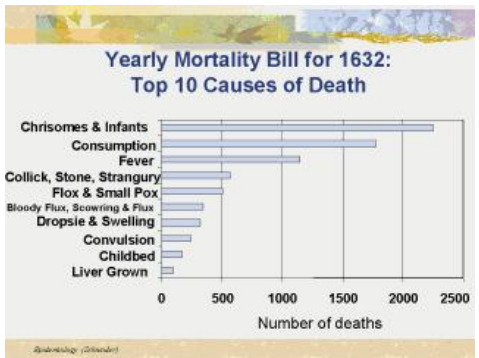
8. Kendall´s Tau, Invarianzeigenschaften
9. Grafische Darstellung von Zusammenhängen
10. Regression (lineare Einfachregression, Streuungszerlegung)
11. Multiple Regression, Scheinkorrelation
12. Interaktive Grafik

Geschichte: „Frühe amtliche Statistik“

2300 v. Chr.	Volkszählung in China
1375 v. Chr.	Volkszählung in Israel 600000 waffenfähige Männer (4. Buch Mose)
seit ca. 550 v. Chr.	Volkszählungen im römischen Reich

Politische Arithmetik

John Graunt (1620 - 1674):
Sterbetafeln der Stadt London nach Ausbruch der Pest



Sir William Petty (1623 - 1687):
„Politische Arithmetik“
Daten für die Verwaltungsreform in Irland

Universitätsstatistik

- **Gottfried Achenwall (1719 - 1722):**

  Begriff „Statistik“ (= den Staat betreffend)

  Fach, das sich mit allerlei Staatsmerkwürdigkeiten beschäftigt.
  Geographie, Wirtschaft und Verwaltung

Entwicklung der amtlichen Statistik in Bayern:
www.historisches-lexikon-bayerns.de/artikel/artikel_44809

Grundlagen

Sir Thomas Bayes (1702 - 1761):
Wichtige theoretische Grundlagen

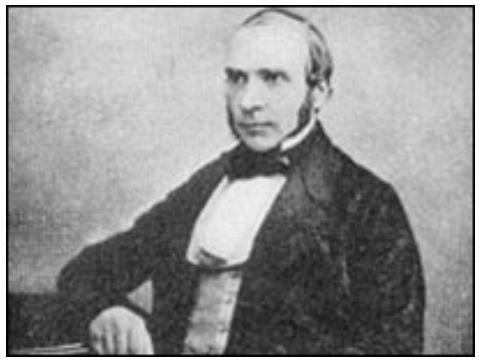
Grundlagen

Sir Thomas Bayes (1702 - 1761):
Wichtige theoretische Grundlagen

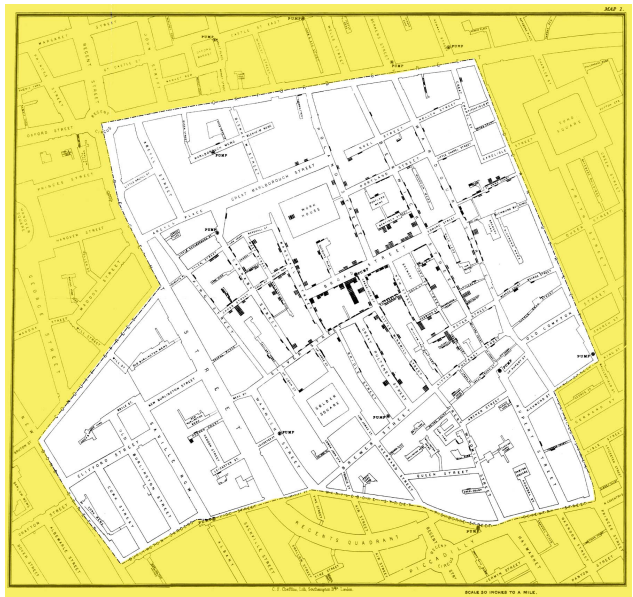
Pierre-Simon Laplace (1749 - 1822):
Wahrscheinlichkeitsrechnung

Epidemiologie

John Snow (1813-1858)

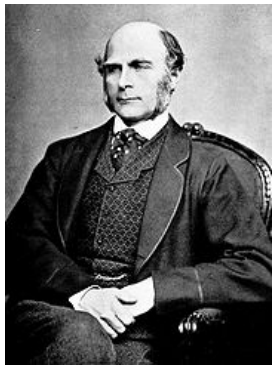


Cholera in London 1854



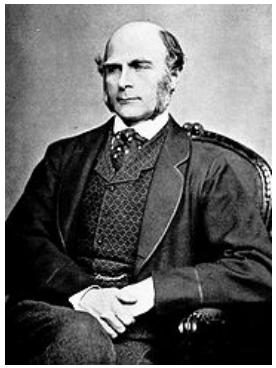
Englische Schule der Inferenzstatistik

Francis Galton (1822 - 1911):
Grundlagen der Regression

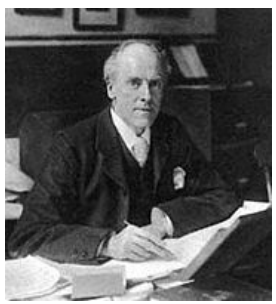


Englische Schule der Inferenzstatistik

Francis Galton (1822 - 1911):
Grundlagen der Regression



Karl Pearson (1857 - 1936):
Theorie der statistischen Tests
(gemeinsam mit J.Neyman)



R.A. Fisher (1890 - 1962):
Grundkonzepte der statistischen Inferenz
Maximum-Likelihood-Prinzip
Statistische Tests

LMU

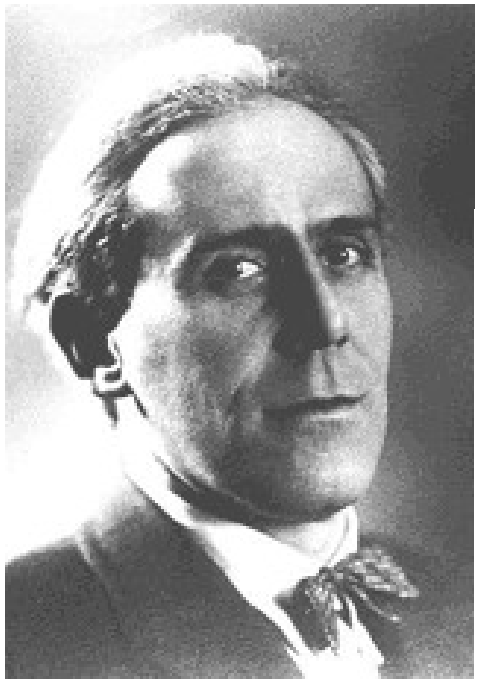
1911	Seminar für Statistik und Versicherungswissenschaft Leitung: Privatdozent Böhm
1978	Studiengang Statistik (Diplom)
1995 - 2006	Sonderforschungsbereich 386 „Statistische Analyse diskreter Strukturen - Modellierung und Anwendung in Biometrie und Ökonometrie“

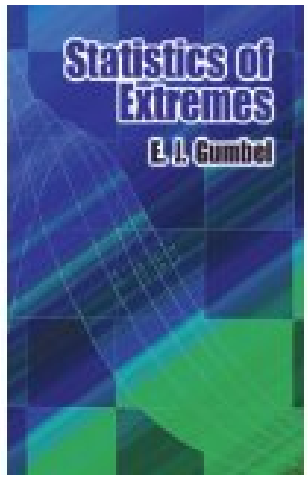
Neuere Entwicklungen

Emil Julius Gumbel	Extremwertstatistik
Cox (1972)	Lebensdauermodelle
Efron (1979)	Computerintensive Verfahren
Mc Cullagh, Nelder (1983)	Verallgemeinerte lineare Modelle
L. Tierney (1994)	Bayesianische Analyse
Hastie, Tibshirani (2001)	Statistical Learning

Emil Julius Gumbel (München 1891- New York 1966

Wichtige Beiträge zur Extremwertstatistik (1958)





Nelder & Wedderburn 1972

Insgesamt 2217 mal zitiert (Stand Okt '15)

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Keywords: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES

INTRODUCTION

Besag 1974

- Insgesamt 2419 mal zitiert (Stand Okt '15)
- Grundidee der räumlichen Statistik

Spatial Interaction and the Statistical Analysis of Lattice Systems

By JULIAN BESAG

University of Liverpool

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, March 13th, 1974, Professor J. DURBIN in the Chair]

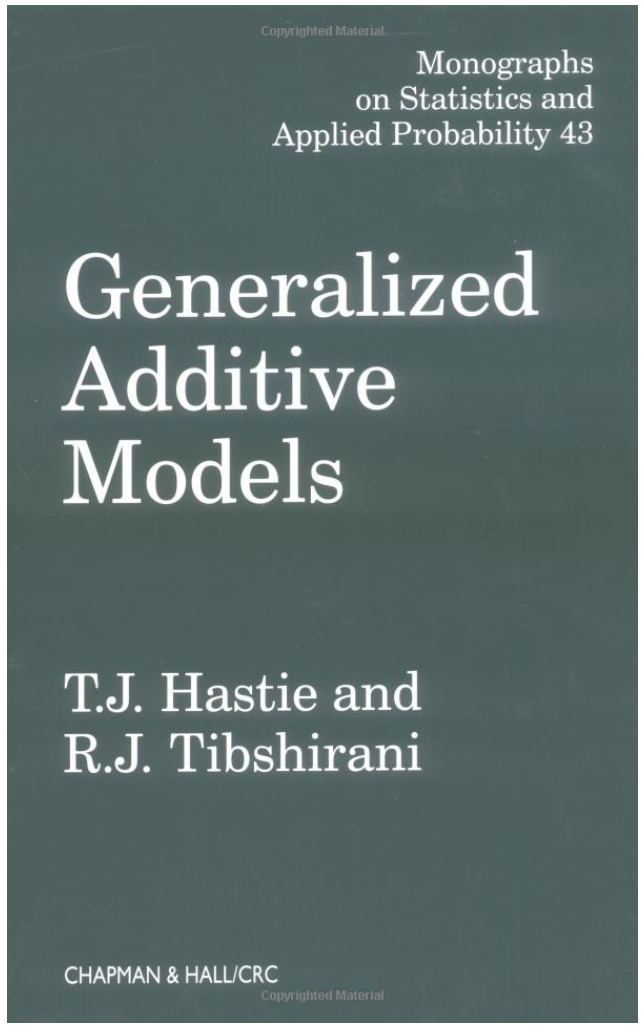
SUMMARY

The formulation of conditional probability models for finite systems of spatially interacting random variables is examined. A simple alternative proof of the Hammersley–Clifford theorem is presented and the theorem is then used to construct specific spatial schemes on and off the lattice. Particular emphasis is placed upon practical applications of the models in plant ecology when the variates are binary or Gaussian. Some aspects of infinite lattice Gaussian processes are discussed. Methods of statistical analysis for lattice schemes are proposed, including a very flexible coding technique. The methods are illustrated by two numerical examples. It is maintained throughout that the conditional probability approach to the specification and analysis of spatial interaction is more attractive than the alternative joint probability approach.

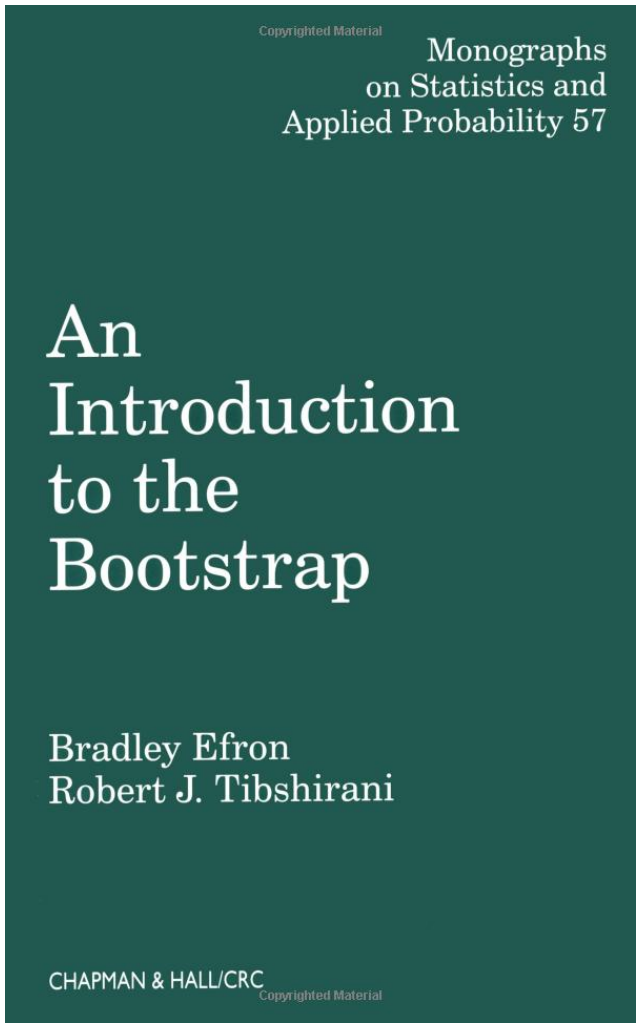
Keywords: MARKOV FIELDS; SPATIAL INTERACTION; AUTO-MODELS; NEAREST-NEIGHBOUR SCHEMES; STATISTICAL ANALYSIS OF LATTICE SCHEMES; CODING TECHNIQUES; SIMULTANEOUS BILATERAL AUTOREGRESSIONS; CONDITIONAL PROBABILITY MODELS

1. INTRODUCTION

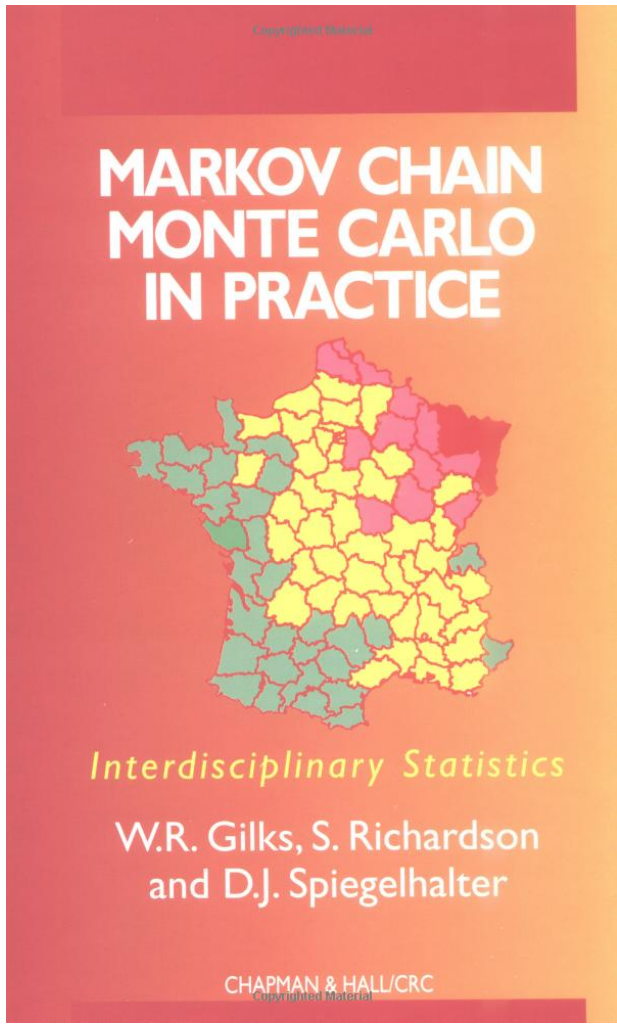
Hastie & Tibshirani 1990



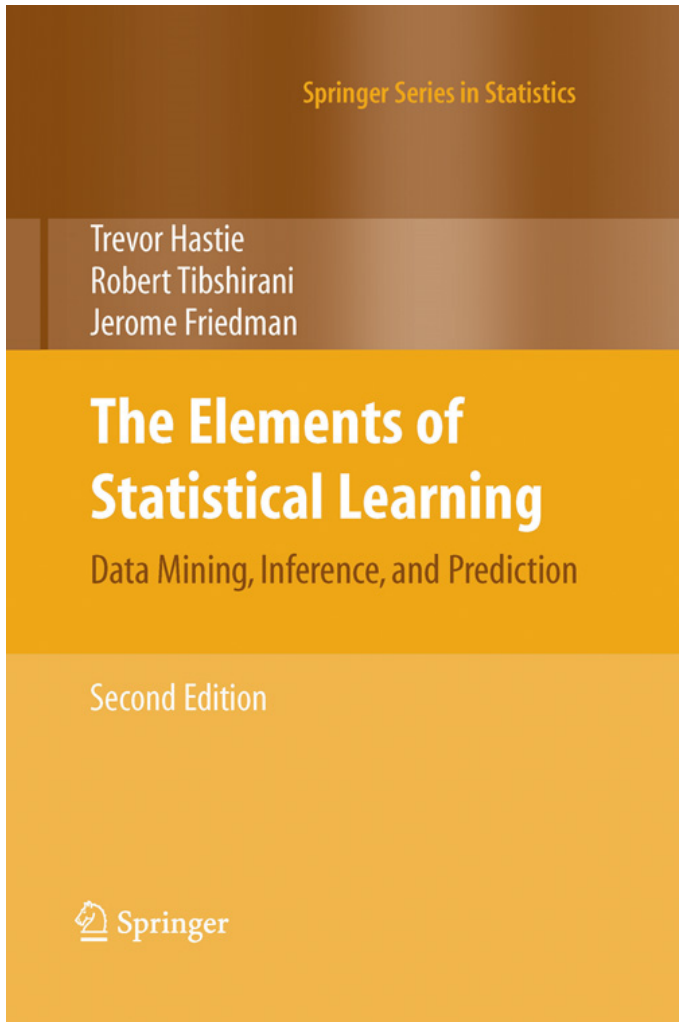
Efron & Tibshirani 1993



Markov chain Monte Carlo (MCMC) 1996



Statistical Learning: Hastie, Tibshirani und Friedmann 2001



Grundbegriffe

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

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“

Lit.: „Measurement theory and practice: The world through quantification“ David J. Hand, Arnold 2004

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

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

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 $\Leftrightarrow 1.84 < 1.91$

Typen von Messungen

1. Messung hat reales (physikalisches) Relativ; direkte Messung („representational measurement“)
   z.B. Länge, Gewicht, Anzahl, Blutzucker, etc.
2. Messung besitzt durch Operationalisierung definiertes Relativ; indirekte Messung
   („pragmatic measurement“), Operationale Messung
   z.B. Intelligenz, Schwere einer Krankheit

Skalen

Homomorphe Abbildung
empirisches Relativ $\Rightarrow$ numerisches Relativ

Skalen

Homomorphe Abbildung
empirisches Relativ $\Rightarrow$ numerisches Relativ

Existenz:	Ist die Struktur der Objekte so, dass eine strukturerhaltende Abbildung existiert? $\Rightarrow$ Axiome von Repräsentationstheoremen müssen erfüllt sein (z.B. Transitivität o.ä.)

Skalen

Homomorphe Abbildung
empirisches Relativ $\Rightarrow$ numerisches Relativ

Existenz:	Ist die Struktur der Objekte so, dass eine strukturerhaltende Abbildung existiert? $\Rightarrow$ Axiome von Repräsentationstheoremen müssen erfüllt sein (z.B. Transitivität o.ä.)
Eindeutigkeit:	Gibt es mehrere zulässige Skalen? (z.B. Länge in cm, m; Temperatur in °C oder K) $\Rightarrow$ Zulässige (strukturerhaltende) Transformationen

Skalentypen (Messniveaus)

Die Skalentypen sind durch die Struktur des empirischen Relativs gegeben. Charakterisierung durch zulässige Transformationen.

$\rightarrow$ Existenz ist nicht immer gegeben

$\rightarrow$ Eindimensionalität könnte verletzt sein

Nominalskala

- Beispiele:
  Diagnosen, Geschlecht
- Struktur:
  keine
- Mögliche Aussagen:
  gleich, ungleich
- Erlaubte Transformationen:
  alle eineindeutigen Transformationen

$$a = b \Leftrightarrow f(a) = f(b)$$

Ordinal- oder Rangskala

- Beispiele:
  Schulbildung, soziale Schicht, Schweregrad einer Erkrankung
- Struktur:
  lineare Ordnung
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner
- Erlaubte Transformationen:
  alle positiv monotonen Transformationen

$$a < b \Rightarrow f(a) < f(b)$$

Intervallskala

- Beispiele:
  Ergebnisse psychometrischer Tests, Scores, Schulnoten?, Häufigkeit der Kommunikation, physiologische Daten (EKG)
- Struktur:
  Abstände definiert mit Axiomen
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner, Differenzen
- Erlaubte Transformationen:
  alle linearen Transformationen $y = ax + b$

$$f(x_1) - f(x_2) = f(x_3) - f(x_4) \Leftrightarrow x_1 - x_2 = x_3 - x_4$$

Verhältnisskala

Intervallskala mit Nullpunkt

- Beispiele:
  Fernsehdauer, Preis, Länge, Gewicht
- Struktur:
  Abstände definiert, Nullpunkt
- Mögliche Aussagen:
  gleich, ungleich, größer, kleiner, Differenzen, Verhältnis
- Erlaubte Transformationen:
  $y = ax$ (Multiplikation)

$$\frac{f(x_1)}{f(x_2)} = \frac{x_1}{x_2}$$

Absolutskala

- Beispiel:
  Häufigkeit
- Struktur:
  Einheit liegt auf natürliche Weise fest
- Erlaubte Transformationen:
  keine

Skalenniveau

Beachte:

- Je höher das Skalenniveau, desto mehr Interpretationen sind möglich
- Sinnvoll interpretierbare Berechnungen sollen invariant bezüglich der zulässigen Transformationen sein

	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

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{rcl} 3 \cdot 10^{-3} & & log(3 \cdot 10^{-3}) \\ 2 \cdot 10^{-4} & \text{oder} & log(2 \cdot 10^{-4}) \\ 2.2 \cdot 10^{-5} & & log(2.2 \cdot 10^{-5}) \end{array}$$

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{rcl} 3 \cdot 10^{-3} & & log(3 \cdot 10^{-3}) \\ 2 \cdot 10^{-4} & \text{oder} & log(2 \cdot 10^{-4}) \\ 2.2 \cdot 10^{-5} & & log(2.2 \cdot 10^{-5}) \end{array}$$

Beispiel: Tumorgröße

$$\begin{array}{rcl} 4mm^3 & & ? \\ 10mm^3 & \text{oder} & ? \\ 2mm^3 & & ? \end{array}$$

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{rcl} 3 \cdot 10^{-3} & & log(3 \cdot 10^{-3}) \\ 2 \cdot 10^{-4} & \text{oder} & log(2 \cdot 10^{-4}) \\ 2.2 \cdot 10^{-5} & & log(2.2 \cdot 10^{-5}) \end{array}$$

Beispiel: Tumorgröße

$$\begin{array}{rcl} 4mm^3 & & ? \\ 10mm^3 & \text{oder} & ? \\ 2mm^3 & & ? \end{array}$$

Skalenwahl $\Leftrightarrow$ Interpretation der Differenz

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{rcl} 3 \cdot 10^{-3} & & log(3 \cdot 10^{-3}) \\ 2 \cdot 10^{-4} & \text{oder} & log(2 \cdot 10^{-4}) \\ 2.2 \cdot 10^{-5} & & log(2.2 \cdot 10^{-5}) \end{array}$$

Beispiel: Tumorgröße

$$\begin{array}{rcl} 4mm^3 & & ? \\ 10mm^3 & \text{oder} & ? \\ 2mm^3 & & ? \end{array}$$

Skalenwahl $\Leftrightarrow$ Interpretation der Differenz

Bei log-Skala: Differenz = Faktor der Veränderung

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{rcl} 3 \cdot 10^{-3} & & log(3 \cdot 10^{-3}) \\ 2 \cdot 10^{-4} & \text{oder} & log(2 \cdot 10^{-4}) \\ 2.2 \cdot 10^{-5} & & log(2.2 \cdot 10^{-5}) \end{array}$$

Beispiel: Tumorgröße

$$\begin{array}{rcl} 4mm^3 & & ? \\ 10mm^3 & \text{oder} & ? \\ 2mm^3 & & ? \end{array}$$

Skalenwahl $\Leftrightarrow$ Interpretation der Differenz

Bei log-Skala: Differenz = Faktor der Veränderung

Verwende log zur Basis 10

Gütekriterien der Messung

Genauigkeit (Accuracy)

Messfehlermodelle: Es gibt einen wahren Wert X und eine Messung X^* z.B.

$$X^* = X + \underbrace{U}_{\textit{Messfehler}}$$

$$\mathbb{E}(U) = 0$$

$\Rightarrow$ klassischer additiver Messfehler
Modell der klassischen (psychologischen) Testtheorie

Gütekriterien der Messung

Genauigkeit (Accuracy)

Messfehlermodelle: Es gibt einen wahren Wert X und eine Messung X^* z.B.

$$X^* = X + \underbrace{U}_{\textit{Messfehler}}$$

$$\mathbb{E}(U) = 0$$

$\Rightarrow$ klassischer additiver Messfehler
Modell der klassischen (psychologischen) Testtheorie

Aspekte:

- Validität = Gültigkeit
- Reliabilität = Zuverlässigkeit

Validität

Frage: Wird das gemessen, was gemessen werden soll?

In der emp. Sozialforschung:

- Inhaltsvalidität
- Kriteriumsvalidität
- Konstruktvalidität

Statistik: systematischer Messfehler? $\mathbb{E}(U) = 0$

Reliabilität

Ist die Messung zuverlässig?
Erhält man bei Wiederholung den gleichen Wert?
Erhält man unter verschiedenen Bedingungen den gleichen Wert?
$\rightarrow$ Interrater Reliabilität

Abhängig von:
Verhältnis von Streuung des Messfehlers zur Gesamtstreuung

$$r = \frac{\sigma_X^2}{\sigma_{X^*}^2} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$$

$\sigma_X^2, \sigma_{X^*}^2$ Varianz von X bzw. X^*
$\rightarrow$ Interne Konsistenz: Cronbachs Alpha (später)

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 ?

Skalierungsverfahren mit latenten Variablen

Grundlage: Modell mit latenter Größe, die das Antwortverhalten oder Lösen von Aufgaben bestimmt

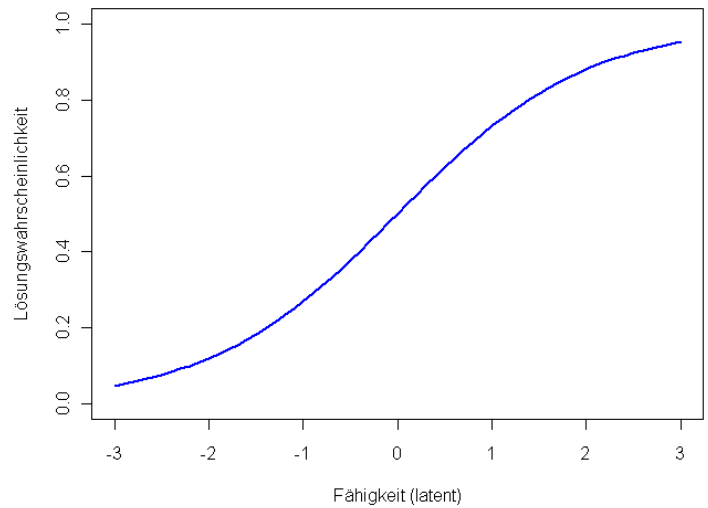
Wichtigste Beispiele :

- Likert Skalen
- Faktorenmodell
- Rasch - Modell

Itemcharakteristik

Grundidee: Wahrscheinlichkeit von Lösung der Aufgabe bzw. Antwort „Ja“ hängt von der latenten Variable ab.

$$P(\text{Item gelöst}) = G(\text{latente Größe})$$



Je größer die Fähigkeit desto größer die Lösungswahrscheinlichkeit

Ausblick: Statistische Verfahren zur Überprüfung der Messung (Skalierung)

- Itemanalyse
- Rasch-Modell
- Faktorenanlyse
- Analyse von Wiederholungs- und Mehrfachmessungen
- Messfähigkeitsanlyse in der Qualitätskontrolle

Weiter: Verfahren zur **Berücksichtung** von Messfehlern

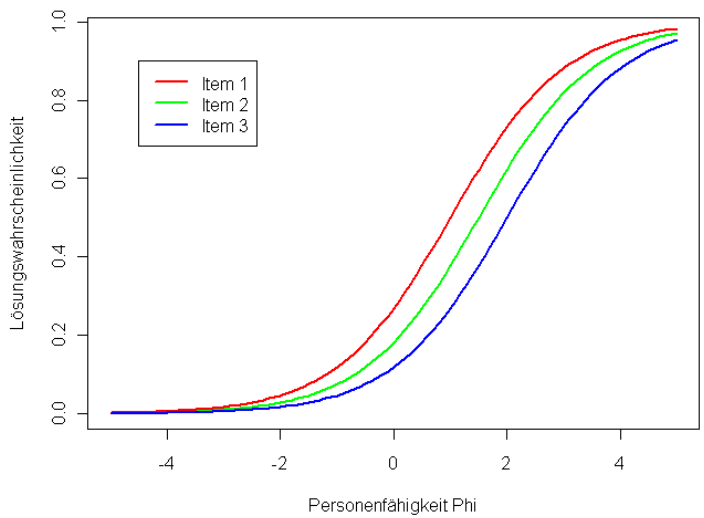
Rasch-Modell

Items haben verschiedene Schwierigkeiten

Lösungswahrscheinlichkeiten lassen sich durch Personenfähigkeit und Itemschwierigkeit beschreiben:

$$P(\text{Item gelöst}) = G(\Phi_i - b_j)$$

Φ_i : Fähigkeit von Person i
b_j : Itemschwierigkeit



Lit.: Carolin Strobl (2010): Das Rasch-Modell, Rainer-Hampp-Verlag.

Merkmalstypen: Stetige und diskrete Merkmale

- Diskretes Merkmal endlich oder abzählbar unendlich viele verschiedene Werte
  Beispiele: Geschlecht, Kinderanzahl,...
- Stetiges Merkmal
  alle Werte eines Intervalls können angenommen werden
  Beispiele: Zeitdauern, Größe, Gewicht,...

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

Epidemiologische Studien

- Kohortenstudien: Längsschnittstudien (Retrospektiv oder prospektiv)
  Beispiel: EPIC Studie (European Prospective Investigation into Cancer) 400000 Personen in neun europäischen Ländern

Epidemiologische Studien

- Kohortenstudien: Längsschnittstudien (Retrospektiv oder prospektiv)
  Beispiel: EPIC Studie (European Prospective Investigation into Cancer) 400000 Personen in neun europäischen Ländern
- Fall - Kontroll-Studien Erhebung von erkrankten (Fälle) und Kontrollen
  Beispiel: Deutsche Radon Studie

Epidemiologische Studien

- Kohortenstudien: Längsschnittstudien (Retrospektiv oder prospektiv)
  Beispiel: EPIC Studie (European Prospective Investigation into Cancer) 400000 Personen in neun europäischen Ländern
- Fall - Kontroll-Studien Erhebung von erkrankten (Fälle) und Kontrollen
  Beispiel: Deutsche Radon Studie
- Querschnittsstudien
  Beispiel: 1997 - 1999 Erstes gesamtdeutsches Gesundheitssurvey (7124 Personen)

Experimente

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

- Randomisierte klinische Studie: Zuordnung von Einheiten zu Behandlungen erfolgt durch Losverfahren (Randomisierung)
- Randomisierte Experimente (Produktion, Landwirtschaft) (Vorlesung Versuchplanung)
- Experimente in Medizin und Biologie
- Naturwissenschaftliche Experimente mit zufälligen Komponenten

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: Befragung von Redakteuren

Bitte füllen Sie diesen Fragebogen nur aus, wenn Sie Chefredakteur bzw. Redaktionsleiter einer Print-Zeitung sind

Sehr geehrter Teilnehmer,
zunächst haben wir einige allgemeine Fragen zur Organisation Ihrer Redaktion:

1. *Die Redaktionen von Print-Zeitungen in Deutschland sind unterschiedlich groß. Wie viele Journalisten (festangestellte und freie) arbeiten in der Stammredaktion Ihrer Print-Tageszeitung?*

_____ festangestellte Redakteure und _____ freie Mitarbeiter.

2. *In jeder Redaktion gibt es verschiedene Positionen zu besetzen. Bitte geben Sie an, welche der folgenden Positionen es in Ihrer Print-Redaktion gibt und wie oft sie gegebenenfalls besetzt sind (also z.B. „2" wenn es zwei Chefs vom Dienst gibt).*

Es gibt.....

___	(Anzahl)	Chefredakteur(e)
	(Anzahl)	Stellvertretende(n) Chefredakteur(e)
___	(Anzahl)	Chef(s) vom Dienst
	(Anzahl)	Ressortleiter
___	(Anzahl)	Leitende(n) Redakteur(e)
___	(Anzahl)	weitere Position und zwar _____

3. *Der Alltag von Journalisten wird durch verschiedene Tätigkeiten bestimmt. Bitte geben Sie an, wie intensiv die Print-Redakteure die folgenden Tätigkeiten im Redaktionsalltag ausüben.*

	täglich	mehrmals pro Woche	einmal pro Woche	mehrmals pro Monat	einmal pro Monat	seltener als einmal pro Monat	nie
Verfassen eigener Artikel	☐	☐	☐	☐	☐	☐	☐
Redigieren von Agenturmeldungen/ Pressemitteilungen	☐	☐	☐	☐	☐	☐	☐
Redigieren von Beiträgen anderer Autoren	☐	☐	☐	☐	☐	☐	☐
Recherche vor Ort	☐	☐	☐	☐	☐	☐	☐
Recherche vom Schreibtisch aus	☐	☐	☐	☐	☐	☐	☐
Bearbeiten von Fotos	☐	☐	☐	☐	☐	☐	☐
Technische Produktion/ Layout der Beiträge	☐	☐	☐	☐	☐	☐	☐

Eindimensionale Häufigkeitsverteilung

- Ordnen der Daten nach einem Merkmal
- Auszählen der Häufigkeiten der einzelnen Merkmalsausprägungen
- Relative Häufigkeiten = Häufigkeit/Anzahl der Untersuchungseinheiten
- Kumulative Häufigkeiten bei ordinal oder metrisch skalierten Merkmalen sinnvoll:
  $F(x)$:=(Summe der relativen Häufigkeiten $\leq$ x)
  empirische Verteilungsfunktion

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, Urliste

Häufigkeiten I

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

Beispiel: Absolventenstudie
Für die Variable D "Ausrichtung der Diplomarbeit" ist die Urliste 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*

Bemerkungen:

- Wenn statt der Urliste bereits die Ausprägungen $a_1, \ldots, a_k$ und die Häufigkeiten $f_1, \ldots, f_k$ bzw. $h_1, \ldots, h_k$ vorliegen, sprechen wir von *Häufigkeitsdaten*.
- Klassenbildung, gruppierte Daten:
  Bei metrischen, stetigen (oder quasi-stetigen) Merkmalen oft Gruppierung der Urliste durch Bildung geeigneter Klassen

Beispiel Nettomieten I

Wir greifen aus dem gesamten Datensatz die Wohnungen ohne zentrale Warmwasserversorgung (zh=1) und mit einer Wohnfläche kleiner als 50 qm (wfl<50) heraus. Die folgende Urliste zeigt, bereits der Größe nach geordnet, die Nettomieten dieser $n = 27$ Wohnungen:

81.28	98.85	109.32	130.35	132.24	151.00	163.41
172.23	181.98	183.09	195.72	203.75	224.61	229.06
244.50	268.36	272.24	275.52	314.09	352.79	353.69
357.05	373.37	388.81	389.23	412.61	463.40	

Alle Werte verschieden
$\Rightarrow k = n$ und $\{x_1, \ldots, x_n\} = \{a_1, \ldots, a_k\}$
$\Rightarrow f_j = \frac{1}{27}, \quad j = 1, \ldots, 27.$

Beispiel Nettomieten II

Selektion dieser Daten in R:

```
daten <- read.table(file="miete03.asc", sep="\t", header=T)
daten <- subset(daten, (zh==1) \& (wfl<50) )
attach(daten)
print(sort(nm))
```

(nm=Nettomiete)

Beispiel Nettomieten III

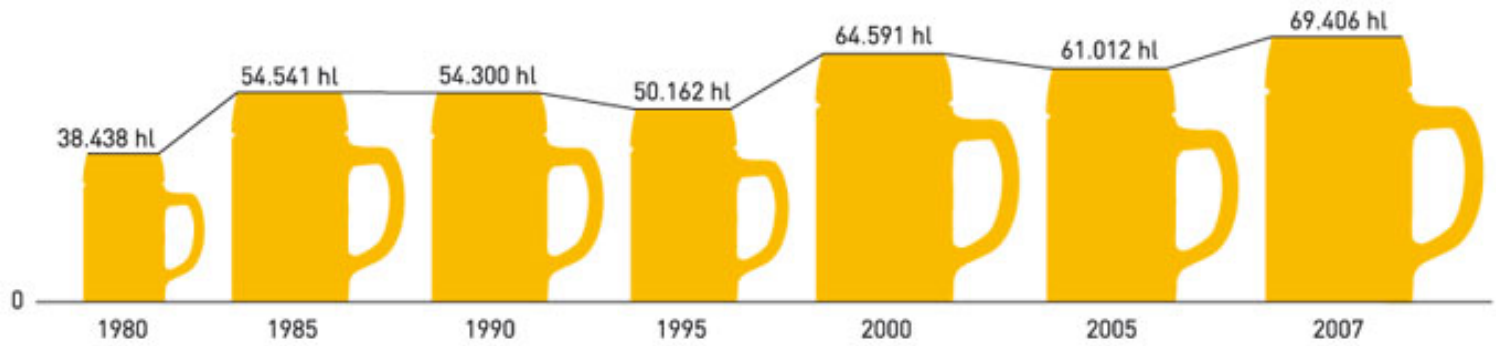
Gruppiert man die Urliste in 5 Klassen mit gleicher Klassenbreite von 100 EURO, so erhält man folgende Häufigkeitstabelle:

Klasse	absolute Häufigkeit	relative Häufigkeit
$50 < \ldots \leq 150$	5	$5/27 = 0.185$
$150 < \ldots \leq 250$	10	$10/27 = 0.370$
$250 < \ldots \leq 350$	4	$4/27 = 0.148$
$350 < \ldots \leq 450$	7	$7/27 = 0.259$
$450 < \ldots \leq 550$	1	$1/27 = 0.037$

Häufigkeiten für gruppierte $n = 27$ Nettomieten

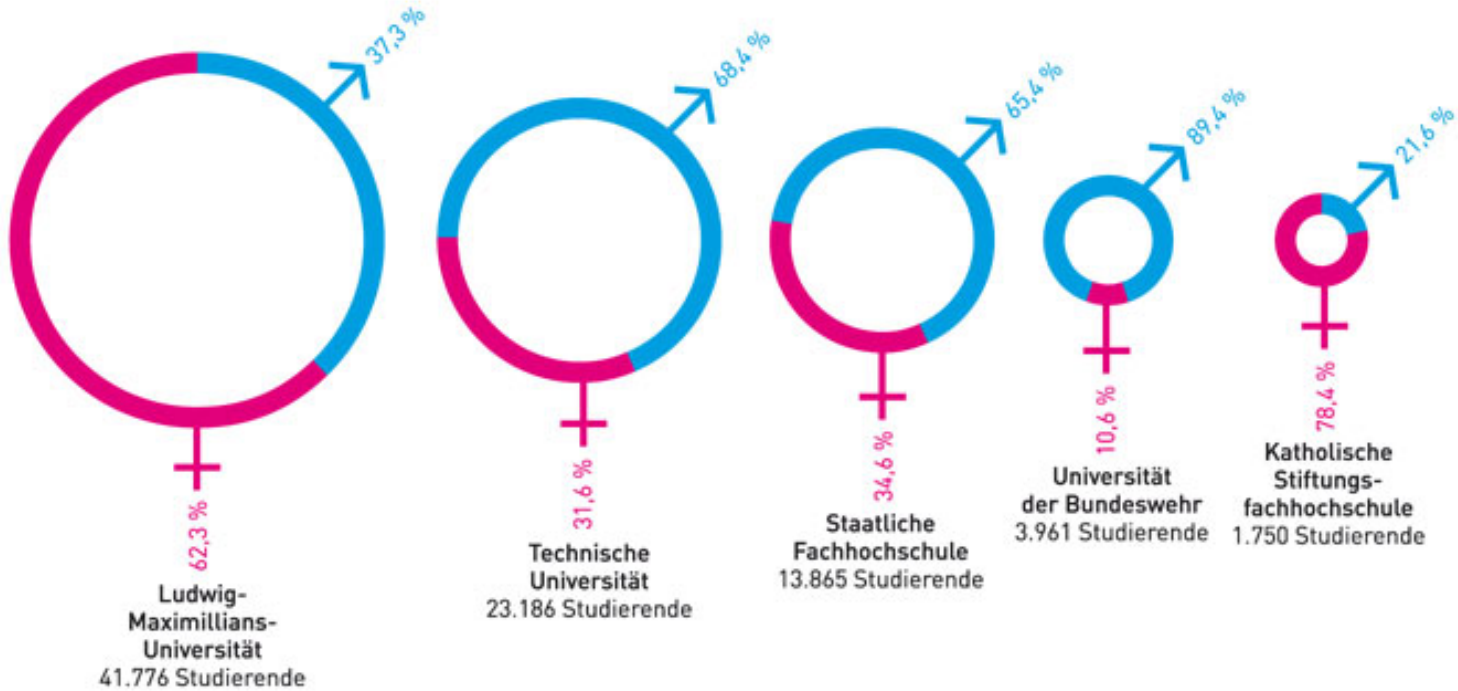
Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“



Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“

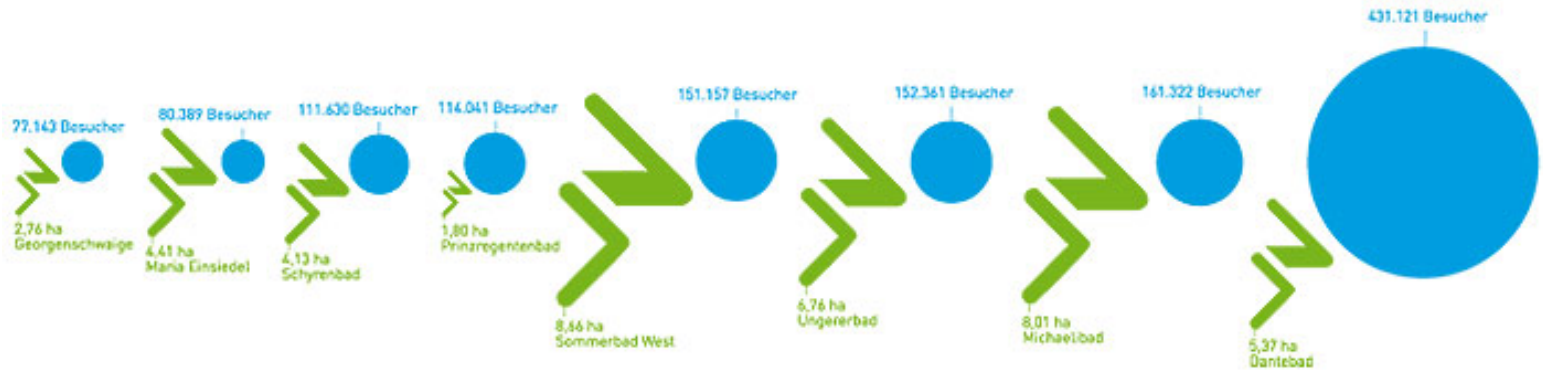


Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“

Grafische Darstellungen

„Ein Bild sagt mehr als tausend Worte“



Statistische Grafik

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

Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data - a matter of *substance*, of *statistics* and of *design*.
- Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Graphical excellence is nearly always multivariate.
- And graphical excellence requires telling the truth about the data.

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)

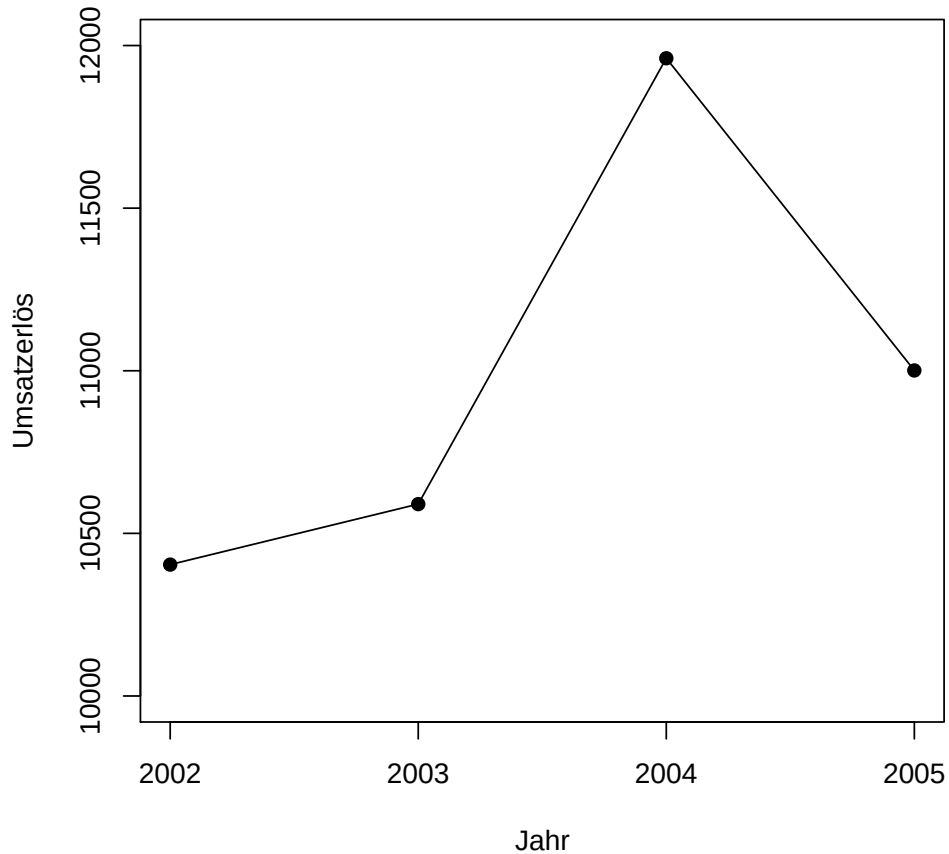
1. Abstände
2. Winkel
3. Flächen
4. Volumen
5. Farbton-Sattheit-Schwärzegrad

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

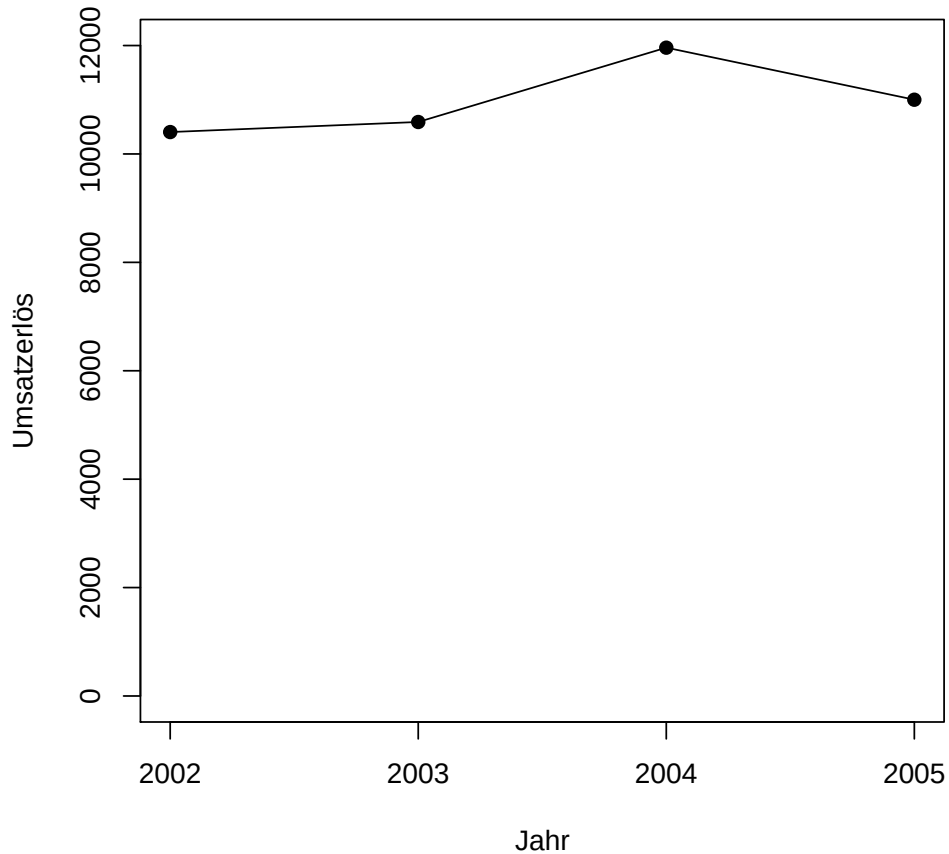
Typen von eindimensionalen Darstellungen

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

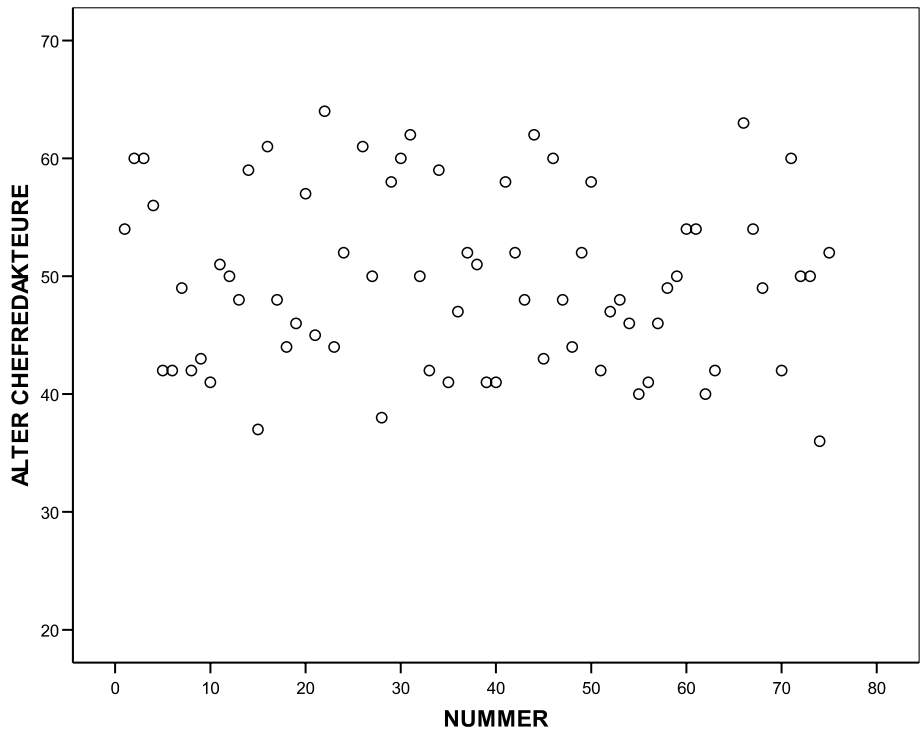
Beispiel: Liniendiagramm (??)



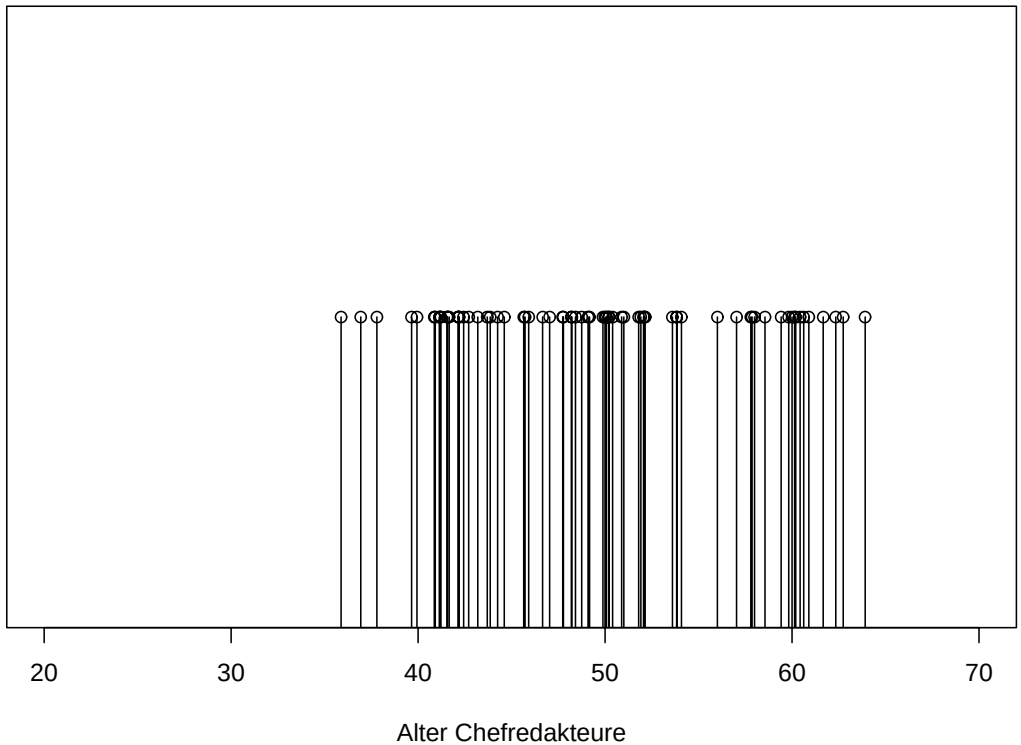
Beispiel: Liniendiagramm (!!)



Beispiel: Streudiagramm



Beispiel: Needleplot



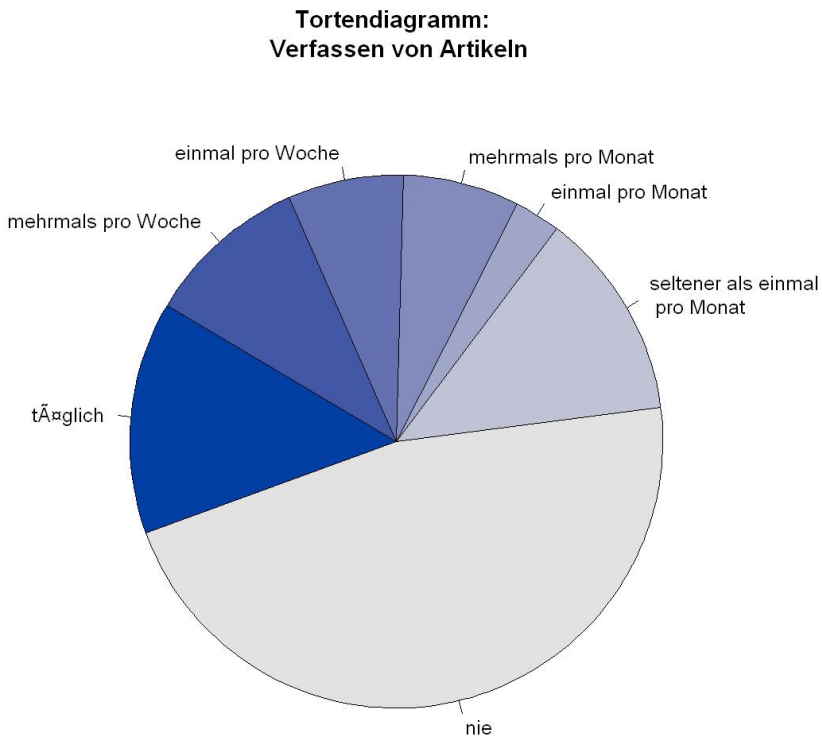
Kreisdiagramm, Tortendiagramm

Darstellung der relativen (absoluten) Häufigkeiten als Fläche eines Kreises

Anwendung:

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

Beispiel Redakteure: Kreisdiagramm



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

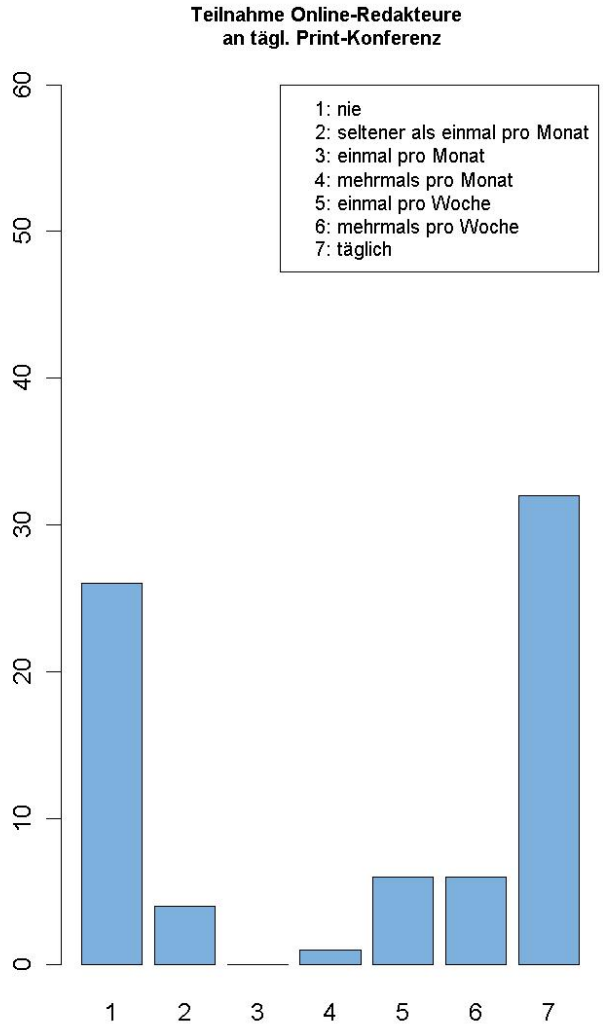
Darstellung der absoluten oder relativen Häufigkeiten als Höhen (Längen)

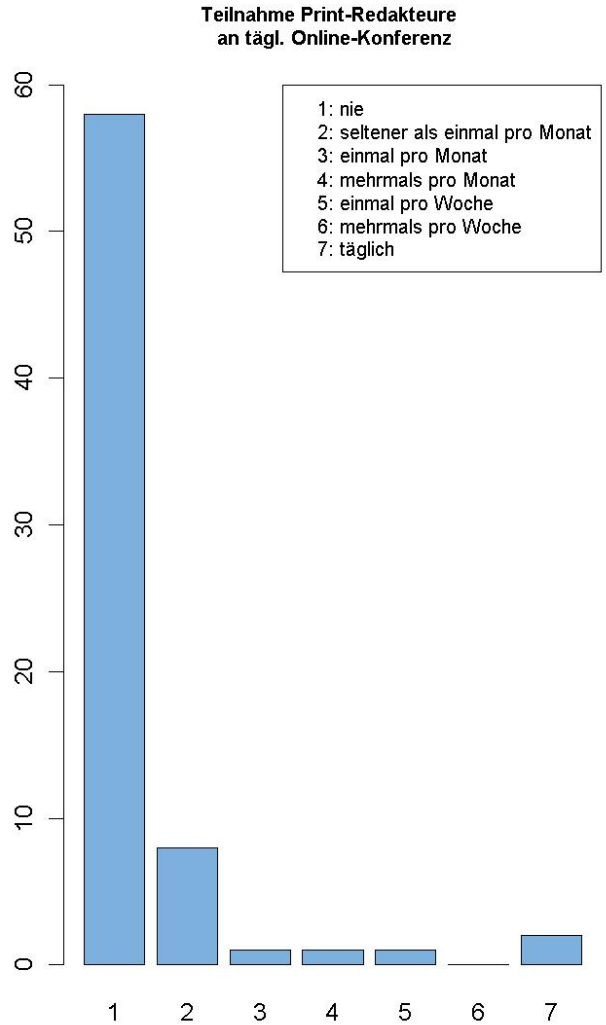
x-Achse: Ausprägungen des Merkmals
y-Achse: absolute/ relative Häufigkeiten

Anwendungen:

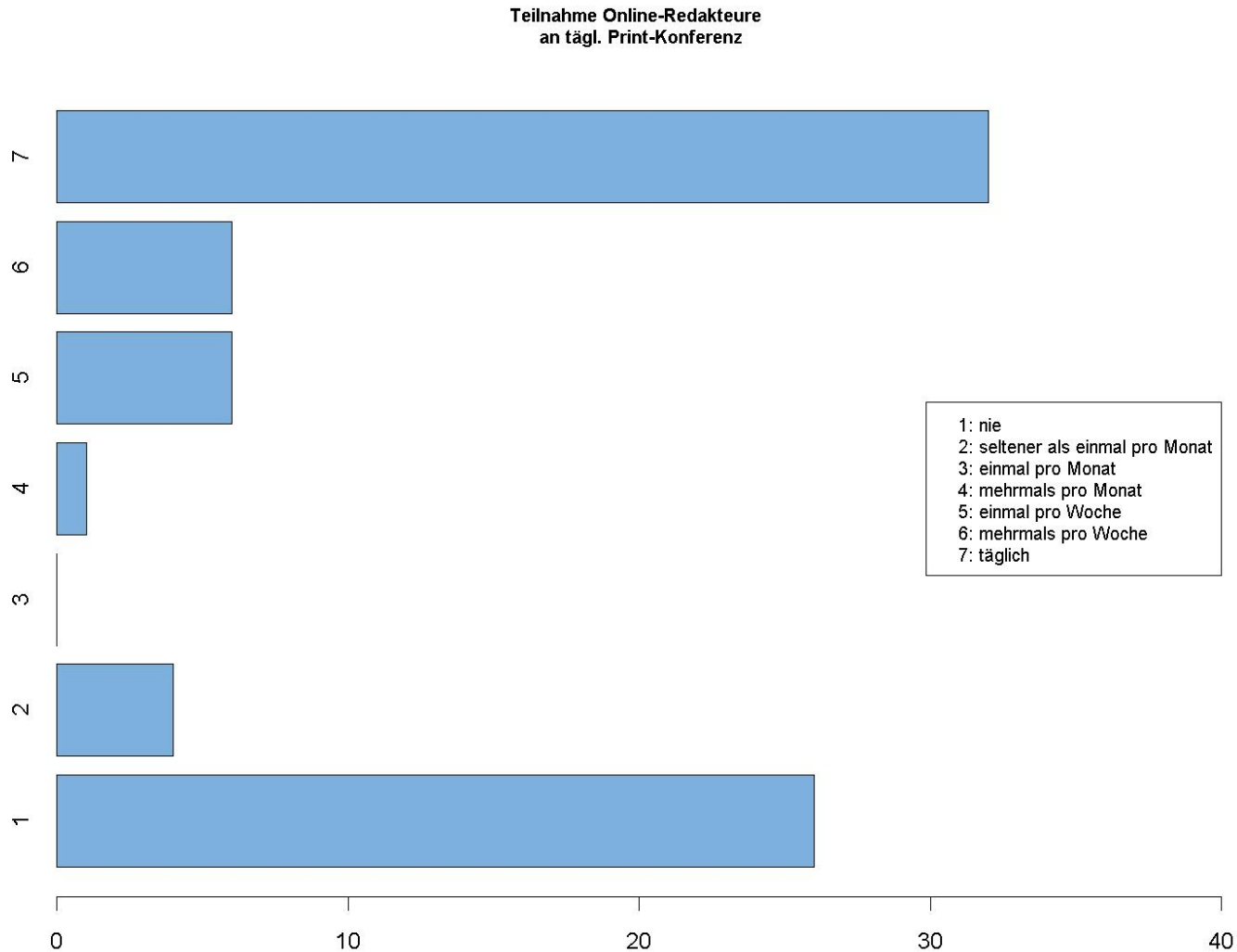
- Ordinale Merkmale
- Metrische Merkmale mit wenigen Ausprägungen
- Nominale Merkmale (Problem: Ordnung nicht vorhanden)

Beispiel Redakteure: Säulendiagramm vertikal





Beispiel Redakteure: Säulendiagramm horizontal



Stapeldiagramm

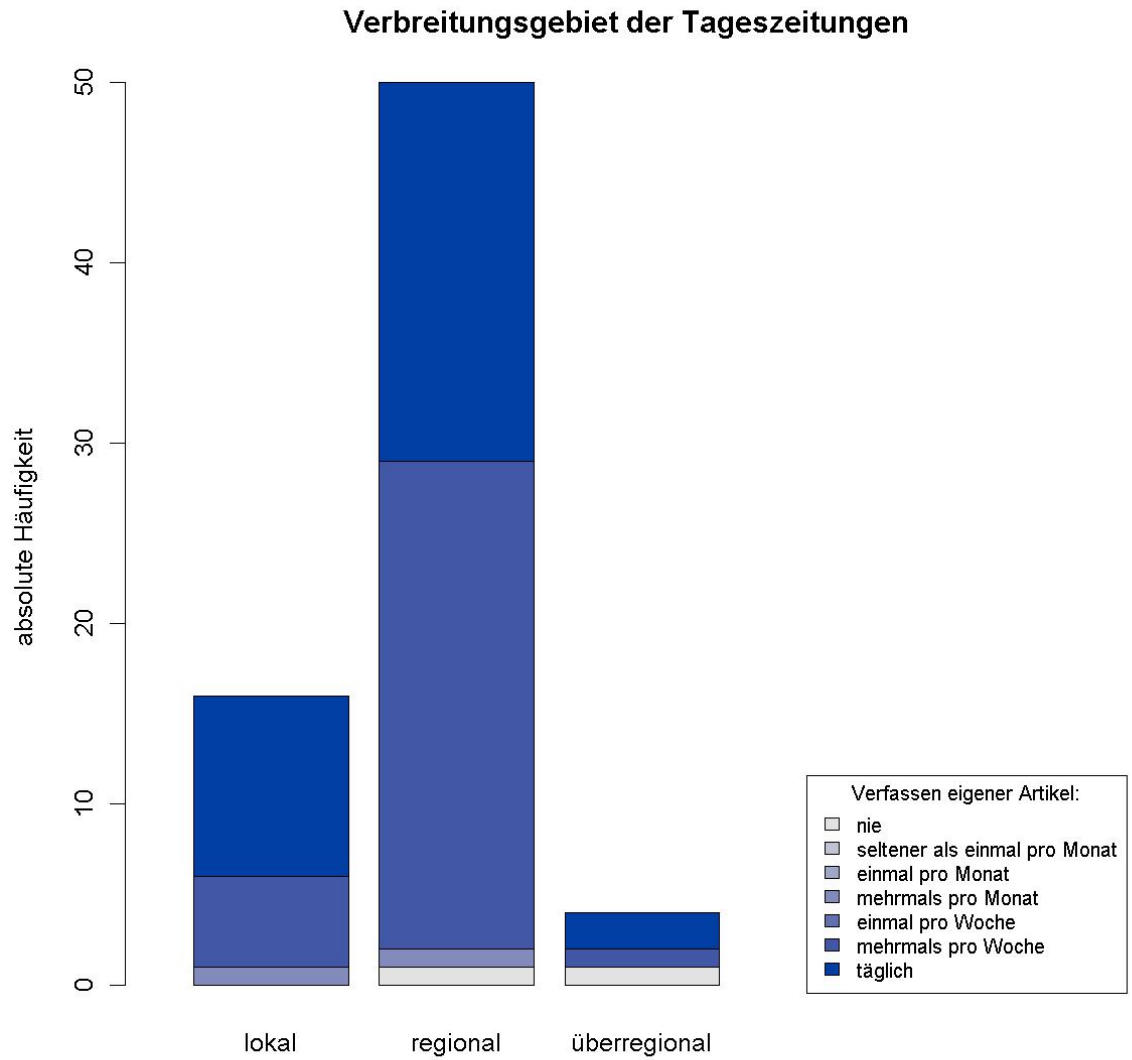
Darstellen der absoluten oder relativen Häufigkeiten als Länge. Die Abschnitte werden übereinander in verschiedenen Farben gestapelt.

Anwendungen:

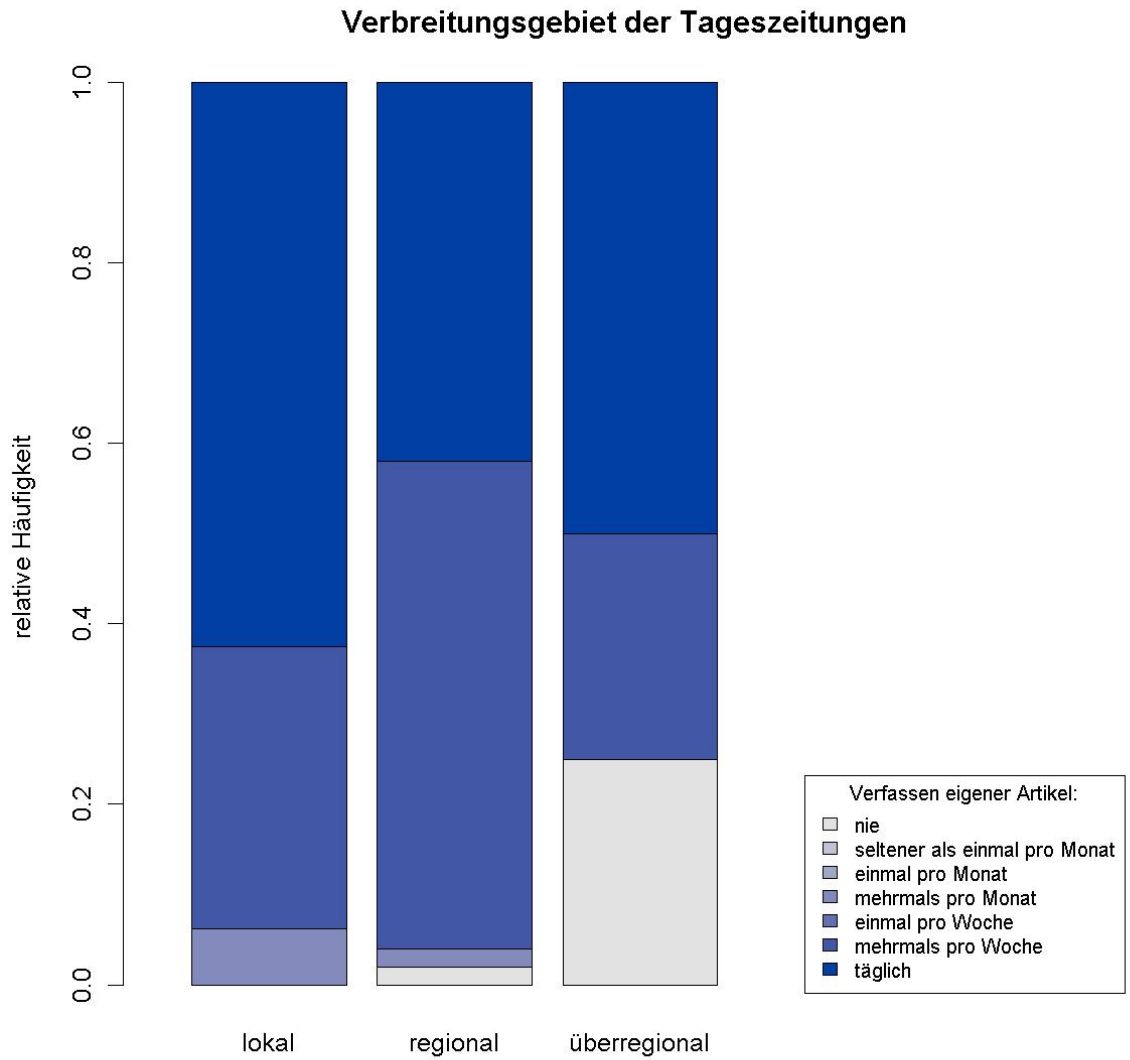
- Ordinale Daten
- Gruppierte Daten
- Metrische Daten mit wenigen Ausprägungen

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

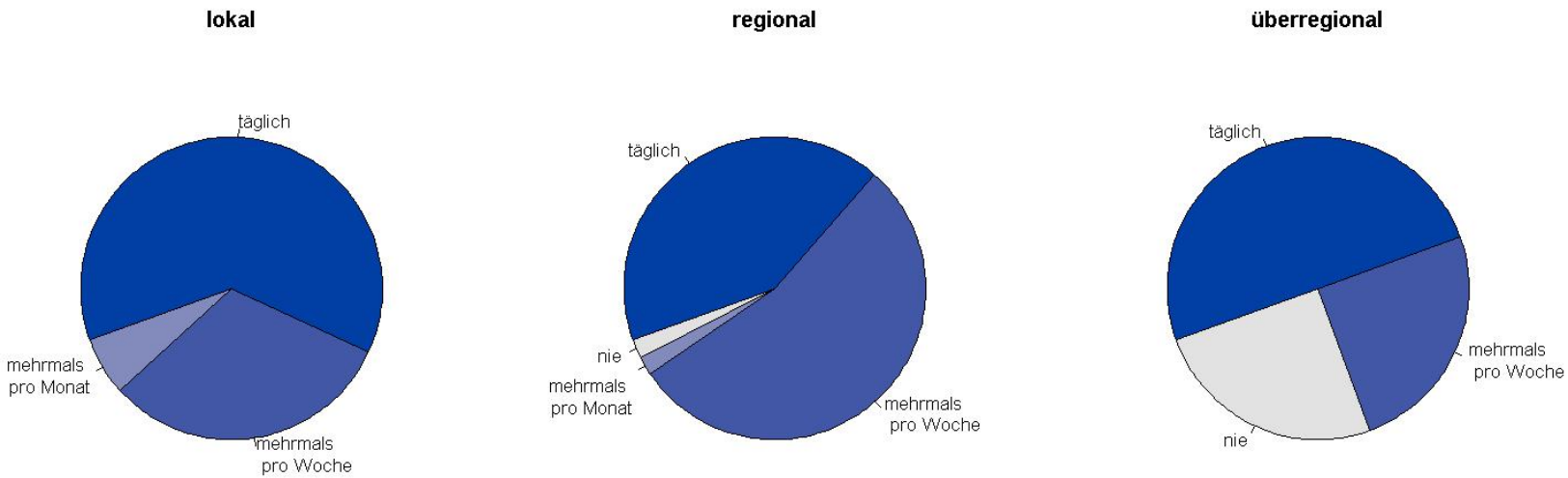
Beispiel Redakteure: Stapeldiagramm I



Beispiel Redakteure: Stapeldiagramm II



Beispiel Redakteure: Vergleich mit Kreisdiagramm



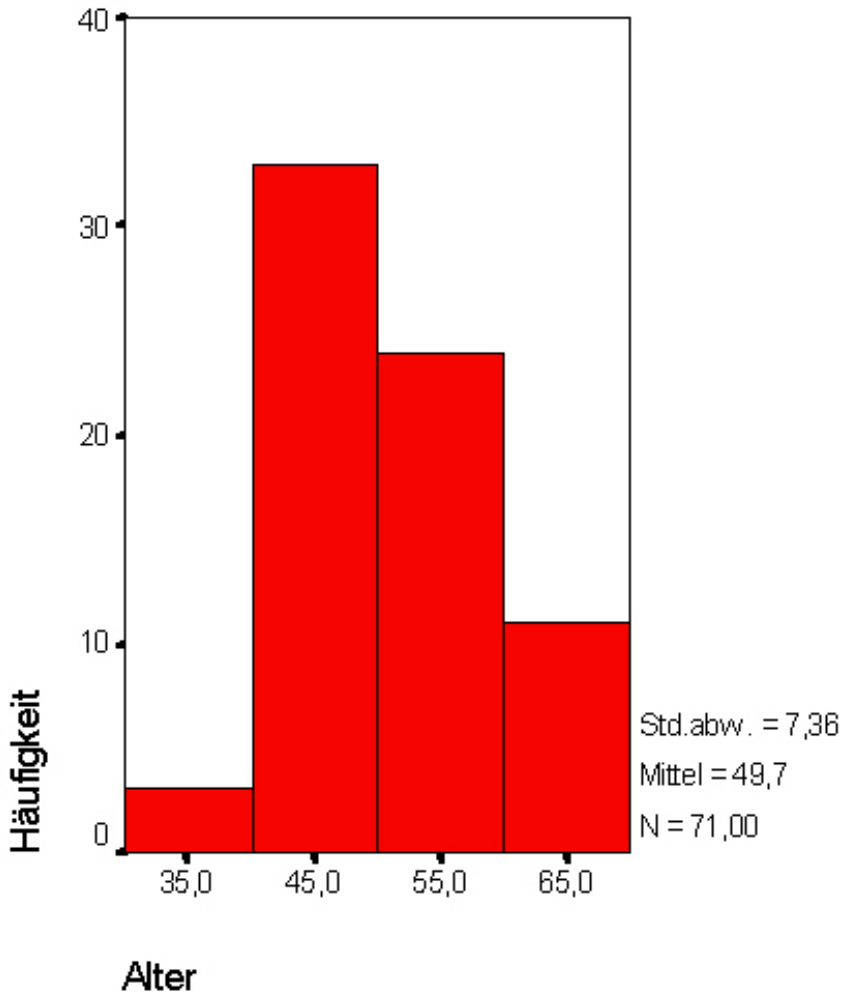
Das Histogramm

Darstellung der relativen Häufigkeiten durch Flächen
(Prinzip der Flächentreue)

Vorgehen:

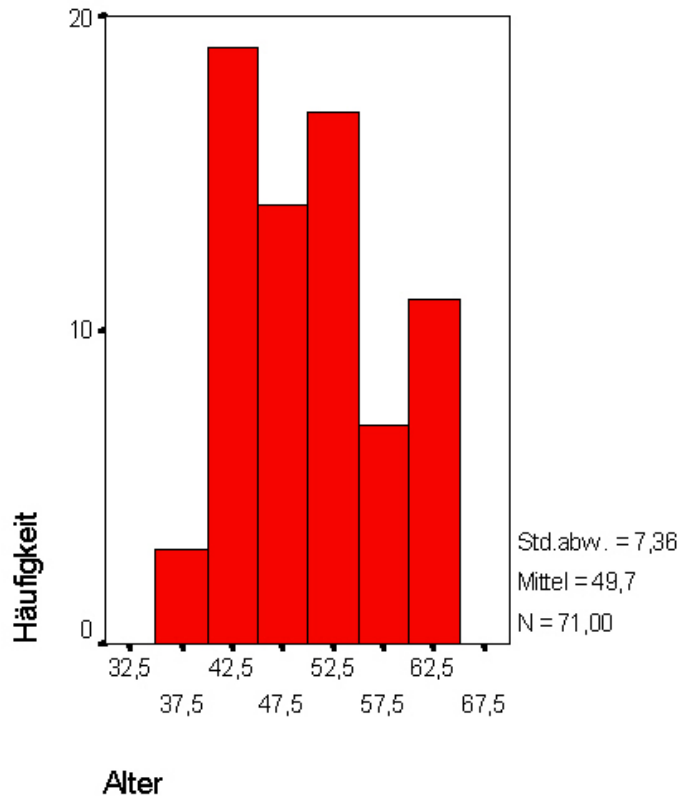
1. Aufteilung in Klassen (falls die Daten noch nicht gruppiert sind)
2. Bestimmung der relativen Häufigkeiten $f_j = \frac{n_j}{n}$
3. Bestimmung der Höhen h_j, so dass gilt $b_j \cdot h_j = f_j$
   wobei b_j: Breite der Klasse j.

Beispiel: Alter der Redakteure

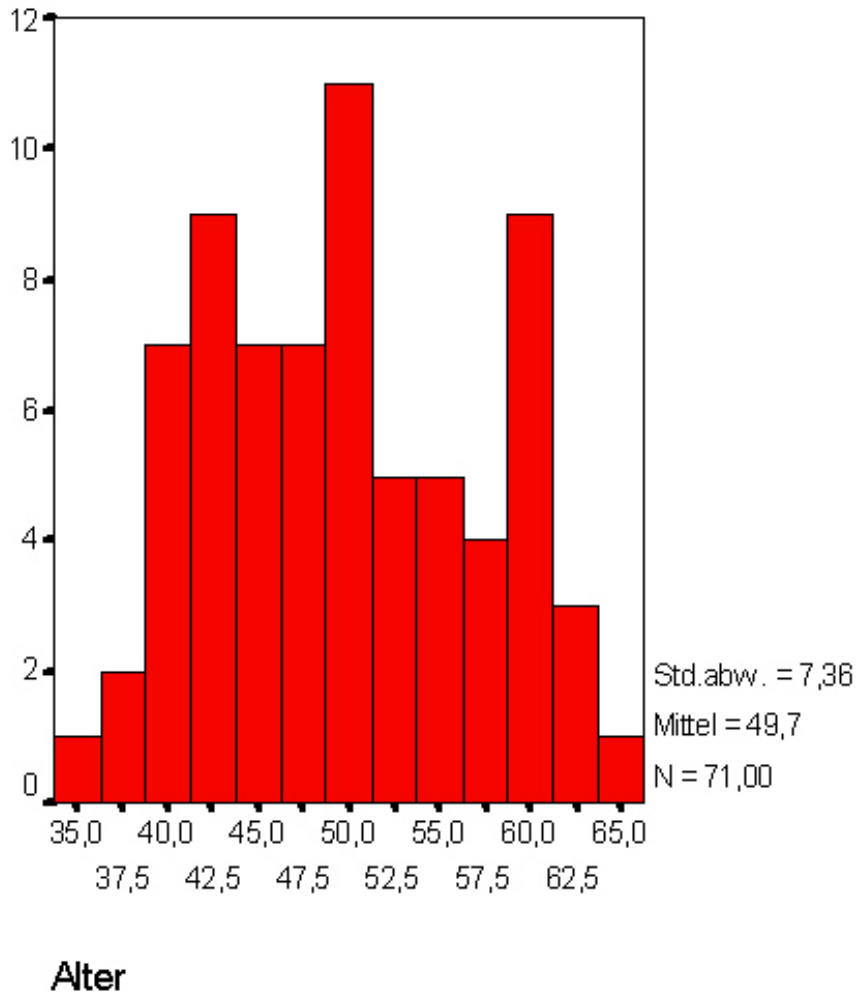


Beispiel: Alter der Redakteure

Altersklassen in Abständen von 5 Jahren



Histogramm mit Standardeinstellung aus SPSS



Histogramm

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

Stamm-Blätter-Diagramm

(Stem and leaf plot)

Spezielles Histogramm mit

- Klassen nach Dezimalsystem
- Einzeldaten reproduzierbar

Beispiel: Alter der Redakteure

Alter Stem-and-Leaf Plot

Frequency	Stem & Leaf
3,00	3 . 678
19,00	4 . 00111112222222233444
14,00	4 . 56667788888999
17,00	5 . 00000011222224444
7,00	5 . 6788899
11,00	6 . 00000112234

Stem width: 10

Each leaf: 1 case(s)

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.

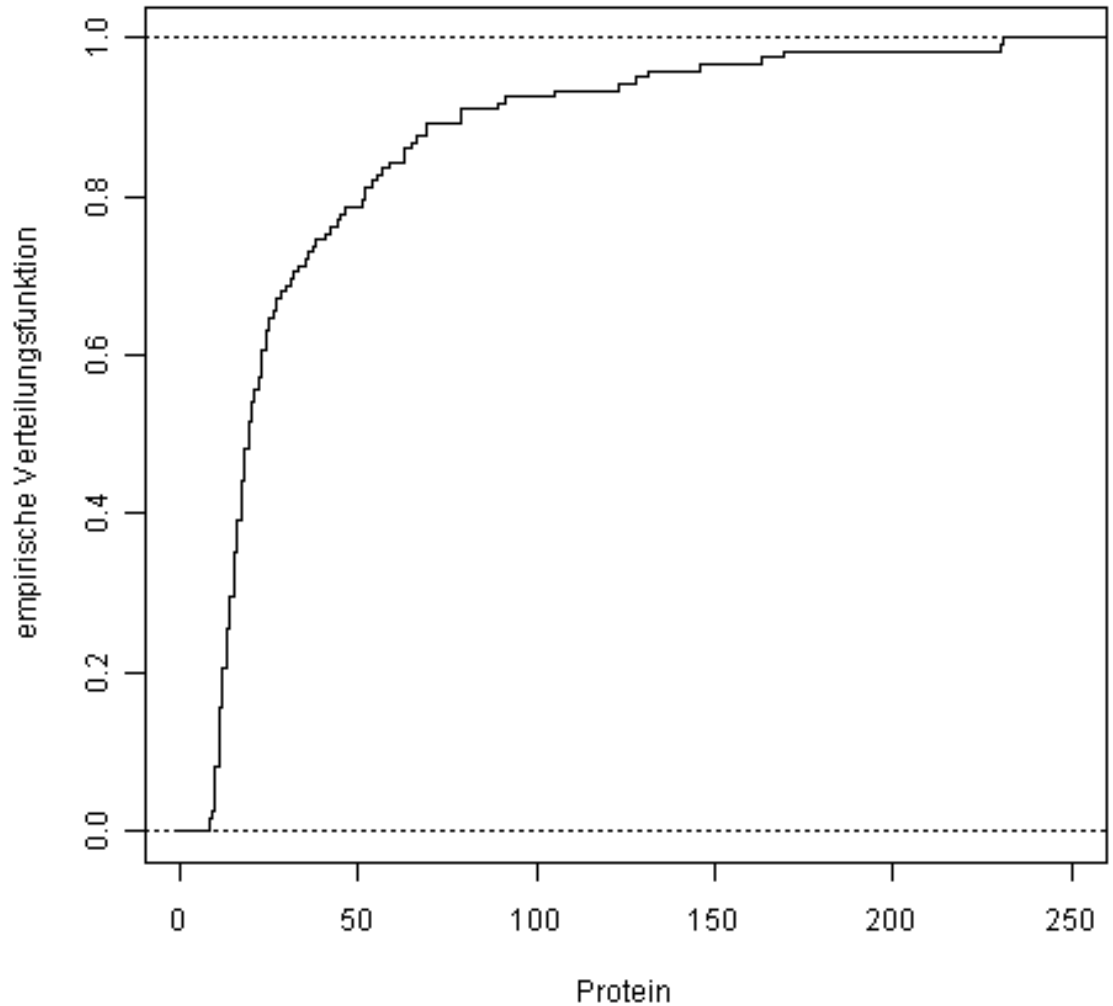
$$F(x) = f(a_1) + \ldots + f(a_j) = \sum_{i:a_i \leq x} f_i,$$

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

Eigenschaften von $F(x)$

- monoton wachsende Treppenfunktionen mit Sprüngen an den Ausprägungen $a_1, \ldots, a_k$
- Sprunghöhen: $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$
- rechtsseitig stetig
- $H(x) = 0$ für $x < a_1$, $H(x) = n$ für $x \geq a_k$
  $F(x) = 0$ für $x < a_1$, $F(x) = 1$ für $x \geq a_k$

Beispiel für eine Empirische Verteilungsfunktion



Eindimensionale statistische Kennwerte

Lagemaßzahlen

- 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?

Statistische Kennwerte

Streumaßzahlen

- Über welchen Bereich erstrecken sich die Daten?
- Wie groß ist die Schwankung der Ausprägungen?

Der Modus

Definition: Häufigster Wert

Eigenschaften:

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

Der Median

Definition: Wert für den gilt

Mindestens 50% der Daten sind kleiner oder gleich *med*
Mindestens 50% der Daten sind größer oder gleich *med*

$$med = \begin{cases} x_{(k)} & \text{falls } k = \frac{n+1}{2} \text{ ganze Zahl} \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & \text{falls } k = \frac{n}{2} \text{ ganze Zahl} \end{cases}$$

$x_{(1)}, \ldots, x_{(n)}$ sind geordnete Werte

Eigenschaften des Medians

- anschaulich
- stabil gegenüber monotonen Transformationen
- geeignet für ordinale Daten
- stabil gegenüber Ausreißern

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.

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}.

Boxplot

Einfacher Boxplot

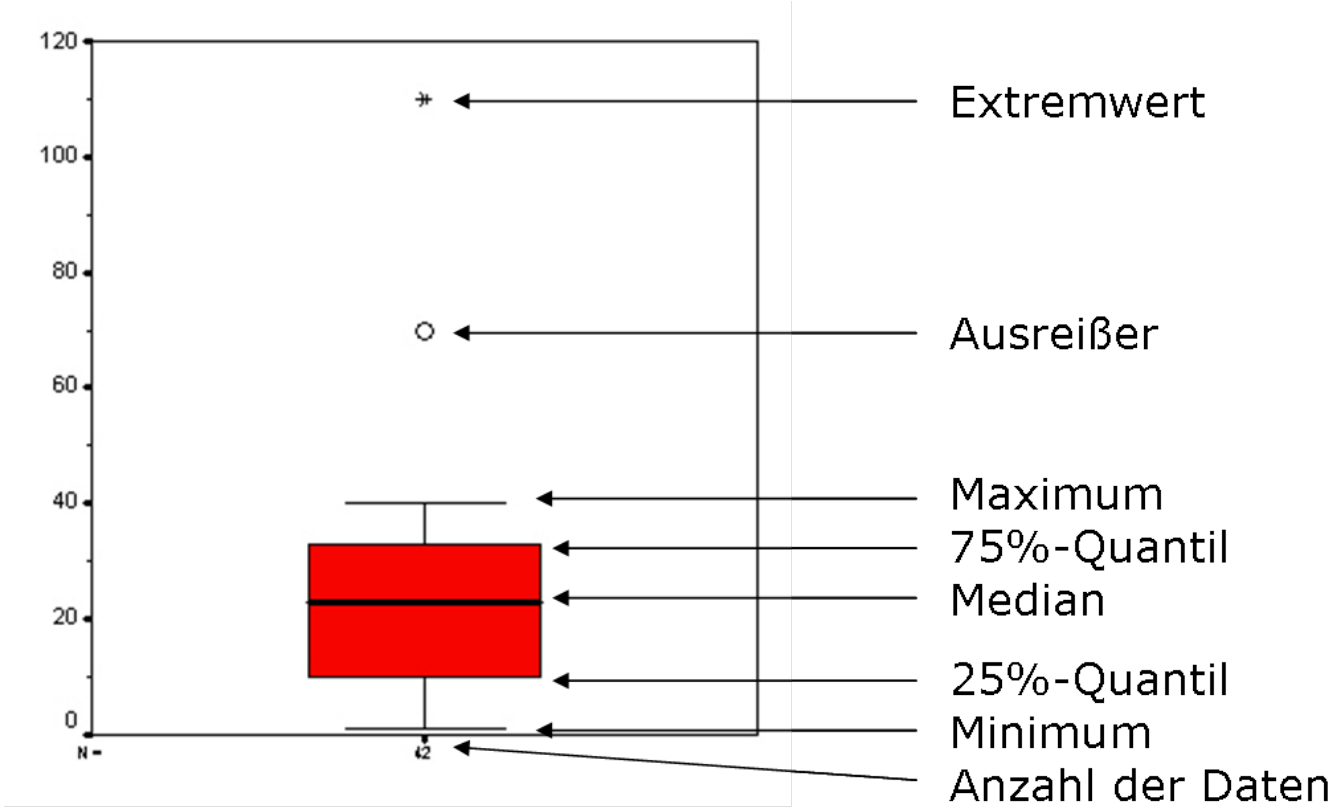
- $\tilde{x}_{0.25}$ = Anfang der Schachtel (Box)
  $\tilde{x}_{0.75}$ = Ende der Schachtel
  d_Q = Länge der Schachtel
- Der Median wird durch den Strich in der Box markiert
- Zwei Linien („whiskers“) außerhalb der Box gehen bis zu x_{min} und x_{max}.

Modifizierter Boxplot

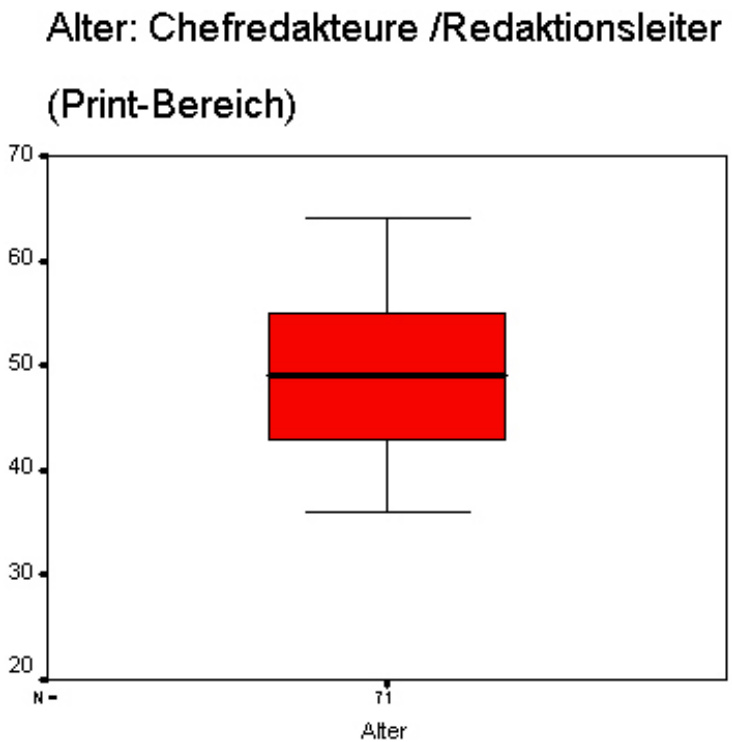
Die Linien außerhalb der Schachtel werden nur bis zu x_{min} bzw. x_{max} gezogen, falls x_{min} und x_{max} innerhalb des Bereichs $[z_u, z_o]$ der Zäune liegen.
$z_u = \tilde{x}_{0.25} - 1,5 d_Q$, $z_o = \tilde{x}_{0.75} + 1,5 d_Q$
Ansonsten gehen die Linien nur bis zum kleinsten bzw. größten Wert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

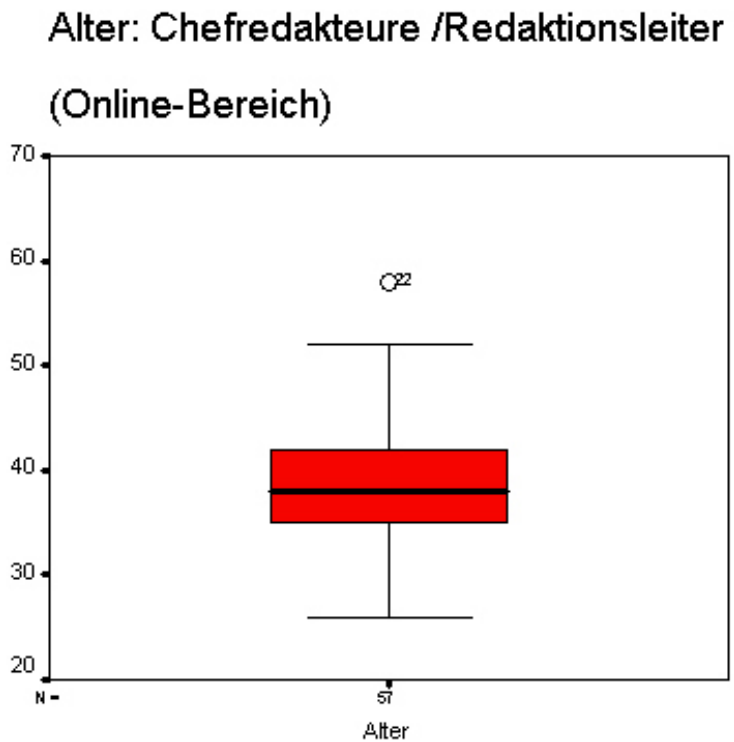
Boxplot

- Eindimensionale Darstellung auf der zugehörigen Skala
- Visualisieren der 5-Punkte-Zusammenfassung

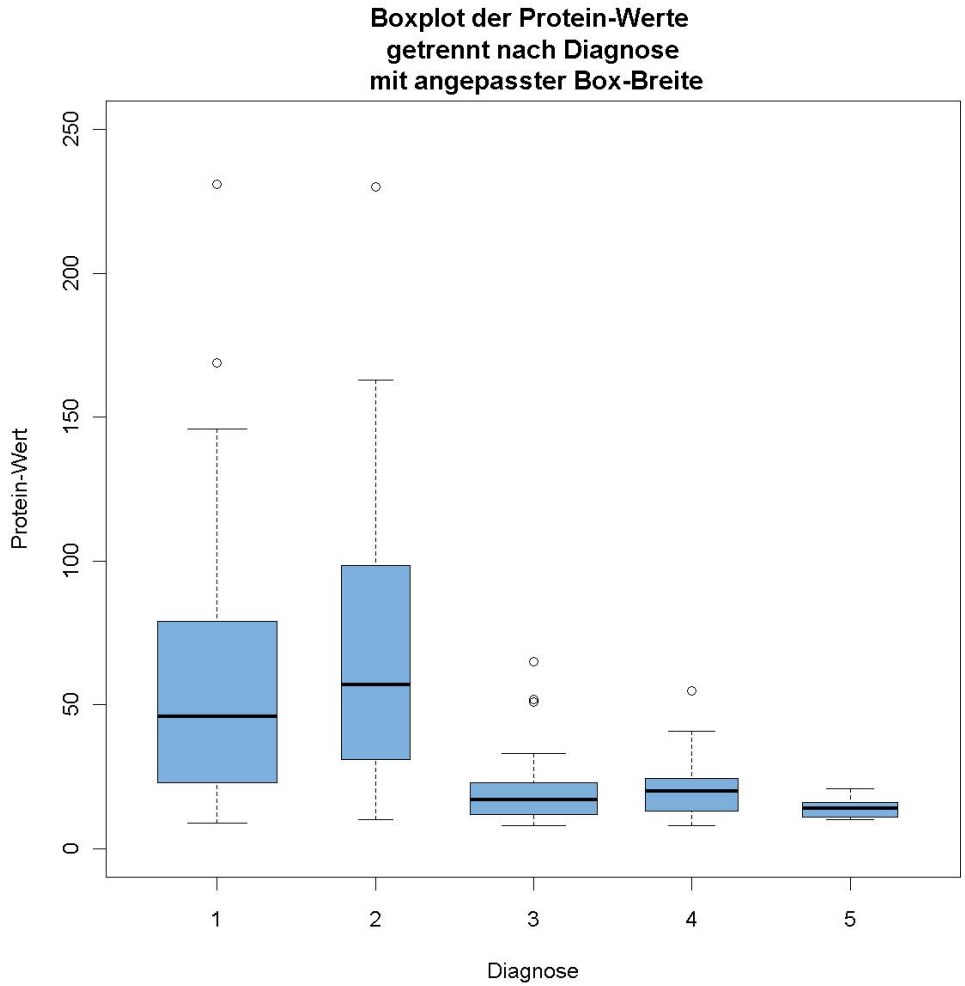


Beispiel Redakteure: Boxplot





Beispiel: Boxplot für Gruppen



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

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:

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

Allgemeine Transformation des Mittelwerts I

Lineare Transformation:

$$\begin{aligned} g(t) &= a + bt \\ y_i &= a + bx_i \quad \Rightarrow \quad \bar{y} = a + b\bar{x} \end{aligned}$$

d.h.

$$\begin{aligned} \overline{a + bx} &= a + b\bar{x} \\ \overline{g(x)} &= g(\bar{x}) \end{aligned}$$

Allgemeine Transformation:

Generell ist $\overline{g(x)} \neq g(\bar{x})$

Allgemeine Transformation des Mittelwerts II

Für konvexe Funktionen g gilt:

$$\begin{aligned} g(\bar{x}) &\leq \overline{g(x)} \\ g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &\leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \qquad \text{(Jensen-Ungleichung)} \end{aligned}$$

g konvex: $\Leftrightarrow$ $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$
$\forall \lambda \in [0,1], \quad x, y \in D_g$

Beispiel:
$\bar{x}^2 \leq \overline{x^2}$

Vergleich I

Es gilt allgemein für positive x_i

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

Beweis:

a) Die Funktion $g : t \to log(t)$ ist konkav,
da $g''(t) = -\frac{1}{t^2} < 0$

$$\Rightarrow log(\bar{x}) \geq \overline{log(x)}$$

$$\Rightarrow \bar{x} \geq exp\left(\overline{log(x)}\right) = exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)$$

$$= \left(\prod_{i=1}^{n} exp\left(log(x_i)\right)\right)^{\frac{1}{n}} = \bar{x}_G$$

Vergleich II

b) Die Funktion $g_2 : t \to \frac{1}{exp(t)}$ ist konvex,
da $g_2''(t) = \frac{1}{exp(t)} \geq 0$

Daten $log(x_1), \ldots, log(x_n)$

$$g_2\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\left(exp(log(x_i))^{-1}\right)$$

$$\Rightarrow \frac{1}{\sqrt[n]{\prod_{i=1}^{n} x_i}} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

$$\Rightarrow \underbrace{\sqrt[n]{\prod_{i=1}^{n} x_i}}_{x_G} \geq \underbrace{\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}}_{x_H}$$

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.

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

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

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

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

Transformationsregel

$y_i = a + bx_i$

$$\begin{aligned} \Rightarrow \widetilde{S}_y^2 &= b^2 \widetilde{S}_x^2 \\ \widetilde{S}_y &= |b| \widetilde{S}_x \quad \text{(Analog für } S_x, S_y) \end{aligned}$$

Varianz und Standardabweichung sind stabil
mit linearen Transformationen verträglich.

Verschiebungssatz

Für jedes $c \in \mathbb{R}$ gilt:

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

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

Beachte:
Verschiebungssatz für numerische Berechnung mit Computer nicht geeignet.

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

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

Schichtmittelwerte:

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

Schichtvarianzen:

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

Streuungszerlegung II

Dann gilt:

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

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

Variationskoeffizient

Das Verhältnis von Standardabweichung und Mittelwert ist gegeben durch

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

Der Variationskoeffizient hat keine Einheit und ist skalenunabhängig.

Er ist eine Maßzahl für die relative Schwankung um den Mittelwert.

Mittlere absolute Abweichung (MAD)

Die mittlere absolute Abweichung ist definiert als

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

$$MedAD := median(|x_i - x_{med}|)$$

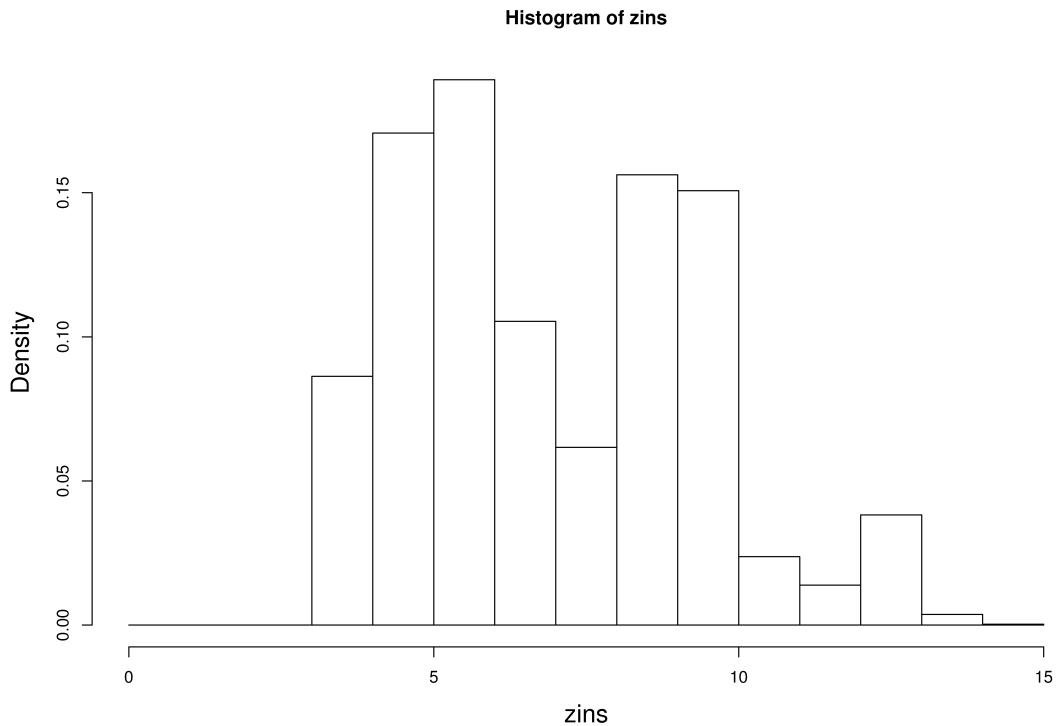
Wegen der Jensen-Ungleichung gilt: $MAD \leq \widetilde{S}$

MAD/ *MedAD* :

- nicht so „schöne“ theoretische Eigenschaften
- klarer interpretierbar als $\widetilde{S}$
- weniger Ausreißer-empfindlich

Uni- und multimodale Verteilungen

unimodal = eingipflig, multimodal = mehrgipflig

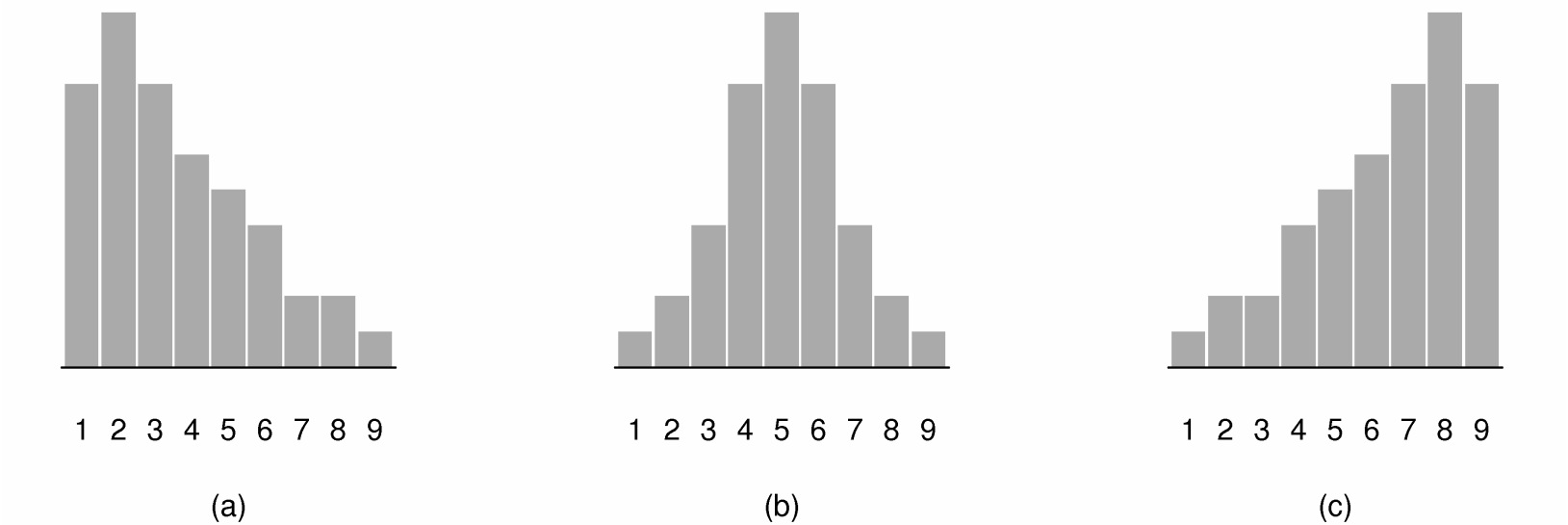


Das Histogramm der Zinssätze zeigt eine bimodale Verteilung.

Symmetrie und Schiefe I

symmetrisch	$\Leftrightarrow$	Rechte und linke Häfte der Verteilung sind annähernd zueinander spiegelbildlich
linkssteil (rechtsschief)	$\Leftrightarrow$	Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab
rechtssteil (linksschief)	$\Leftrightarrow$	Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab

Symmetrie und Schiefe II



Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)

Lageregeln

Symmetrische und unimodale Verteilung:

$\bar{x} \approx x_{med} \approx x_{mod}$

Linkssteile Verteilung:

$\bar{x} > x_{med} > x_{mod}$

Rechtssteile Verteilung:

$\bar{x} < x_{med} < x_{mod}$

Bei gruppierten Daten: Auch für Histogramme gültig

Beachte:
Form der Verteilung bleibt bei linearen Transformationen gleich.
Änderung bei nichtlinearen Transformationen.

Maßzahlen für die Schiefe I

Quantilskoeffizient:

$$g_p = \frac{(x_{1-p} - x_{med}) - (x_{med} - x_p)}{x_{1-p} - x_p}$$

$p = 0.25$ Quartilskoeffizient

Werte des Quantilskoeffizienten:

$g_p = 0$ für symmetrische Verteilungen
$g_p > 0$ für linkssteile Verteilungen
$g_p < 0$ für rechtssteile Verteilungen

Maßzahlen für die Schiefe II

Momentenkoeffizient der Schiefe

$$g_m = \frac{m_3}{\widetilde{s}^3} \quad \text{mit} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Werte des Momentenkoeffizienten

$g_m = 0$ für symmetrische Verteilungen
$g_m > 0$ für linkssteile Verteilungen
$g_m < 0$ für rechtssteile Verteilungen

Dichtefunktion I

Histogramm:

Anteil = Fläche unter der Kurve

Histogramm ist stückweise konstante Funktion

Probleme: Abhängigkeit von der Wahl der Klassengrenzen

Ersetze Histogramm durch glatte Funktion f

Dichtefunktion II

Dichte

Eine positive stetige Funktion heißt Dichte(-funktion), wenn $f(x) \geq 0$ und

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Die Flächen unter der Dichte sollen den approximativen Häufigkeiten entsprechen, d.h.

$$\int_{a}^{b} f(x)dx \approx \frac{1}{n}\#\{x_i | a < x_i \leq b\}$$

Dichte und Verteilung

$$F(x_0) = \frac{1}{n}\#\{x_i | x_i \leq x_0\} \approx \int_{-\infty}^{x_0} f(x)dx$$

$$\hat{F}(x_0) = \int_{-\infty}^{x_0} \hat{f}(x)dx$$

Quantile

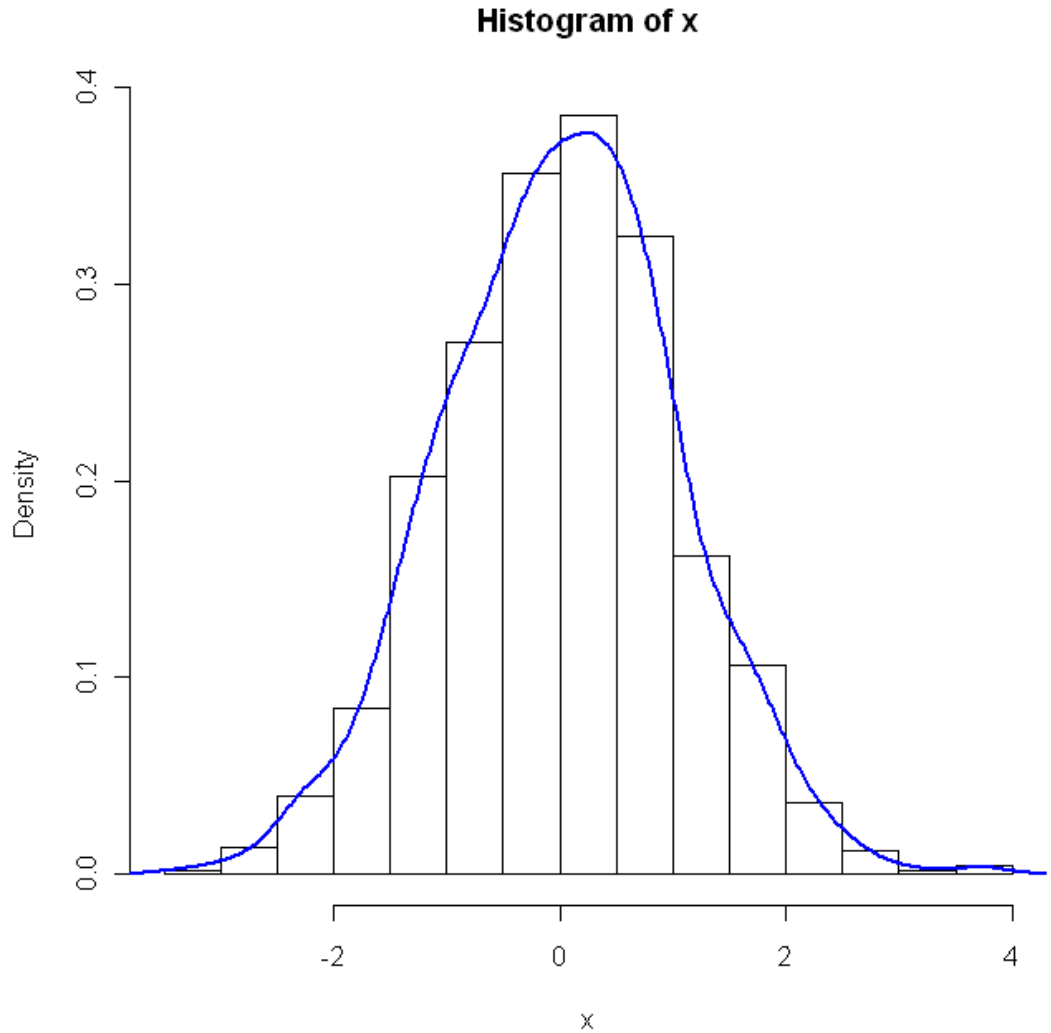
$0 < p < 1$

x_p ist der Wert auf der x-Achse, für den gilt:

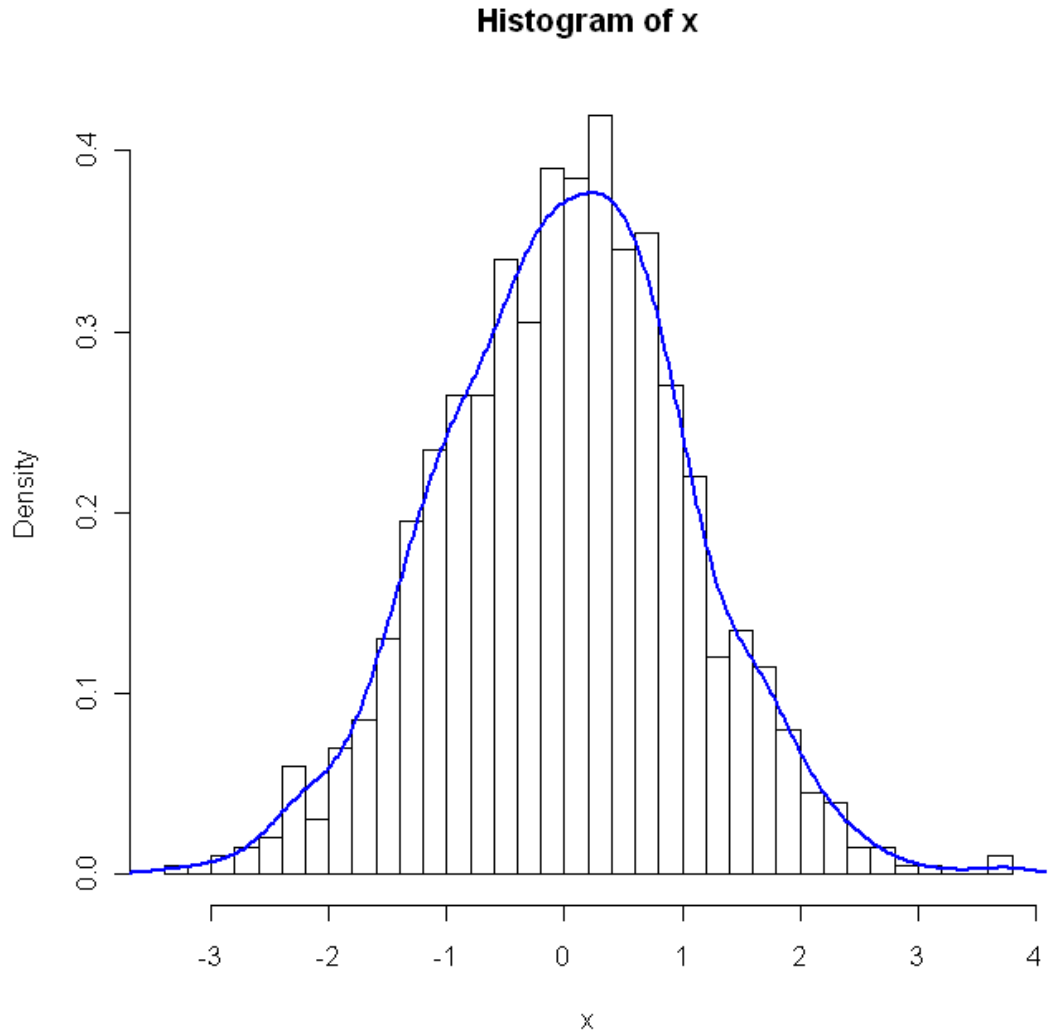
$$\int_{-\infty}^{x_p} f(x)dx = p$$

Der Median teilt die Fläche in 2 gleich große Teile.

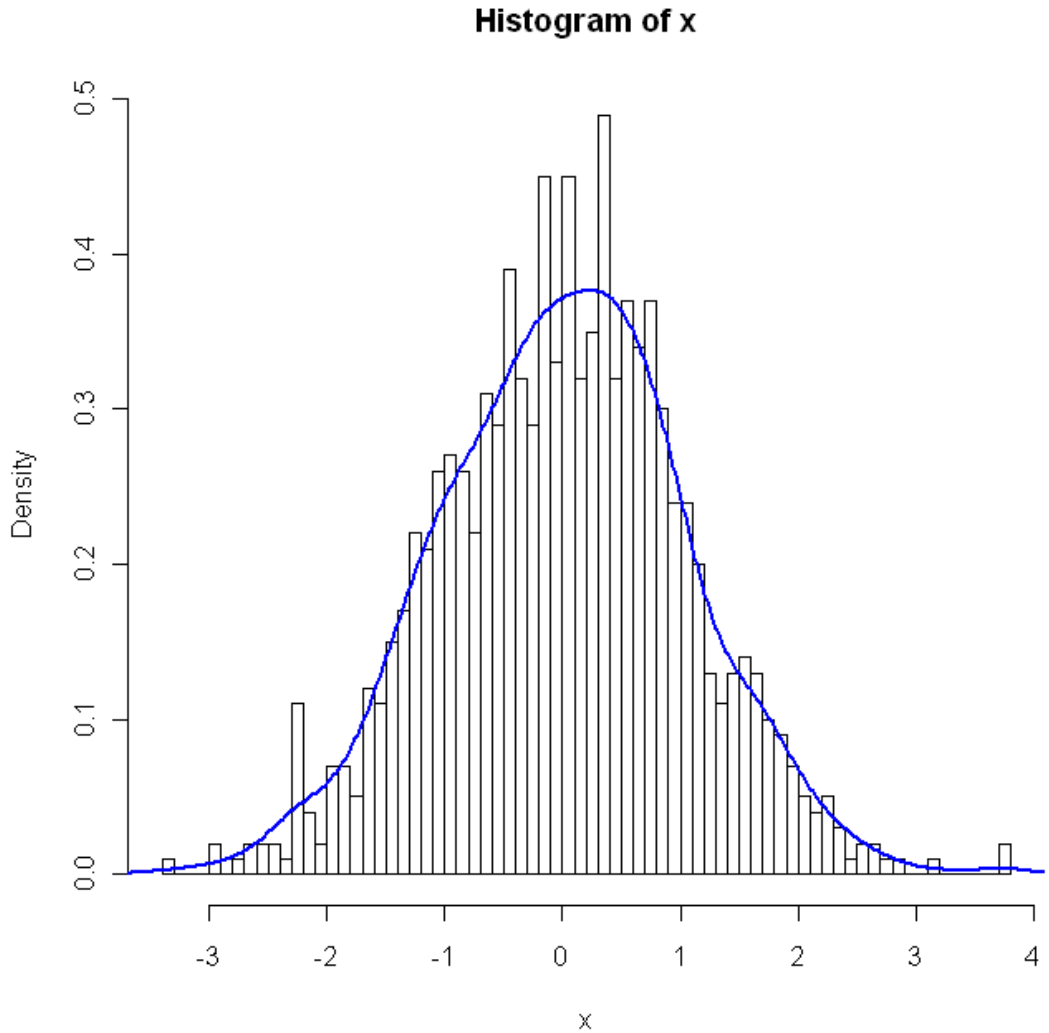
Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Berechnung von Dichte-Kurven

$$\hat{f}(x) = \frac{\frac{1}{n}\#\{x_i | x_i \in [x-h, x+h)\}}{2h}$$

$\Rightarrow$ „Gleitendes Histogramm“

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

mit $K(u) = \begin{cases} \frac{1}{2} & \text{für } -1 \leq u < 1 \\ 0 & \text{sonst} \end{cases}$

K : Kernfunktion

Kern-Dichteschätzer

$K(u)$ sei Kernfunktion, d.h.

$K(u) \geq 0$ und $\int_{-\infty}^{\infty} K(u)du = 1$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

heißt Kern-Dichteschätzer

Kerne:

Epanechnikov-Kern	$K(u) = \frac{3}{4}(1 - u^2)$	für $-1 \leq u < 1$, 0 sonst.
Bisquare-Kern	$K(u) = \frac{15}{16}(1 - u^2)^2$	für $-1 \leq u < 1$, 0 sonst.
Gauß-Kern	$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right)$	für $u \in \mathbb{R}$

Bemerkungen zur Dichteschätzung

- Abhängigkeit von der Bandweite $h \rightarrow$ Verfahren zur Bestimmung von h aus den Daten
- Abhängigkeit von der Wahl des Kerns eher unbedeutend
- Kerndichteschätzungen sind insbesondere bei größeren Datenmengen Histogrammen vorzuziehen

Normalverteilung

Für $x \in \mathbb{R}$ heißt

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Normalverteilungsdichte mit Mittelwert μ und Standardabweichung σ

Für $\mu = 0$ und $\sigma = 1$ erhät man

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}x^2\right)$$

Quantile I

Quantile der Standardnormalverteilung:

p	50%	75%	90%	95%	97.5%	99%
z_p	0.0 (Median)	0.67	1.28	1.64	1.96	2.33

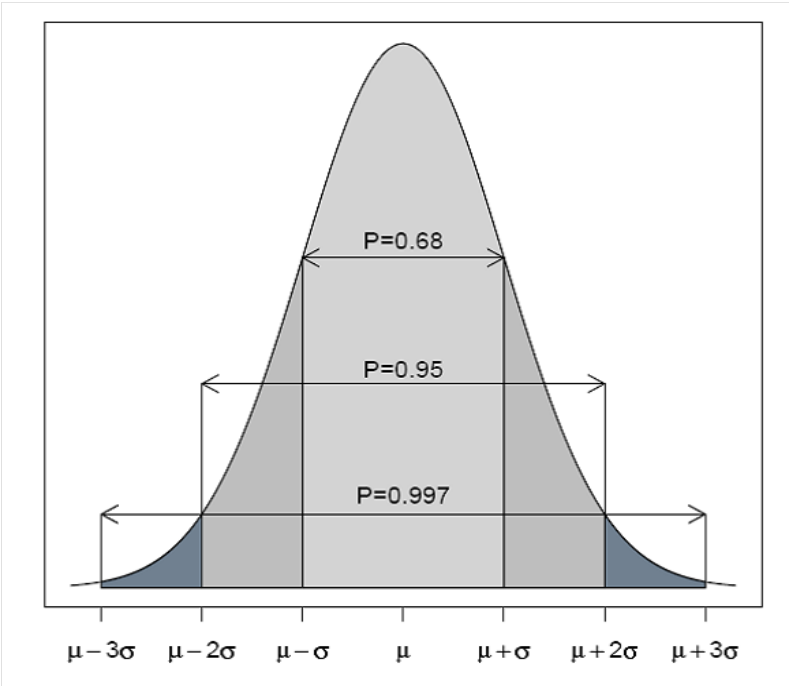
Quantile II

68-95-99.7-Prozent-Regel:

68% der Beob. liegen im Interv. $\mu \pm \sigma$

95% der Beob. liegen im Interv. $\mu \pm 2\sigma$

99.7% der Beob. liegen im Interv. $\mu \pm 3\sigma$

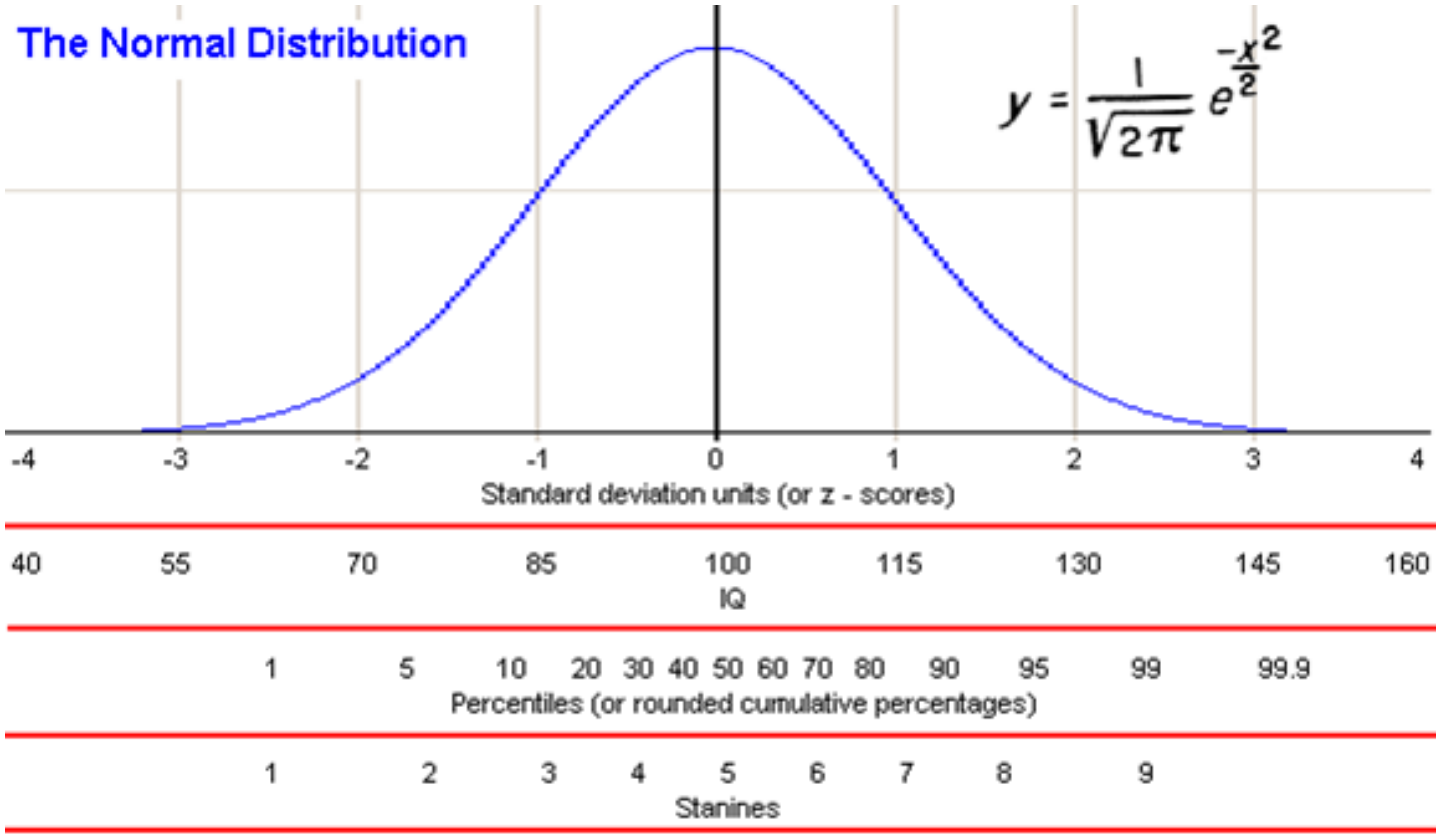


Normalverteilung in der Psychologie

Skalen werden so konstruiert, dass die Verteilung in der Population einer Normalverteilung genügt.

IQ :	Mittelwert = 100, Standardabweichung 15
T- Werte:	Mittelwert 50, Standardabweichung 10
Sta(ndard)nine	Mittelwert 5 Standardabweichung 2

Normalverteilung in der Psychologie II



Strategie zur Skalenbildung

- Ziehe grosse Stichprobe aus der Population
- Ordne Ergebnisse
- Zuordnung von Stanine

Schema

4%	7%	12%	17%	20%	17%	12%	7%	4%
1	2	3	4	5	6	7	8	9

Strategie zur Skalenbildung II

- Ziehe grosse Stichprobe aus der Population
- Standardisierung (Z- Werte)

$$z_i = (x_i - \bar{x})/s_x$$

- Reskalieren durch Multiplikation mit gewünschter Standardabweichung und Addition des gewünschten Mittelwertes

Normalverteilung in der technischen Statistik

- Größen in der Produktion (Längen etc.)
- Messfehler
- Größen nach geeigneter Transformation
- 6 Sigma

Überprüfung der Annahme der Normalverteilung

Fragestellung: Passen die Daten zu einer Normalverteilung?

a) Vergleiche Histogramm, Dichteschätzer mit NV-Dichte ($\mu = \bar{x}, \sigma = S$)

Überprüfung der Annahme der Normalverteilung

Fragestellung: Passen die Daten zu einer Normalverteilung?

a) Vergleiche Histogramm, Dichteschätzer mit NV-Dichte ($\mu = \bar{x}, \sigma = S$)

b) Vergleiche Verteilungsfunktion, d.h. empirische Verteilungsfunktion F mit $\Phi(t) = \int_{-\infty}^{t} f(\mu, \sigma, t)dt$

Überprüfung der Annahme der Normalverteilung

Fragestellung: Passen die Daten zu einer Normalverteilung?

a) Vergleiche Histogramm, Dichteschätzer mit NV-Dichte ($\mu = \bar{x}, \sigma = S$)

b) Vergleiche Verteilungsfunktion, d.h. empirische Verteilungsfunktion F mit $\Phi(t) = \int_{-\infty}^{t} f(\mu, \sigma, t)dt$

c) Schiefe = 0 ?

d) Q-Q-Plot

Normal-Quantil-Plot

Sei $x_{(1)}, \ldots, x_{(n)}$ die geordneten Daten. Für $i = 1, \ldots, n$ werden die $(i - 0.5)/n$-Quantile $z_{(i)}$ der Standardnormalverteilung berechnet.

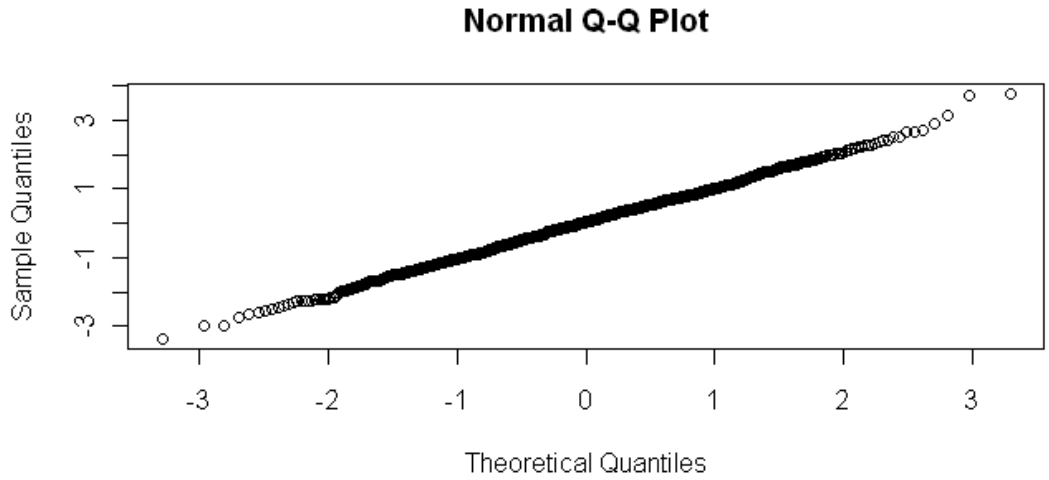
Der Normal-Quantil-Plot (Normal-Q-Q-Plot) besteht aus den Punkten

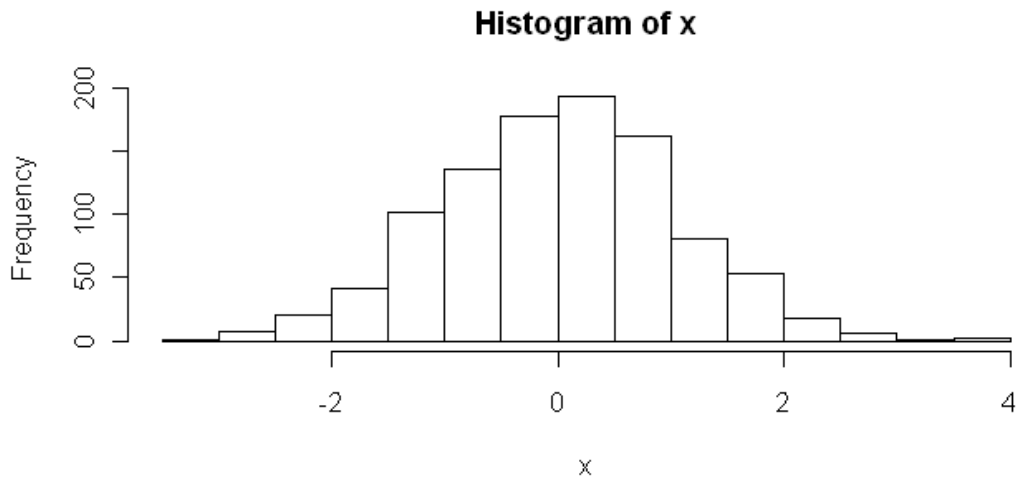
$$(z_{(1)}, x_{(1)}), \ldots, (z_{(n)}, x_{(n)})$$

im $z - x-$Koordinatensystem.

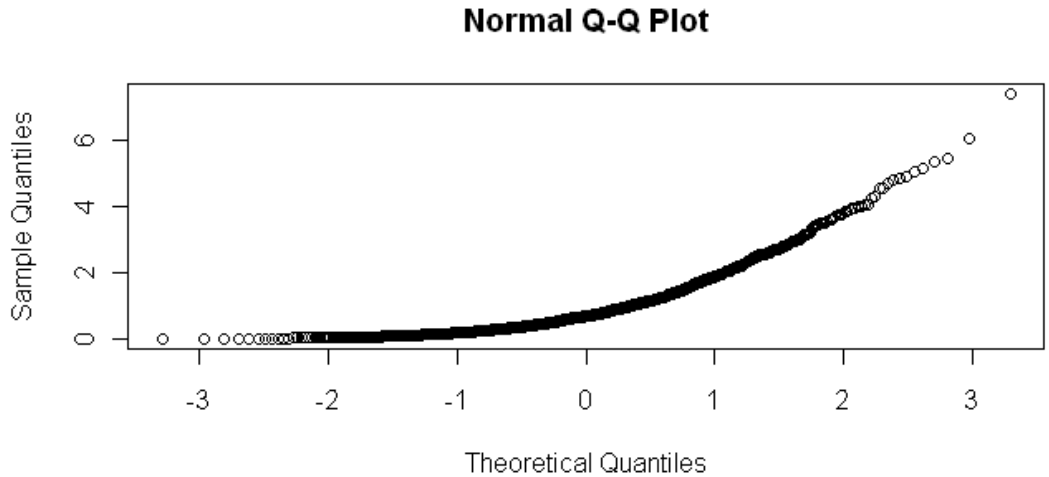
Liegen die Punkte auf einer Geraden, so passt die Normalverteilung gut zu den Daten

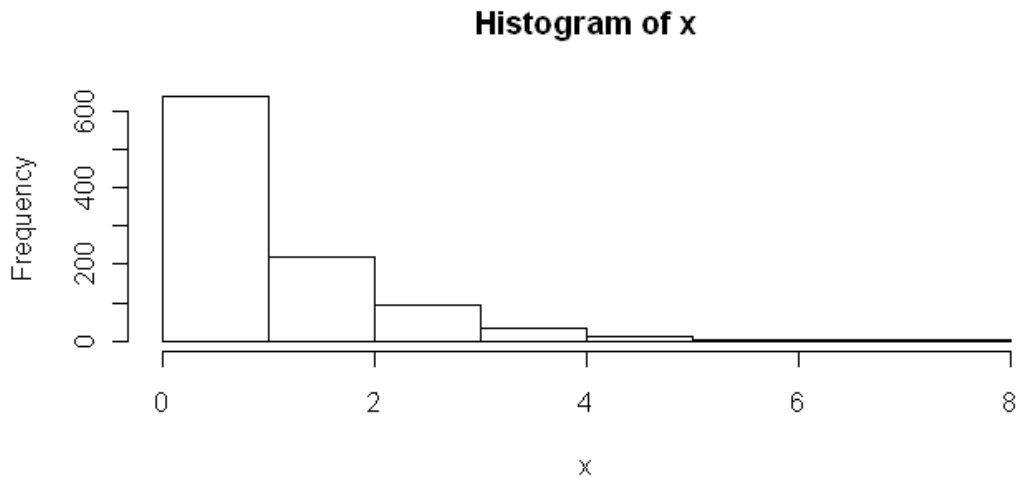
Beispiele I


Normal Q-Q Plot
Sample Quantiles
-3
-1
1
3
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram of x
Frequency
0
50
100
200
-2
0
2
4
x

Beispiele II





Konzentrationsmaße

Motivation

Existiert eine Menge, die auf viele Individuen verteilt ist, kann es hilfreich sein zu wissen, wie diese Menge verteilt ist.

Beispiele

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

Lorenzkurve

Grundidee

Es sollen folgende Aussagen grafisch dargestellt werden:

- Die „Ärmsten“ x% besitzen einen Anteil von y%.
- Die „Reichsten“ x% besitzen einen Anteil von y%.

Lorenzkurve

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 x-Achse wird in *gleiche Längen* aufgeteilt, deren Anzahl der der Individuen (Merkmalsausprägungen) entspricht:

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

- Die y- Werte werden wie folgt berechnet:

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

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

Lorenzkurve

Beispiel

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

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

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

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

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

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

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

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

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

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

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

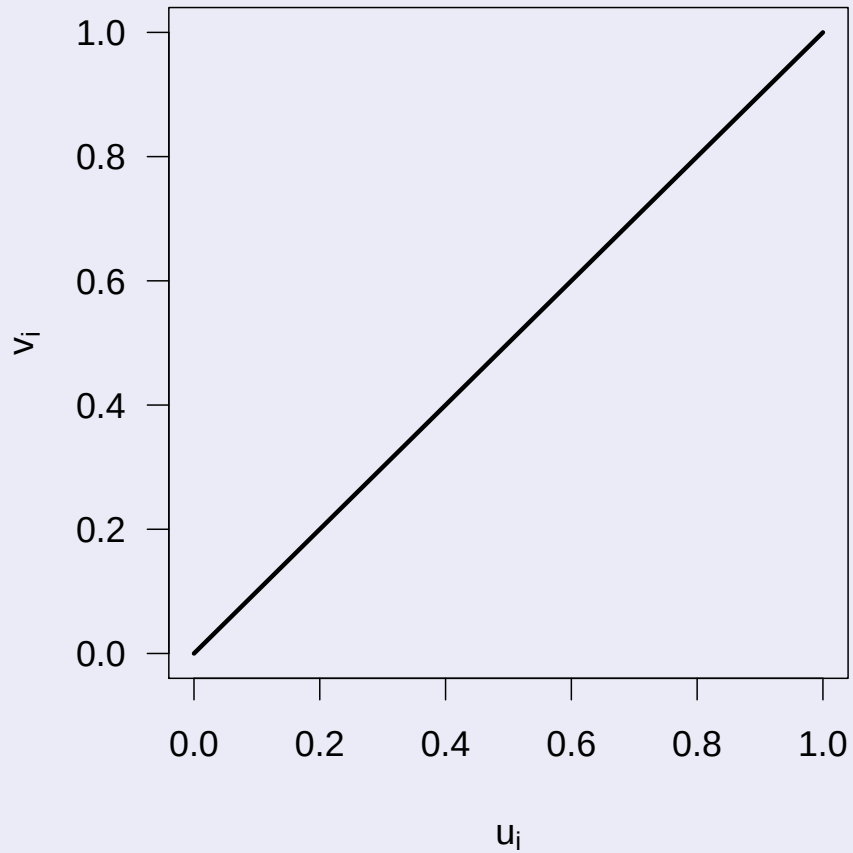
i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	0,2	0,2
2	20	0,4	0,4
3	20	0,6	0,6
4	20	0,8	0,8
5	20	1	1

Lorenzkurve

Beispiel

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.

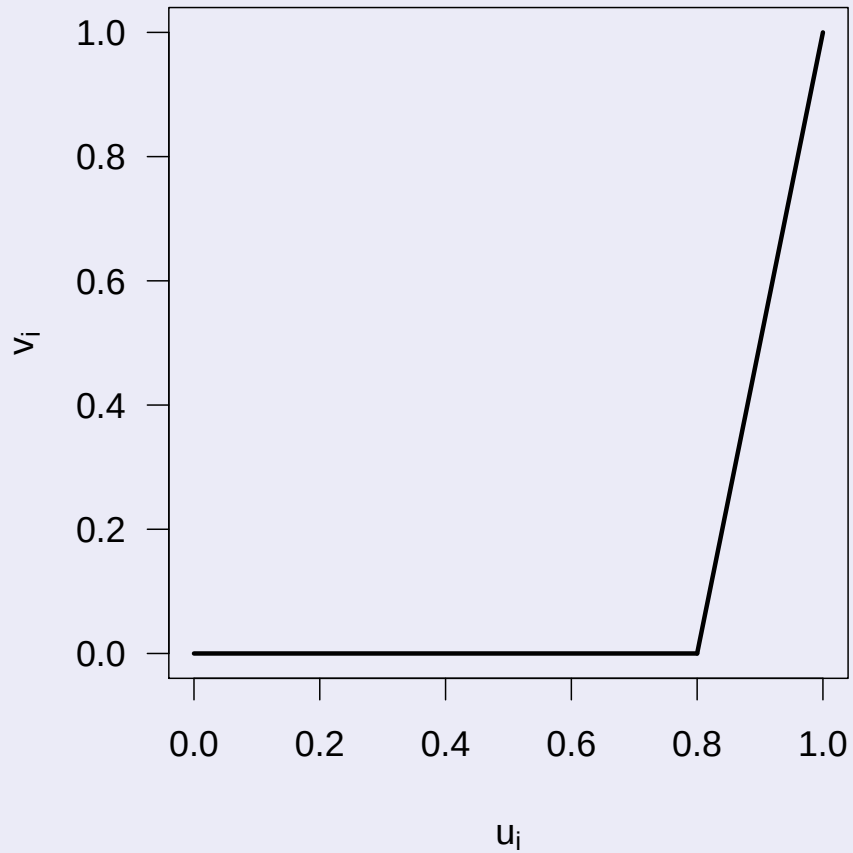
i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	0,2	0
2	0	0,4	0
3	0	0,6	0
4	0	0,8	0
5	100	1	1

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

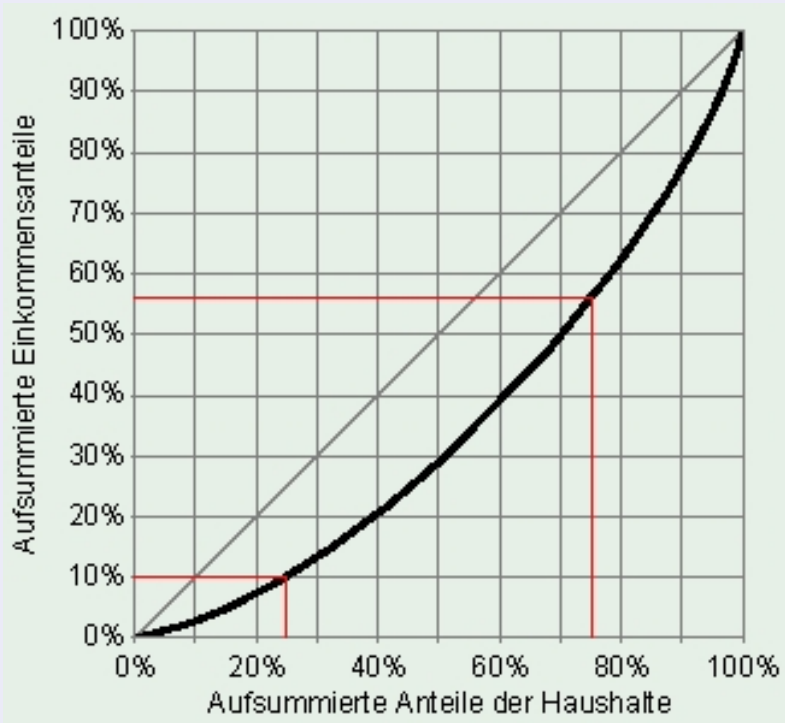
Erscheinungsbild von Lorenzkurven

- 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

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.

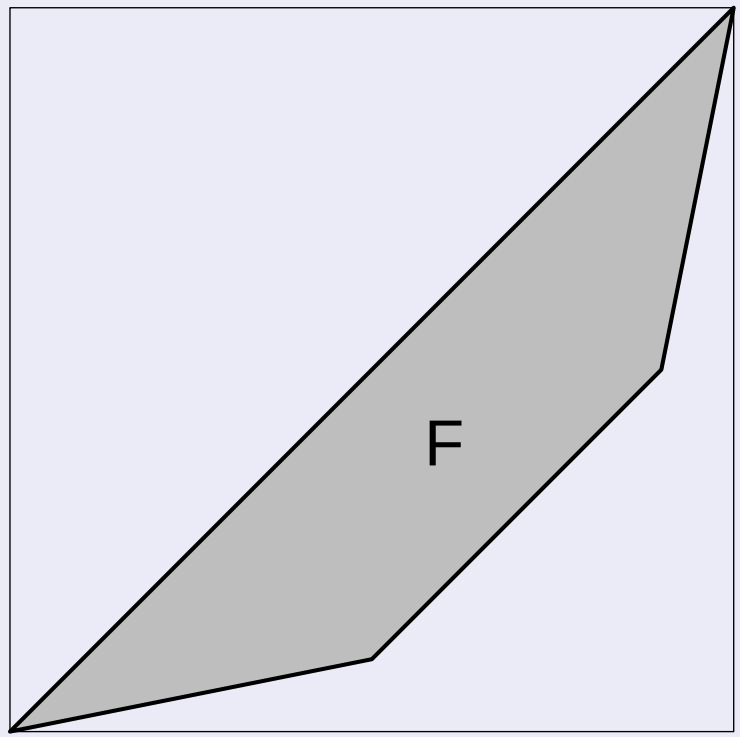
Gini-Koeffizient

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

Normierter 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.

Herfindahl-Index

Herfindahl-Index

$x_1, \ldots x_n$ seien die Daten mit $x_i \geq 0$.
Die Anteile der Einheiten i sind wie folgt definiert:

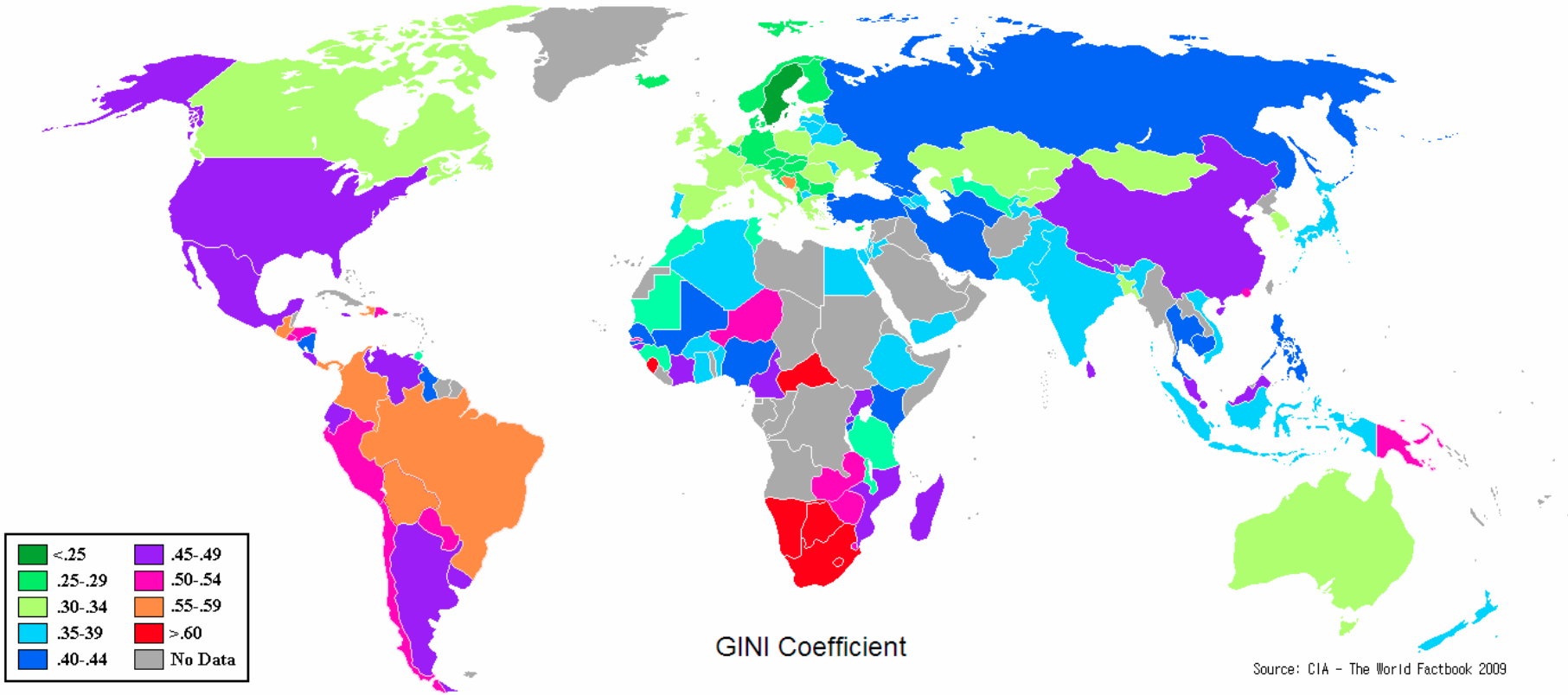
$$p_i := \frac{x_i}{\sum_{j=1}^{n} x_j}$$

Der Herfindahl-Index ist

$$H := \sum_{i=1}^{n} p_i^2$$

Der Wertebereich ist von $1/n$ (Identische x) bis 1 (Monopol)

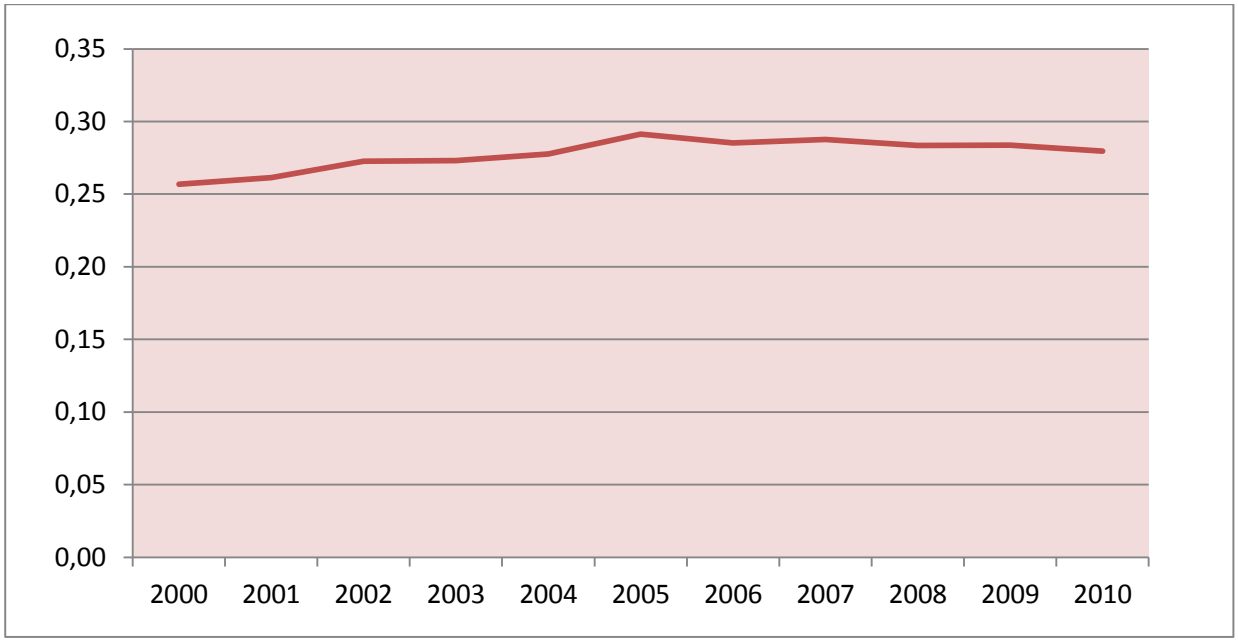
GINI Einkommen nach CIA report 2009



Bericht der Bundesregierung

- X -

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.

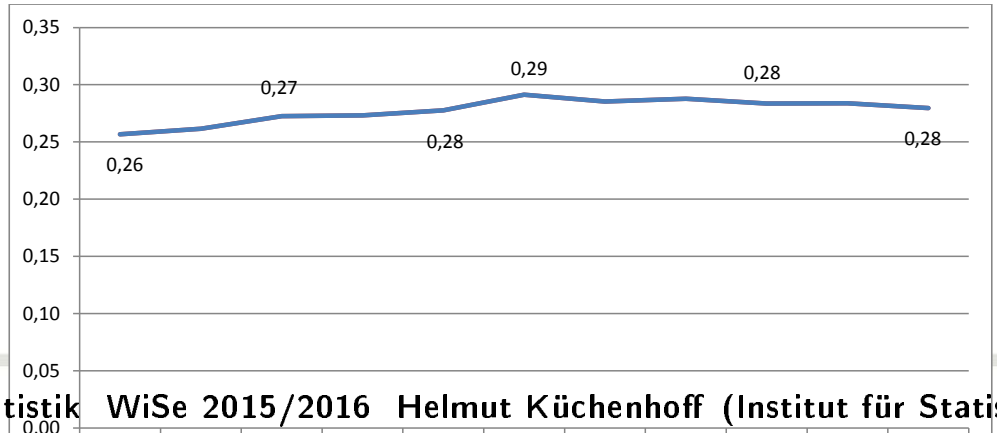
Tabelle C I.1.2:
Verteilung der Nettoäquivalenzeinkommen 2003 und 2008

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)



Einführung

In den meisten Anwendungen und in den Beispielen werden an jeder Einheit *gleichzeitig* mehrere Merkmale $X, Y, Z, \ldots$ erhoben:

$\Rightarrow$ mehrdimensionale oder multivariate Daten

Grundgesamtheit

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

Einheit i

Daten

- Daten $(x_1, y_1, z_1), \ldots, (x_i, y_i, z_i), \ldots, (x_n, y_n, z_n)$

  Im weiteren: Meistens nur *zwei* Merkmale X, Y
- Fragestellungen:

  $X \leftrightarrow Y$ Wie hängen X und Y zusammen?
  Assoziation, Korrelation

  $X \rightarrow Y$ Wie beeinflusst X das (Ziel-)Merkmal Y ?
  Regression

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.

Kontingenztabellen

Sonntagsfrage: "Welche Partei würden Sie wählen, wenn am nächsten Sonntag Bundestagswahlen wären?" werden üblicherweise in Prozenten (%) wiedergegeben. Für den Befragungszeitraum ergab sich folgende Tabelle:

	CDU/CSU	SPD	FDP	Grüne	Rest	
Männer	33	35	4	6	22	100
Frauen	40	29	6	10	15	100
insgesamt	37	32	5	8	18	100

Aus den ersten beiden Zeilen ergibt sich, dass die Parteipräferenzen für Männer und Frauen unterschiedlich sind.

Sonntagsfrage

In der angegebenen Tabelle sind die ursprünglichen Daten bereits in Prozenten für die geschlechtsspezifischen Populationen angegeben. Die Rückrechnung auf 435 Männer und 496 Frauen ergibt:

	CDU/CSU	SPD	FDP	Grüne	Rest	
Männer	144	153	17	26	95	435
Frauen	200	145	30	50	71	496
	344	298	47	76	166	931

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

Noation

$f_{ij} = h_{ij}/n$ 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

Beispiel: Sonntagsfrage Prozentzahlen für Parteipräferenz in den

Schichten (Subpopulationen) „weibliche Wähler“ und „männliche Wähler“ $\stackrel{\wedge}{=}$ bedingte relative Häufigkeiten für Parteipräferenzen gegeben das Geschlecht.

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: Sonntagsfrage

- Zeile $X = a_1 =$ Männer

  Bedingte Häufigkeiten für Männer $(X = a_1)$:

  1.Zeile / Randhäufigkeit für Männer

  $$\frac{h(a_1,b_1)}{h(a_1)} = f(b_1|a_1), \ldots, \frac{h(a_j,b_j)}{h(a_1)} = f(b_j|a_1), \ldots$$

  $\frac{144}{435} \approx 33\%$, $\frac{153}{435} \approx 35\%$ usw.

- Zeile $X = a_2 =$ Frauen analog,

  z.B. $\frac{200}{496} \approx 40\%$ usw.

Bedingte und gemeinsame Häufigkeiten

Man kann auch umgekehrt aus bedingten Häufigkeiten und Randhäufigkeiten die gemeinsamen Häufigkeiten ausrechnen. Bei der Sonntagsfrage ist die Tabelle der bedingten Häufigkeiten gegeben und dazu die Randhäufigkeiten

$$h(a_1) = 435 \text{ Männer}, \quad h(a_2) = 496 \text{ Frauen}; \quad n = 931.$$

$$h(a_1) \cdot f(b_1|a_1) = h(a_1, b_1)$$

$$\Rightarrow$$

$$435 \cdot 33\% \approx 144 \quad \text{usw.}$$

So wurde die Tabelle der gemeinsamen Häufigkeiten h_{ij} rekonstruiert.

Beispiel: Arbeitslosigkeit

$f(\quad|a_i), \quad X = a_i, \;\; i = 1, \ldots, 4$ 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.

Zusammenhangsanalyse in Kontingenztabellen

Bisher: Tabellarische / graphische 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

	ApoE3	ApoE4	Summe
Kontrolle	2258	803	3061
Fall	593	620	1213
	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

Verallgemeinerung auf $k \times m$ Kontingenztafeln, Anmerkungen

- Verallgemeinerung des Verfahres auf mehr als zwei Ausprägungen mindestens eines Merkmals: Man beschränkt sich auf jeweils zwei Zeilen $X = a_i$ und $X = a_j$ und zwei Spalten $Y = b_r$ und $Y = b_s$ und die zugehörigen vier Zellen einer $(k \times m)$-Kontingenztafel.
- Verwendung einer Referenzkategorie
- Statt Odds Ratio wird oft der logarithmierte Odds Ratio verwendet

Anwendung: Apolipoprotein E und Morbus Alzheimer

Etablierter Zusammenhang zwischen Apolipoprotein Eε4 und Morbus Alzheimer

Daten aus Metaanalyse

ApoE genotype	$\varepsilon2\varepsilon2$	$\varepsilon2\varepsilon3$	$\varepsilon2\varepsilon4$	$\varepsilon3\varepsilon3$	$\varepsilon3\varepsilon4$	$\varepsilon4\varepsilon4$
Clinical controls	27	425	81	2258	803	71
Clinical Alzheimer	7	74	41	593	620	207
PM controls	3	75	18	358	120	8
PM Alzheimer	1	20	17	249	373	97

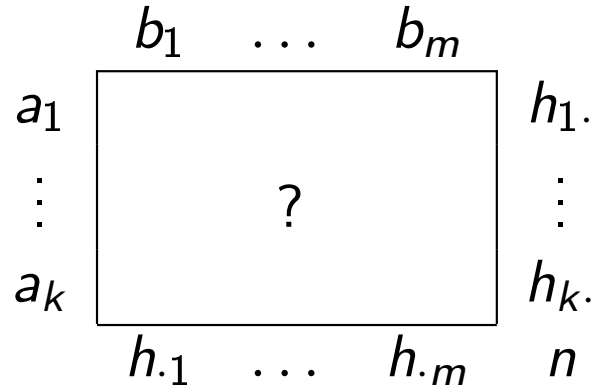
Anwendung: Apolipoprotein E und Morbus Alzheimer

OR im Vergleich zu $\varepsilon 3\varepsilon 3$ (Referenz)

	$\varepsilon 2\varepsilon 2$	$\varepsilon 2\varepsilon 3$	$\varepsilon 2\varepsilon 4$	$\varepsilon 3\varepsilon 3$	$\varepsilon 3\varepsilon 4$	$\varepsilon 4\varepsilon 4$
OR (klinisch)	1	0.7	2.94	1	2.94	11.1
OR (post mortem)	0.5	0.4	1.4	1	4.5	17.4

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

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

Bemerkung: Lokale Odds Ratios sind alle 1

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

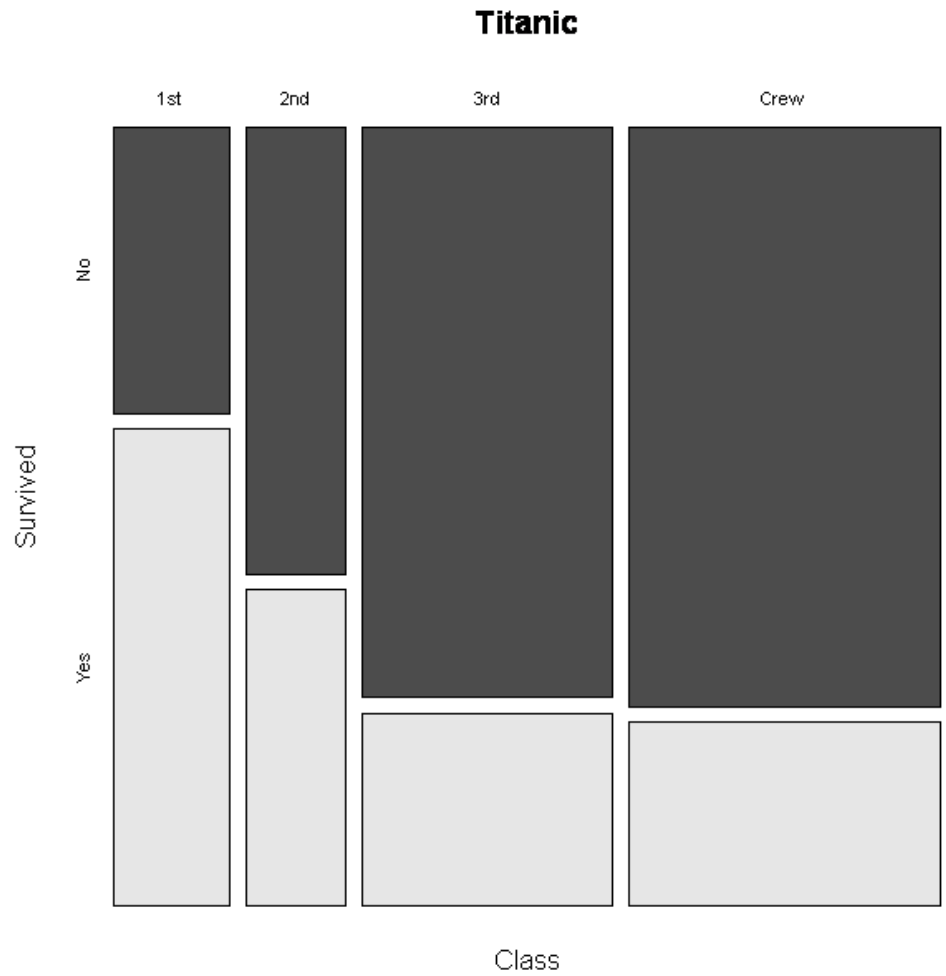
Beispiel: Überleben beim Titanic-Untergang

- Mehrere diskrete Merkmale: Geschlecht, Klasse, Kind/Erwachsene, Überleben (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

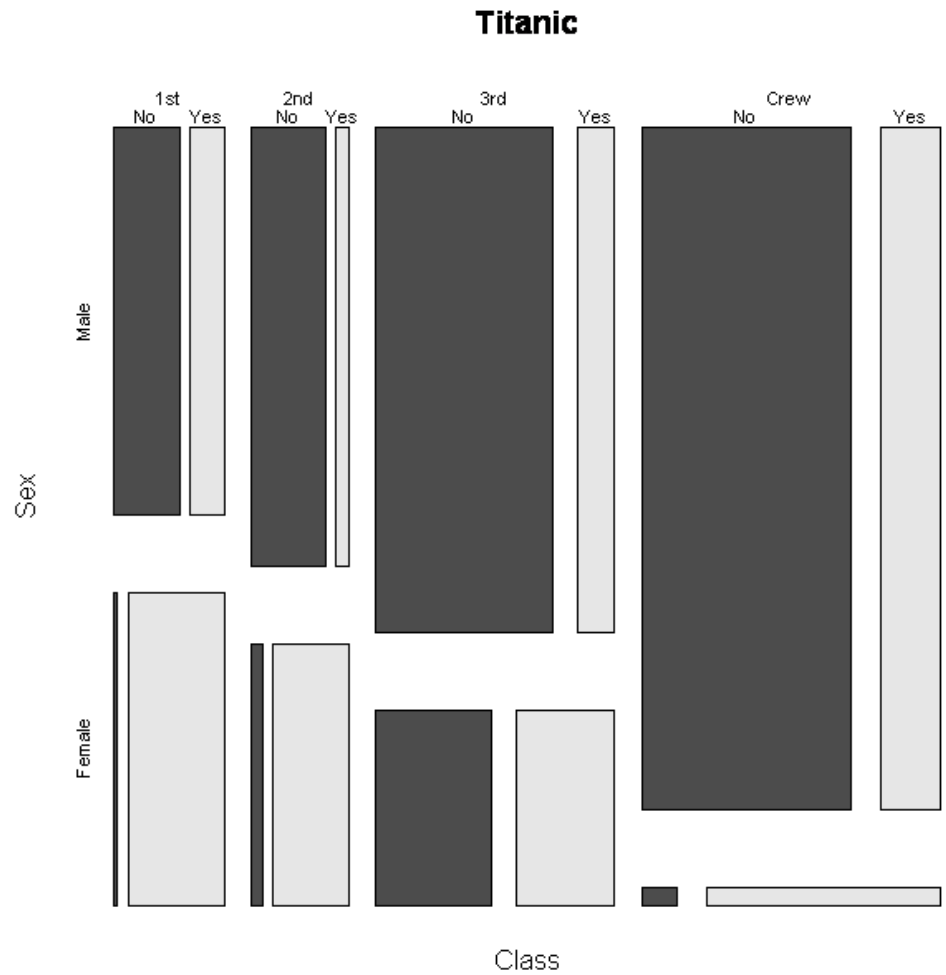
Mosaik-Plot

- 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: Überlebende bei Titanic



Beispiel: Überlebende bei Titanic



Maße zur Übereinstimmung

Problemstellung

- Stimmen zwei oder mehrere Beobachter in ihrer Einschätzung überein? (im engl.: *rater agreement* oder *interrater agreement*)
- Beispiel: Zwei Professoren beurteilen die Referate oder Seminararbeiten von Studenten. Stimmen Sie in ihrer Bewertung (durch Noten) überein?
- Beispiel: Stimmen zwei oder mehrere Ärzte in ihrer Diagnose überein?
- Wahre Diagnose ("gold standard") unbekannt

Medizinisches Beispiel

Eine Klinik macht computertomografische Aufnahmen (CT–Bilder) oder Röntgenaufnahmen von sogenannten Kalkschultern, also Schultern, die Kalkablagerungen aufweisen. Mehrere Ärzte sollen anhand von 56 Patienten beurteilen, um welchen Typ es sich bei diesen Ablagerungen handelt (sog. Gärtner–Skala, ordinal):

- Typ 1: Aufbauphase (kann Monate oder Jahre dauern). Der Patient hat chronische Beschwerden, Schmerztherapie und Krankengymnastik bringen keine Linderung, eine Operation muss erwogen werden.
- Typ 2: Beginnende Auflösungsphase.
- Typ 3: Auflösung des Kalkdepots (kann Wochen oder Monate dauern). Behandlung meist konservativ durch Krankengymnastik und Schmerztherapie.

Medizinisches Beispiel (Fortsetzung)

- Die Beobachtungen lassen sich für zwei Ärzte in einer quadratischen Kontingenztafel veranschaulichen
- Die folgenden Kontingenztafeln geben die Ergebnisse für jeweils zwei verschiedene Ärzte bei den gleichen 56 Patienten wieder

Medizinisches Beispiel (Fortsetzung)

Tabelle: 3×3–Tafel für die Einschätzung von Arzt A und Arzt B

	Arzt A			
Arzt B	1	2	3	$\sum$
1	8	20	3	31
2	2	15	1	18
3	0	5	2	7
$\sum$	10	40	6	56

Medizinisches Beispiel (Fortsetzung)

Tabelle: 3×3–Tafel für die Einschätzung von Arzt C und Arzt D

	Arzt C			
Arzt D	1	2	3	$\sum$
1	27	6	3	36
2	2	6	3	11
3	0	4	5	9
$\sum$	29	16	11	56

Medizinisches Beispiel (Fortsetzung)

- Vollständige Übereinstimmung in der Einschätzung des Patienten liegt vor, wenn beide Ärzte den gleichen Typ (1, 2 oder 3) zuordnen
- *Bemerkung:* Gleiche Einstufung bedeutet nicht unbedingt, dass diese auch *richtig* im Sinne einer validen Einstufung ist. Beide können sich irren!

Üblichen Maßzahlen?

- Üblich bei Kontingenztafeln: χ^2, Kontingenzkoeffizient, etc.
- Man geht von vornherein davon aus, dass ein Zusammenhang bezüglich der Bewertung der Beobachter besteht
- Beobachter führen Bewertung *unabhängig voneinander durch*, d.h. kein Beobachter kennt die Bewertung des oder der anderen Beobachter, die Beurteilungen hängen aber zusammen, da sie jeweils am gleichen Subjekt (hier: Patient) durchgeführt werden

Medizinisches Beispiel (Fortsetzung)

- Für die Tabelle der Ärzte A und B erhält man $(8 + 15 + 2) = 25$ vollständige Übereinstimmungen
- Für die Tabelle der Ärzte C und D erhält man $(27 + 6 + 5) = 38$ vollständige Übereinstimmungen
- Kann man daraus sofort schließen, dass Übereinstimmung von C und D größer ist als von A und B? Im Prinzip ja, allerdings müssen wir beachten, dass ein gewisser *Teil der Übereinstimmung zufällig sein kann*

Maßzahl: Kappa–Koeffizient

- Kappa–Koeffizient nach Cohen (1960) dient zur Messung der Übereinstimmung in quadratischen $I \times I$–Kontingenztafeln
- Betrachtet wird primär die Hauptdiagonale der Kontingenztafel (vollständige Übereinstimmung)

Allgemeine Darstellung

Tabelle: Schema einer $I \times I$-Kontingenztafel. Die fettgedruckten Häufigkeiten liegen auf der Diagonalen und werden zur Berechnung von Kappa verwendet.

		Beobachter 1					
		1		i		I	$\sum$
	1	$\mathbf{n_{11}}$	$\cdots$	n_{1i}	$\cdots$	n_{1I}	n_{1+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Beobachter 2	i	n_{i1}	$\cdots$	$\mathbf{n_{ii}}$	$\cdots$	n_{iI}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	$\cdots$	n_{Ii}	$\cdots$	$\mathbf{n_{II}}$	n_{I+}
	$\sum$	n_{+1}	$\cdots$	n_{+i}	$\cdots$	n_{+I}	n

Allgemeine Darstellung

- Kappa–Koeffizient berücksichtigt darüber hinaus die *zufällige Übereinstimmung* die man auch bekommen würde, wenn die Einschätzungen der Beobachter keinen Zusammenhang aufweisen würden
- Wir berechnen daher zwei Größen

Berechnung von Kappa

- Relativer Anteil der Übereinstimmung beider Beobachter:

$$f_o = \sum_{i=1}^{I} f_{ii} = \sum_{i=1}^{I} \frac{n_{ii}}{n} = \frac{\sum_{i=1}^{I} n_{ii}}{n} \tag{3.1}$$

- Zufällige Übereinstimmung, wenn kein Zusammenhang besteht: dies ist äquivalent zur Bestimmung der sogenannten erwarteten relativen Häufigkeiten unter Unabhängigkeit (wie bei χ^2):

$$f_e = \sum_{i=1}^{I} f_{i+} f_{+i} = \sum_{i=1}^{I} \frac{n_{i+}}{n} \frac{n_{+i}}{n} = \sum_{i=1}^{I} \frac{n_{i+} n_{+i}}{n^2} = \frac{\sum_{i=1}^{I} n_{i+} n_{+i}}{n^2} \tag{3.2}$$

Berechnung von Kappa (Fortsetzung)

Der Kappa–Koeffizient ist definiert durch

$$\kappa = \frac{f_o - f_e}{1 - f_e} \ , \tag{3.3}$$

Interpretation von Kappa

- Der Zähler ist die Differenz aus der beobachteten Übereinstimmung und der unter Zufälligkeit zu erwartenden Übereinstimmung. Dies ist damit ein Maß für die über die Zufälligkeit hinausgehende Übereinstimmung der Beobachter (*engl.: chance corrected agreement*)
- Die Eins im Nenner stellt die maximal mögliche relative Häufigkeit für Übereinstimmung dar, nämlich wenn alle Beobachtungen auf der Diagonalen der Kontingenztafel liegen und sämtliche Nebendiagonalen nur Nullen enthalten

Interpretation von Kappa (Fortsetzung)

- Der Kappa–Koeffizient ist damit ebenfalls Eins, wenn alle Beobachtungen auf der Diagonalen der Kontingenztafel liegen und sämtliche Nebendiagonalen nur Nullen enthalten. Er kann auch negativ werden, wenn zum Beispiel keine Übereinstimmung da ist (im Extremfall: Nullen auf der Diagonalen)
- Der Kappa–Koeffizient ist Null, wenn für alle $i = 1, \ldots, I$ gilt: $f_{ii} = f_{i+} f_{+i}$, das heißt, wenn exakte Unabhängigkeit in der beobachteten Tafel vorliegt

Medizinisches Beispiel (Fortsetzung), Arzt A und B

$$f_o = \frac{8 + 15 + 2}{56} = \frac{25}{56}$$

und

$$\begin{aligned} f_{11} &= \frac{31 \cdot 10}{56^2} \\ f_{22} &= \frac{18 \cdot 40}{56^2} \\ f_{33} &= \frac{7 \cdot 6}{56^2} \\ f_e &= \frac{31 \cdot 10 + 18 \cdot 40 + 7 \cdot 6}{56^2} \\ &= \frac{1072}{56^2} \end{aligned}$$

Damit erhalten wir

$$\kappa = \frac{\frac{25}{56} - \frac{1072}{56^2}}{1 - \frac{1072}{56^2}} = 0.159$$

Medizinisches Beispiel (Ergebnis)

- Entsprechend: Arzt D und C: $\kappa = 0.445$
- Die Einschätzung von A und B ist „schwach übereinstimmend“
- Die Einschätzung von C und D ist „mäßig übereinstimmend“

Veranschaulichung

Zwei mögliche Ergebnisse für die Einschätzung zweier Beobachter für 20 Objekte bezüglich eines dichotomen Merkmals:

Tabelle: Problematik von Kappa

	1	0
1	10	1
0	0	9

	1	0
1	18	1
0	0	1

In beiden Fällen erhält man eine Übereinstimmung in 19 von 20 Objekten, oder einen Wert von $f_o = 0.95$. Allerdings ist Kappa für die linke Tafel 0.9 und für die rechte Tafel nur 0.64

Erklärung

- Die Randverteilungen der beiden Beobachter unterscheiden sich dabei jeweils nur gering, d.h. sie scheinen gut kalibriert zu sein (daran liegts also nicht)
- Offenbar ist die Prävalenz (d.h. der Grundanteil in der untersuchten Population/Stichprobe) in der rechten Tafel für die Ausprägung „1“ wesentlicher größer ist als für die Ausprägung „0“. Damit ist aber auch die zufällige Übereinstimmung wahrscheinlicher! Genau dieser Effekt wird bei Kappa berücksichtigt und herausgerechnet
- Befürworter von Kappa sehen diesen Effekt als wünschenswert an!

Kritik an Kappa

Zwei Aspekte werden vermischt:

- Die Beobachter können einen Bias aufweisen, das heisst, die Nicht–Übereinstimmung beruht darauf, dass zum Beispiel Lehrer 1 generell bessere Noten vergibt als Lehrer 2. Man sagt dann auch, dass die Beobachter nicht *kalibriert* sind
- Die Beobachter schätzen die Subjekte verschieden ein. Die Nicht–Übereinstimmung beruht darauf, dass Beobachter 1 zum Beispiel Subjekt 1 höher einstuft als Subjekt 2, Beobachter 2 dagegen Subjekt 2 höher als Subjekt 1. Beispiel: Lehrer 1 gibt Schüler 1 eine bessere Note als Schüler 2, Lehrer 2 dagegen gibt Schüler 2 eine bessere Note als Schüler 1.

Aspekt 2 ist der eigentlich uns interessierende Aspekt, während Aspekt 1 (Bias) nach Möglichkeit durch Kalibrierung vermieden werden sollte

Veranschaulichung der Kritik an Kappa

Tabelle: Problematik von Kappa

9	3		5	7
5	3		1	7

In der linken Tafel ergibt sich ein Kappa von 0.13, in der rechten Tafel ein Kappa von 0.26, obwohl wieder in beiden Fällen 12 von 20 Objekten ($f_o = 0.6$) übereinstimmend eingestuft wurden

Veranschaulichung der Kritik an Kappa

Erklärung:

- Die Randverteilungen in der rechten Tafel divergieren wesentlich stärker als in der linken Tafel (d.h. hier kann ein Kalibrierungsproblem vorliegen)
- Rechte Tafel:
  - Beobachter 1: $((5+7)/20, (1+7)/20) = (0.6, 0.4)$
  - Beobachter 2: $((5+1)/20, (7+7)/20) = (0.3, 0.7)$
- Linke Tafel:
  - Beobachter 1: $(12/20, 8/20) = (0.6, 0.4)$
  - Beobachter 2: $(14/20, 6/20) = (0.7, 0.3)$

Fazit: Beobachter müssen unbedingt kalibriert werden!

Erweiterungen von Kappa

- Ein weiterer Nachteil von Kappa ist, dass nur die Diagonale berücksichtigt wird
- Wenn die Bewertungsskala sehr viele verschiedene Merkmalsausprägungen besitzt, so liegen aufeinanderfolgende Ausprägungen oft nicht so weit auseinander
- Wenn zwei Beobachter sich nur gering in der Bewertung unterscheiden, so sollte eine Maßzahl dies auch berücksichtigen können
- Eine solche Maßzahl ist das *gewichtete Kappa*

Gewichtetes Kappa

- Das gewichtete Kappa wurde von Cohen (1968) vorgeschlagen
- Formal gehen alle Zellen der Kontingenztafel in die Berechnung ein
- Die Zellen auf der Hauptdiagonalen erhalten das höchste Gewicht (in der Regel Gewicht Eins), während die anderen Zellen ein geringeres Gewicht erhalten
- Die Idee ist, die Zellen umso geringer zu gewichten, je schlechter die Übereinstimmung der beiden Beobachter ist

Definition: Gewichtetes Kappa

Das gewichtete Kappa ist definiert als

$$\kappa_w = \frac{f_o^* - f_e^*}{1 - f_e^*} \ , \tag{3.4}$$

mit

$$f_o^* = \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} f_{ij}$$

$$f_e^* = \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} f_{i\cdot} f_{\cdot j}$$

Dabei wird f_o^* wie beim ungewichteten Kappa als relativer Anteil der Übereinstimmung beider Beobachter aufgefasst, während f_e^* die zufällige Übereinstimmung darstellt, wenn kein Zusammenhang bestehen würde

Wahl der Gewichte w_{ij}

Zwei populäre Vorschläge sind

$$w_{ij} = 1 - \frac{(i-j)^2}{(I-1)^2} \tag{3.5}$$

und

$$w_{ij}^* = 1 - \frac{|i-j|}{I-1} \; . \tag{3.6}$$

Gewichte w_{ij} und w_{ij}^* im 3×3-Fall

Tabelle: Gewichte w_{ij} einer 3×3–Tafel

1.0	0.75	0.0
0.75	1.0	0.75
0.0	0.75	1.0

Tabelle: Gewichte w_{ij}^* einer 3×3–Tafel

1.0	0.5	0.0
0.5	1.0	0.5
0.0	0.5	1.0

Die Zellen der größten Nichtübereinstimmung (Zelle $(1, 3)$ und $(3, 1)$ bei einer 3×3–Tafel) werden in beiden Fällen mit 0 gewichtet, die Zellen auf der Diagonalen mit Gewicht 1

Gewichte w_{ij} und w_{ij}^* im 4×4-Fall

Tabelle: Gewichte w_{ij} einer 4×4–Tafel

1.0	0.89	0.56	0.0
0.89	1.0	0.89	0.56
0.56	0.89	1.0	0.89
0.0	0.56	0.89	1.0

Tabelle: Gewichte w_{ij}^* einer 4×4–Tafel

1.0	0.67	0.33	0.0
0.67	1.0	0.67	0.33
0.33	0.67	1.0	0.67
0.0	0.33	0.67	1.0

Medizinisches Beispiel (Fortsetzung), Arzt A und B

Verwendung der Gewichte gemäß Formel (3.5):

$$\begin{aligned} nf_o &= 8 \cdot 1.0 + 20 \cdot 0.75 + 3 \cdot 0.0 \\ &\quad + 2 \cdot 0.75 + 15 \cdot 1.0 + 1 \cdot 0.75 \\ &\quad + 0 \cdot 0.0 + 5 \cdot 0.75 + 2 \cdot 1.0 \end{aligned}$$

und damit

$$f_o = \frac{46}{56} = 0.8214 \ ,$$

sowie

$$\begin{aligned} n^2 f_e &= 31 \cdot 10 \cdot 1.0 + 31 \cdot 40 \cdot 0.75 + 31 \cdot 6 \cdot 0.0 \\ &\quad + 18 \cdot 10 \cdot 0.75 + 18 \cdot 40 \cdot 1.0 + 18 \cdot 6 \cdot 0.75 \\ &\quad + 7 \cdot 10 \cdot 0.0 + 7 \cdot 40 \cdot 0.75 + 7 \cdot 6 \cdot 1.0 \end{aligned}$$

Medizinisches Beispiel (Fortsetzung), Arzt A und B

Also

$$f_e = \frac{2428}{56^2} = 0.7742$$

Damit erhalten wir

$$\kappa_w = \frac{0.8214 - 0.7742}{1 - 0.7742} = 0.209$$

Medizinisches Beispiel (Fortsetzung), Arzt A und B

Verwendung der Gewichte gemäß Formel (3.6):

$$\begin{aligned} nf_o &= 8 \cdot 1.0 + 20 \cdot 0.5 + 3 \cdot 0.0 \\ &\quad + 2 \cdot 0.5 + 15 \cdot 1.0 + 1 \cdot 0.5 \\ &\quad + 0 \cdot 0.0 + 5 \cdot 0.5 + 2 \cdot 1.0 \end{aligned}$$

und damit

$$f_o = \frac{39}{56} = 0.6964 \ ,$$

sowie

$$\begin{aligned} n^2 f_e &= 31 \cdot 10 \cdot 1.0 + 31 \cdot 40 \cdot 0.5 + 31 \cdot 6 \cdot 0.0 \\ &\quad + 18 \cdot 10 \cdot 0.5 + 18 \cdot 40 \cdot 1.0 + 18 \cdot 6 \cdot 0.5 \\ &\quad + 7 \cdot 10 \cdot 0.0 + 7 \cdot 40 \cdot 0.5 + 7 \cdot 6 \cdot 1.0 \end{aligned}$$

Medizinisches Beispiel (Fortsetzung), Arzt A und B

Also

$$f_e = \frac{1976}{56^2} = 0.6301$$

Damit erhalten wir

$$\kappa_{W^*} = \frac{0.6964 - 0.6301}{1 - 0.6301} = 0.179$$

Für dieses Beispiel erhalten wir also

$$\kappa = 0.159 < \kappa_{W^*} = 0.179 < \kappa_W = 0.209$$

Medizinisches Beispiel (Fortsetzung), Arzt C und D

Wir erhalten unter Verwendung der Gewichte aus (3.5)

$$\kappa_w = 0.601$$

und unter Verwendung der Gewichte aus (3.6)

$$\kappa_{w^*} = 0.525$$

Auch hier erhalten wir

$$\kappa = 0.445 < \kappa_{w^*} = 0.525 < \kappa_w = 0.601$$

Cohens Kappa: Zusammenfassung

- Nützliches Maß zur Übereinstimmung bei binären und metrischen Merkmalen
- Im ordinalen Fall sollte gewichtetes Kappa verwendet werden
- Zusätzlich sollte Kalibrierung geprüft werden

Zusammenhänge zwischen metrischen Merkmalen

Darstellung des Zusammenhangs, Korrelation und Regression

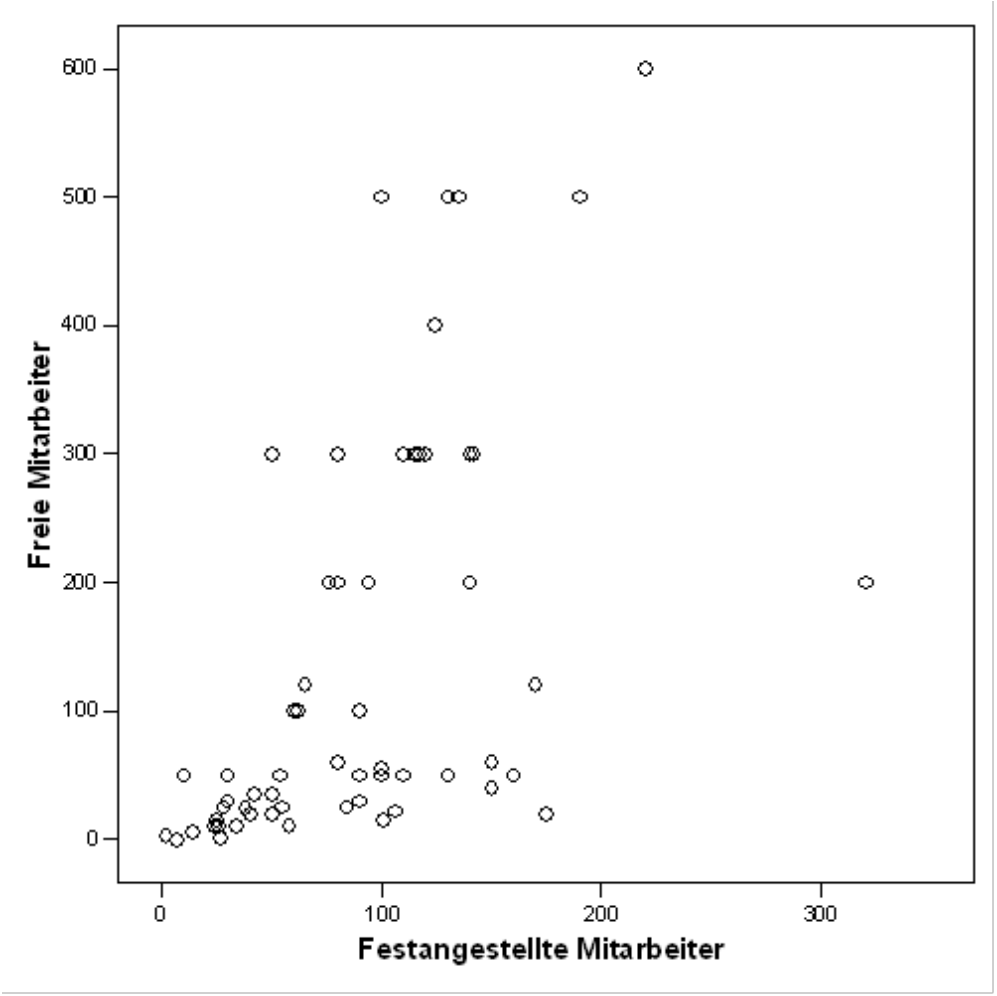
Daten liegen zu zwei metrischen Merkmalen vor:
Datenpaare $(x_i, y_i), \quad i = 1, \ldots, n$

Beispiel:
x: Anzahl der fest angestellten Mitarbeiter
y: Anzahl der freien Mitarbeiter

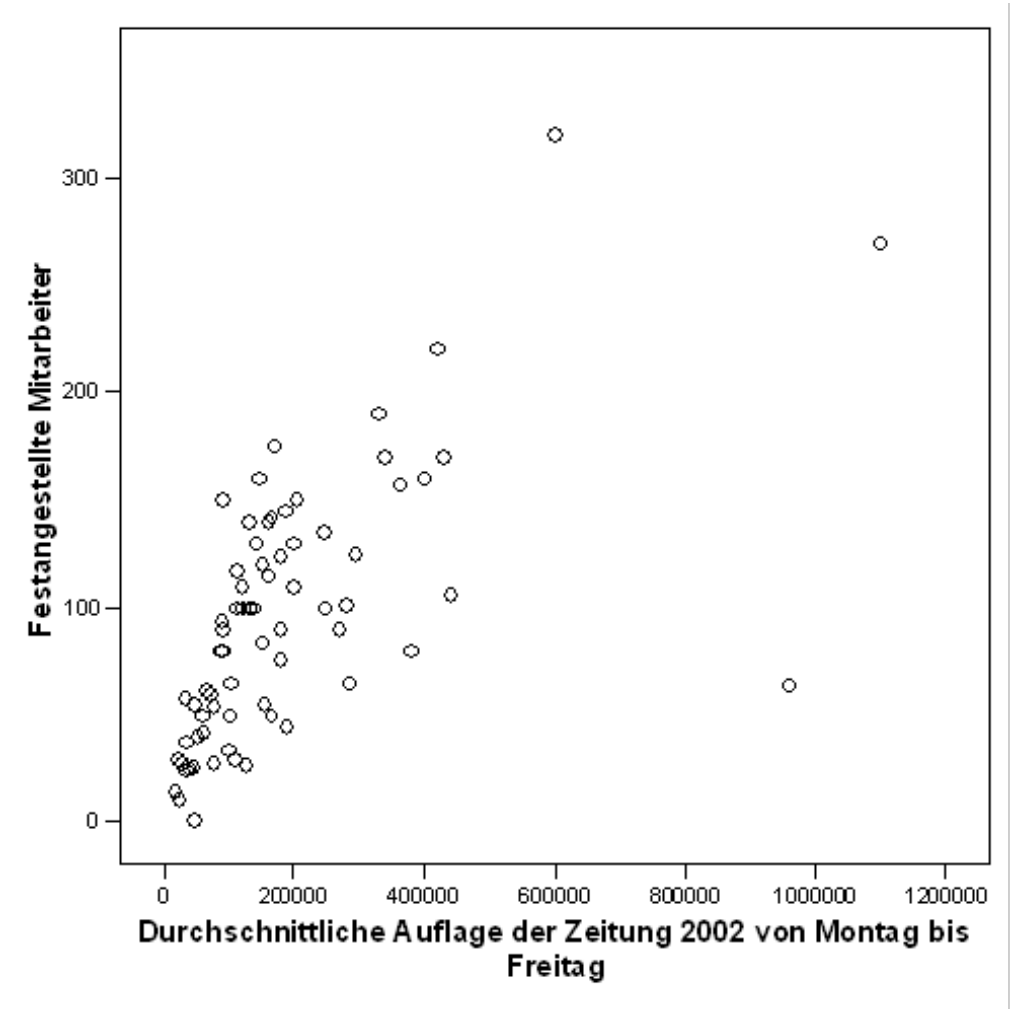
Frage:
Gibt es einen Zusammenhang zwischen diesen Merkmalen?
Wie lässt sich dieser Zusammenhang beschreiben?

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

Beispiel 1: Streudiagramm (mit SPSS)



Beispiel 2

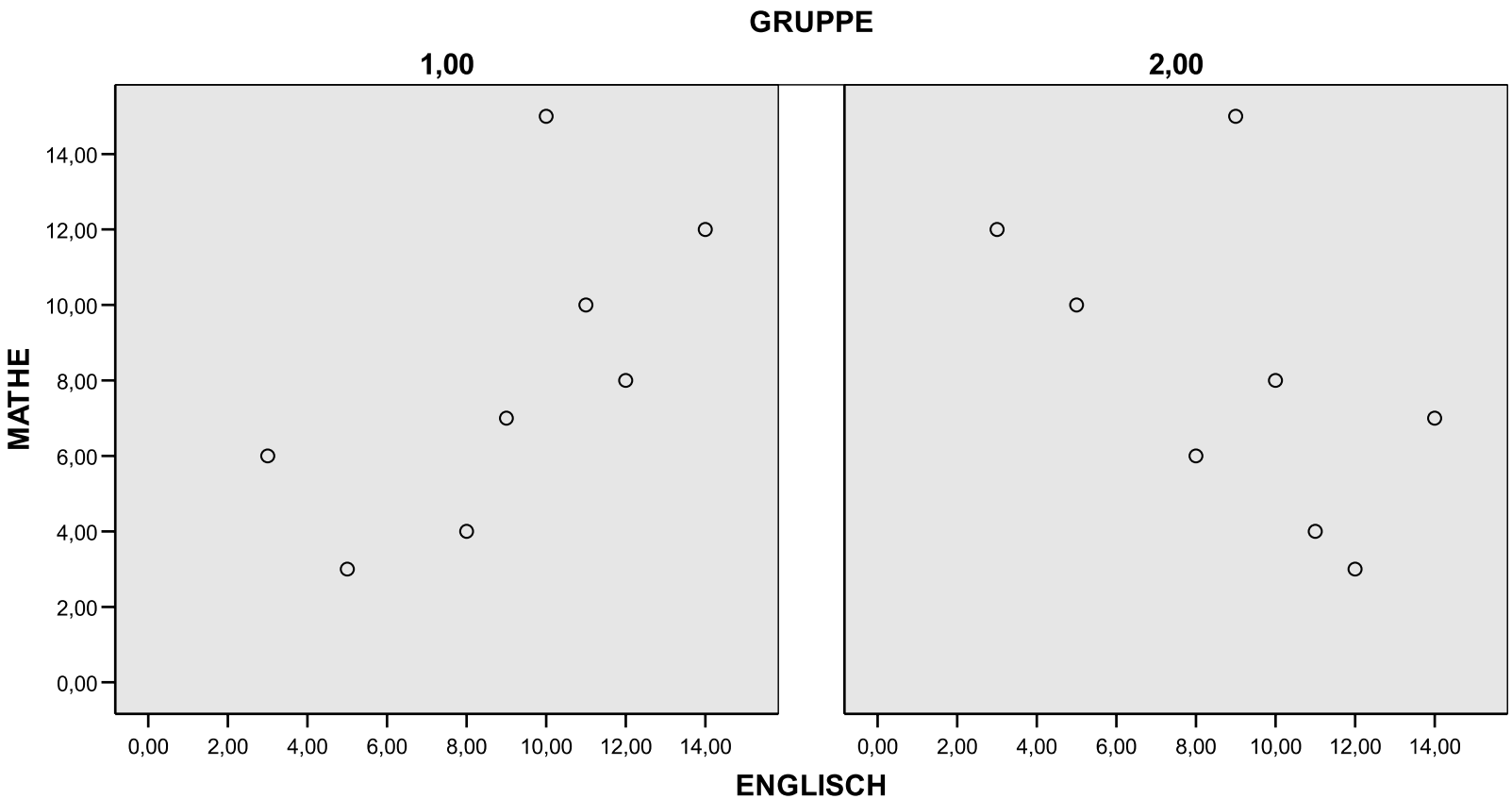

Festangestellte Mitarbeiter
300
200
100
0
0
200000
400000
600000
800000
1000000
1200000
Durchschnittliche Auflage der Zeitung 2002 von Montag bis Freitag

Beispiel 3

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 3 (Streudiagramme)



Kovarianz

Maß für den Zusammenhang der beiden Merkmale:

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

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

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.

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

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

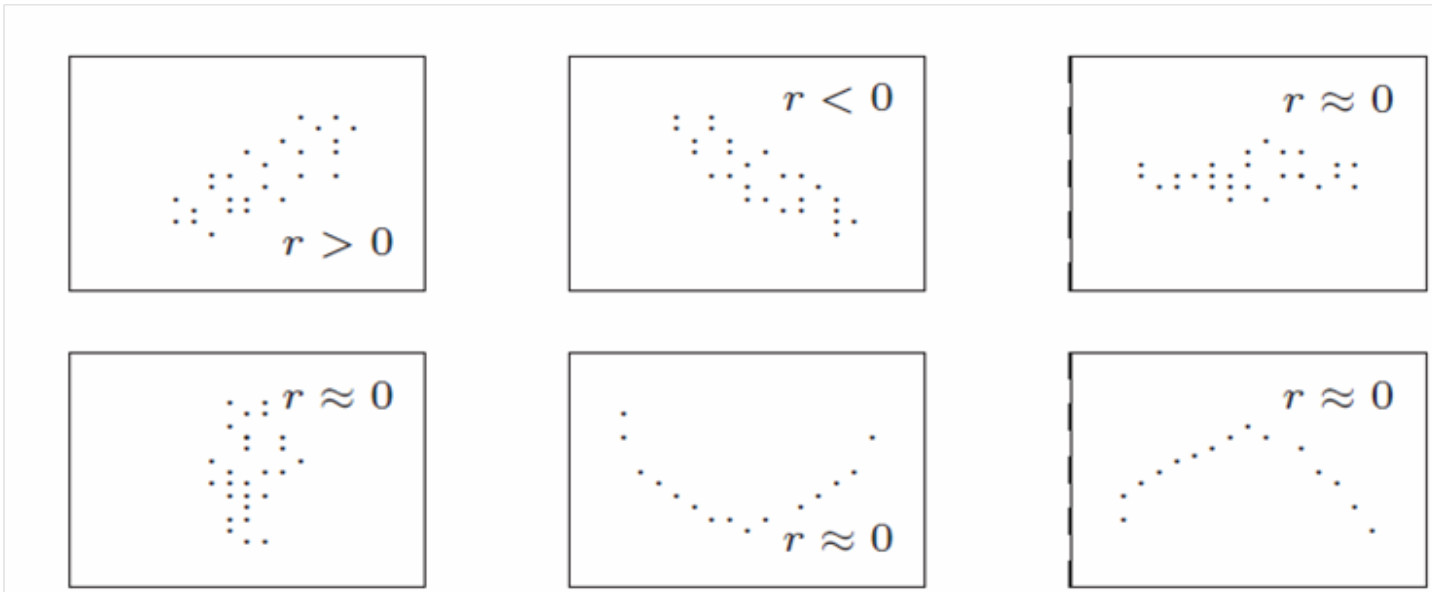
Gruppe 1: positiver linearer Zusammenhang
Gruppe 2: negativer linearer Zusammenhang

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

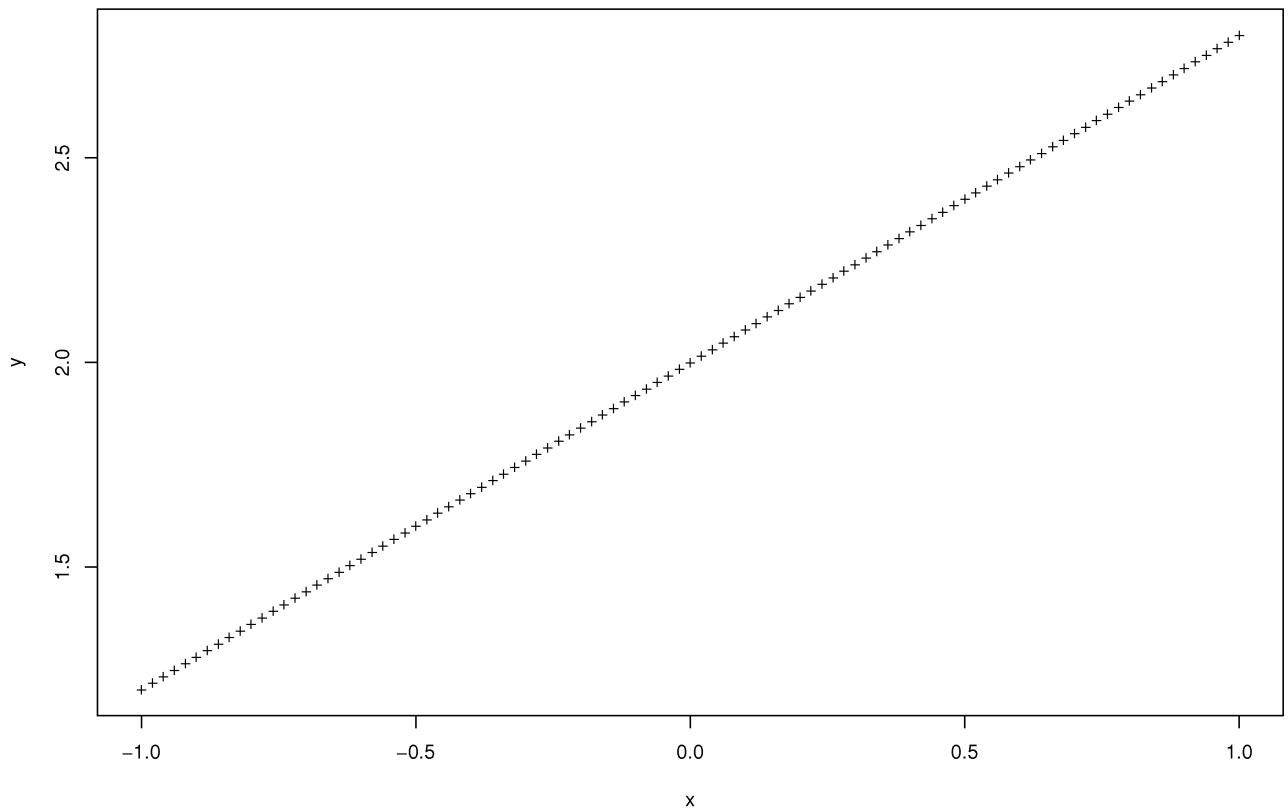
- r misst Stärke des *linearen* Zusammenhangs.



Punktkonfigurationen und Korrelationskoeffizienten (qualitativ)

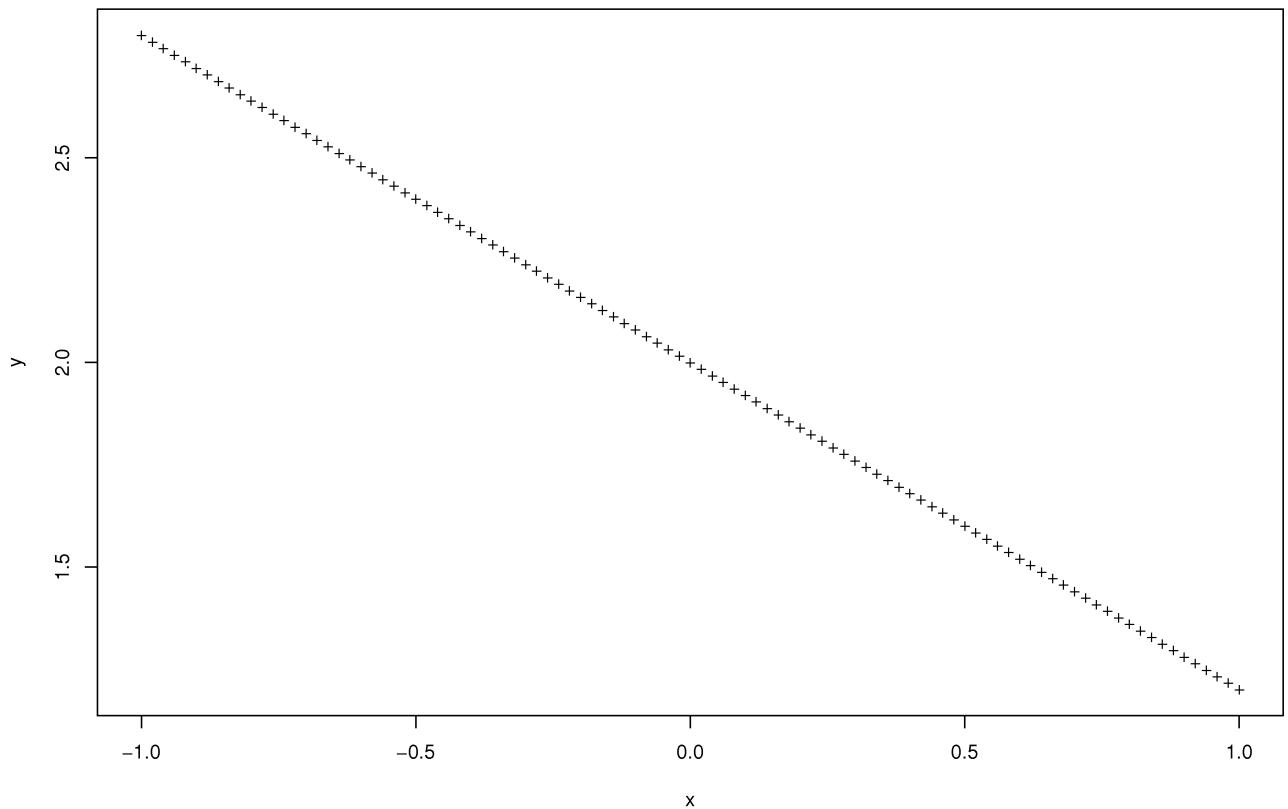
Einige Beispiele von exakten und verrauschten Zusammenhängen

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



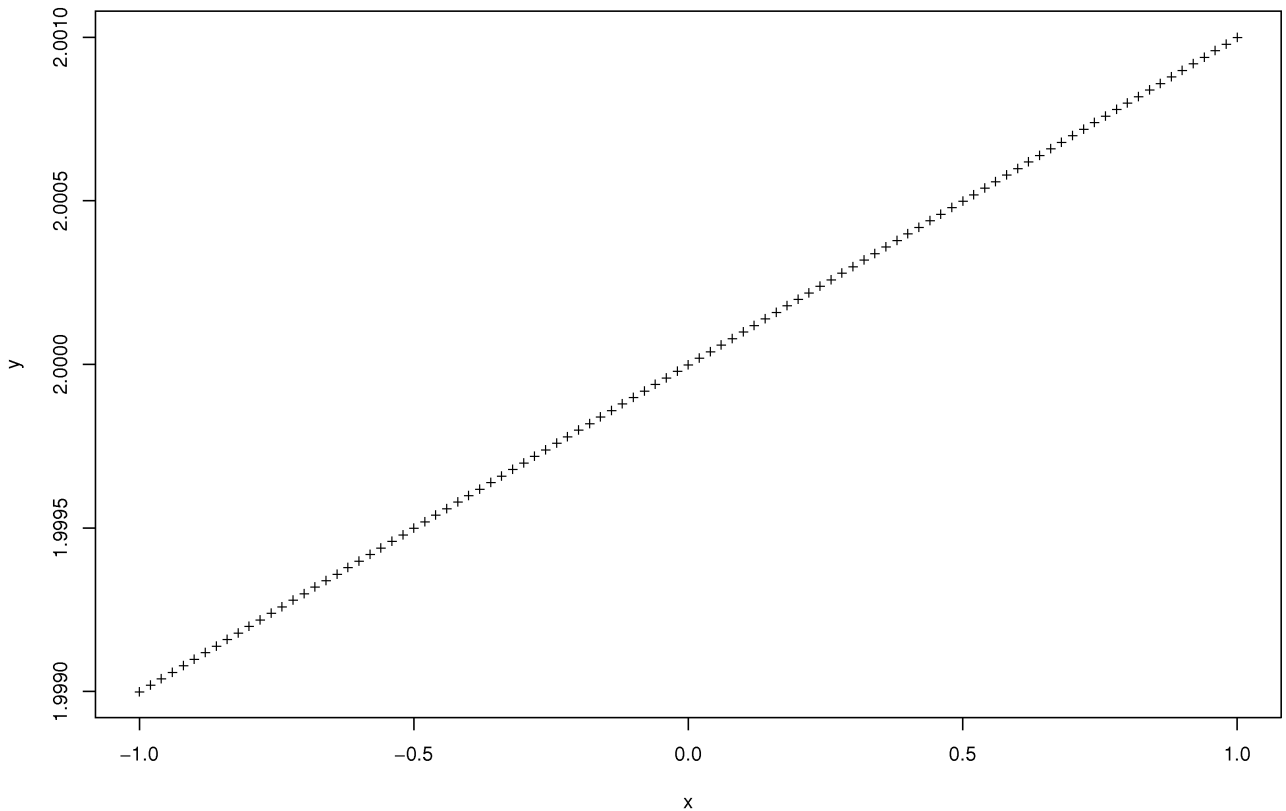
Einige Beispiele von exakten und verrauschten Zusammenhängen

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



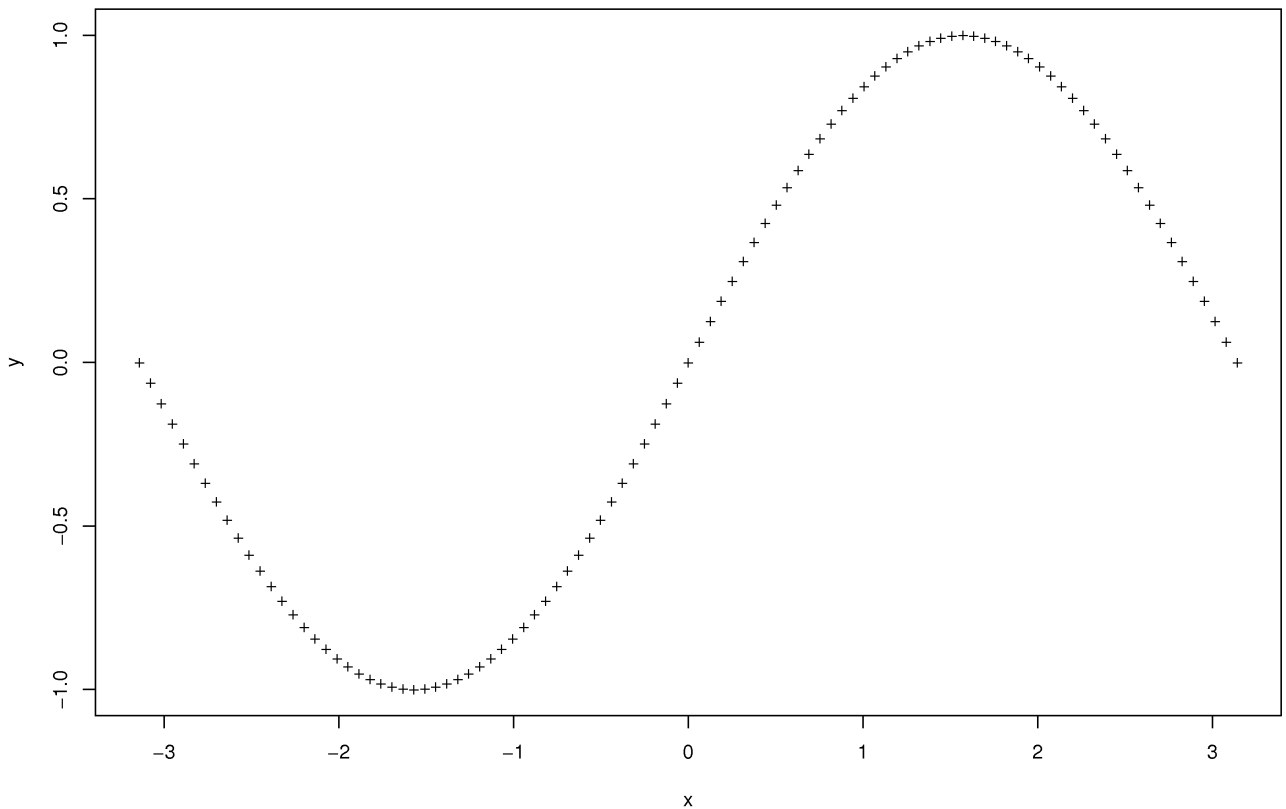
Einige Beispiele von exakten und verrauschten Zusammenhängen

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



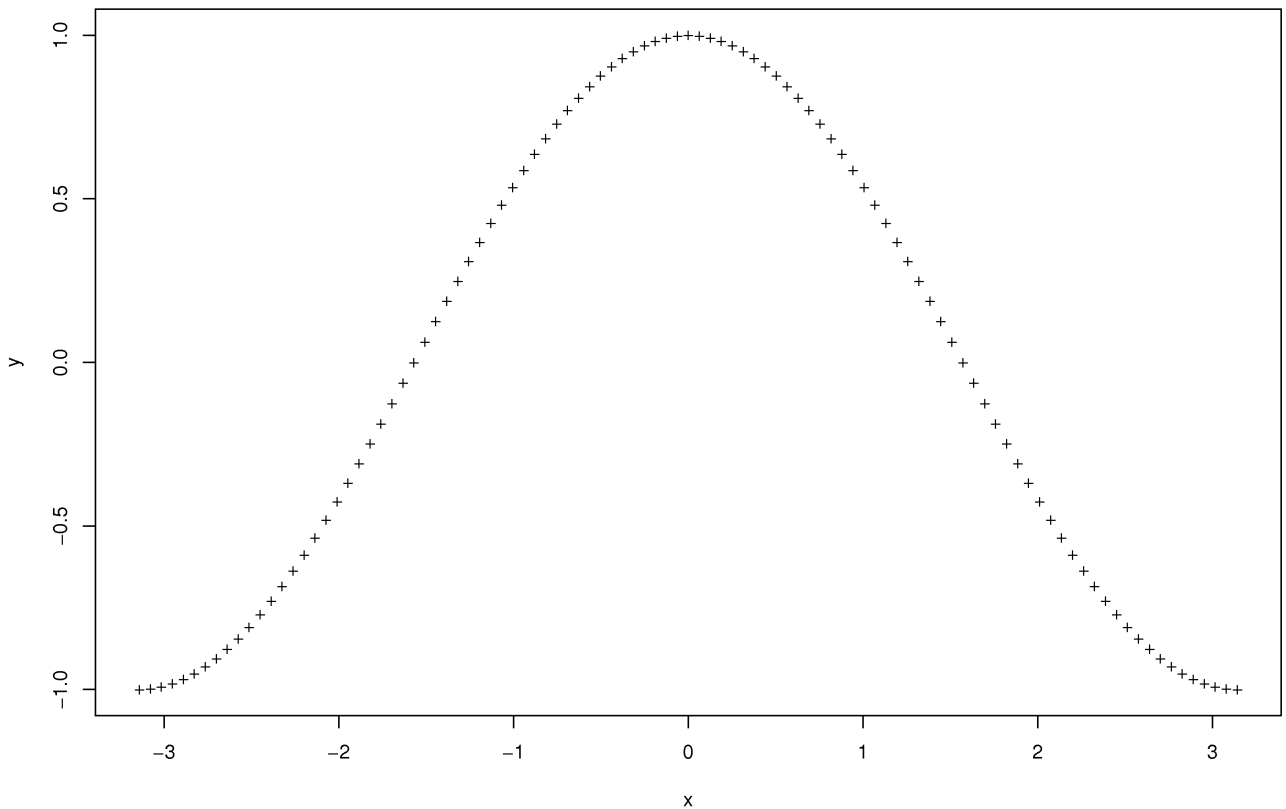
Einige Beispiele von exakten und verrauschten Zusammenhängen

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



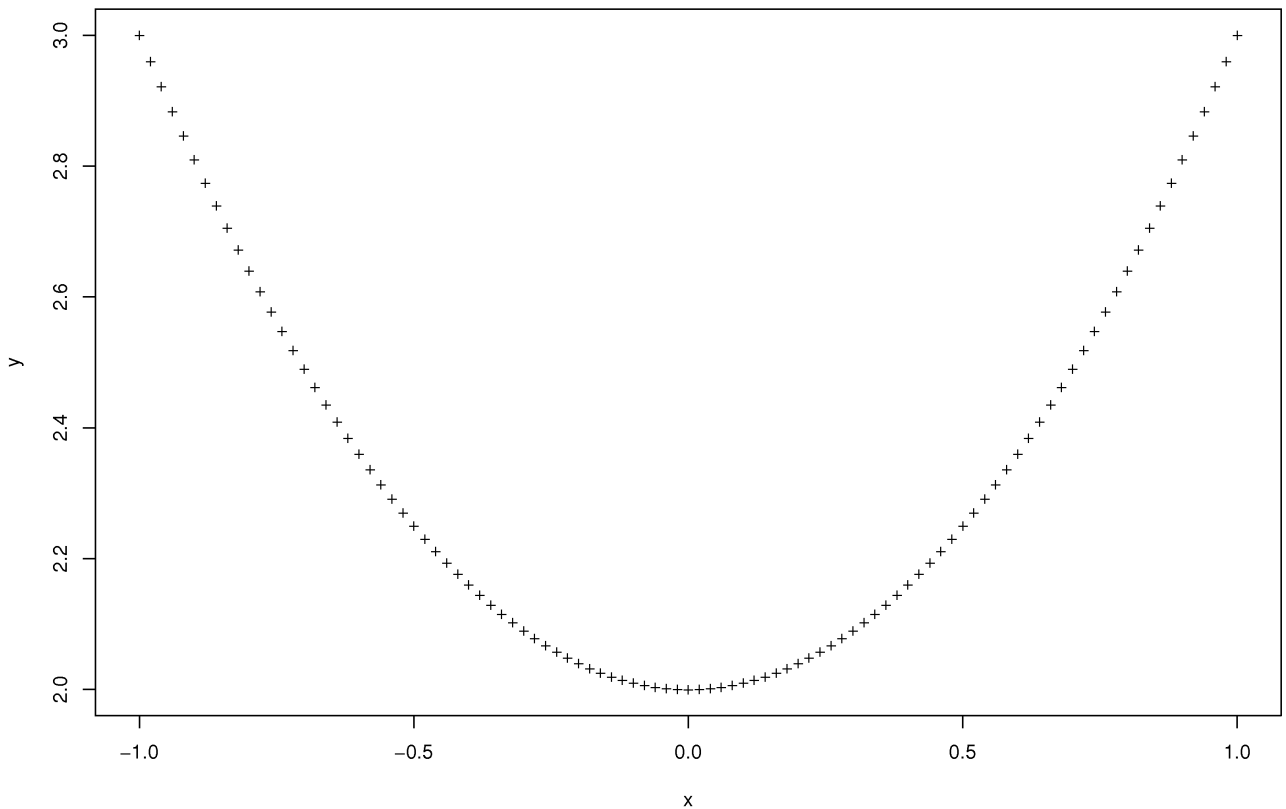
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 5: Periodische (unverrauschte) Funktion, $y = \cos(x)$, 101 equidistante Stützstellen im Intervall $[-\pi, \pi]$, $r =$



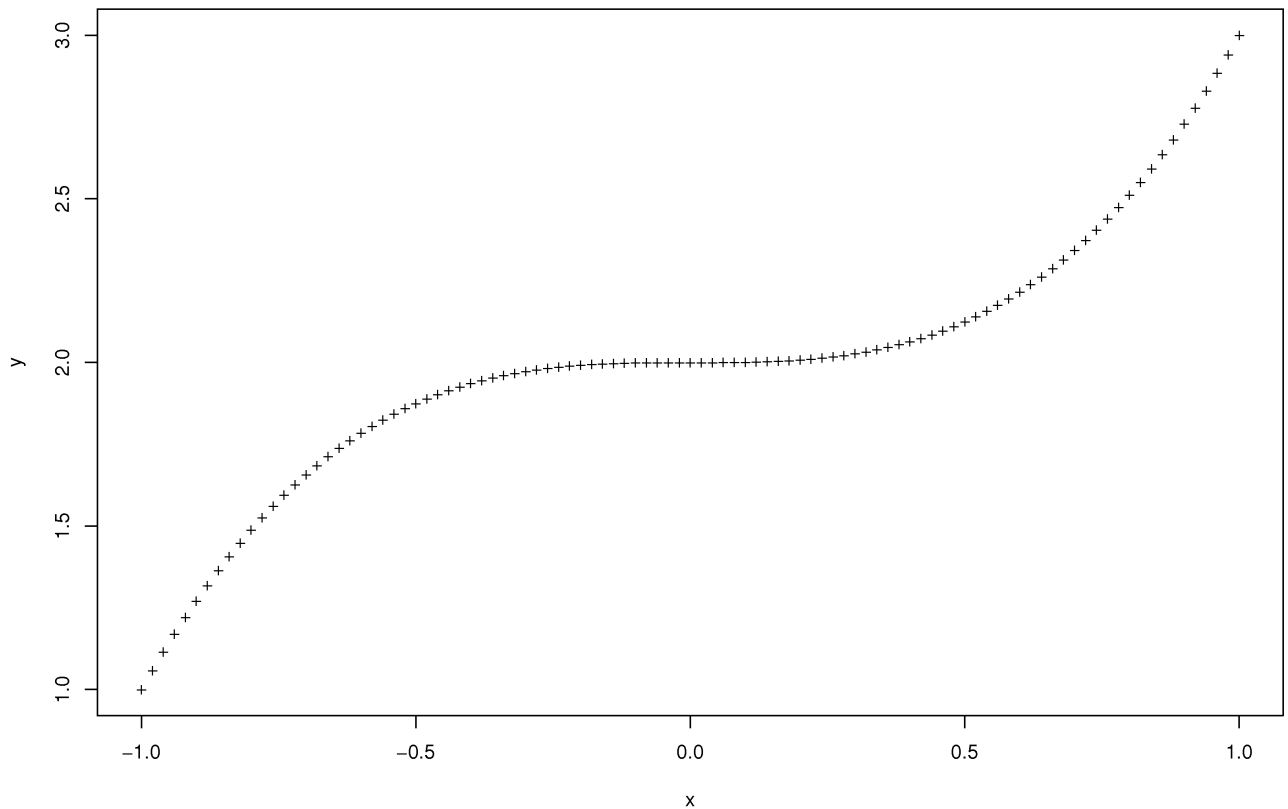
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 6: Quadratische (unverrauschte) Funktion, $y = x^2 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r =$



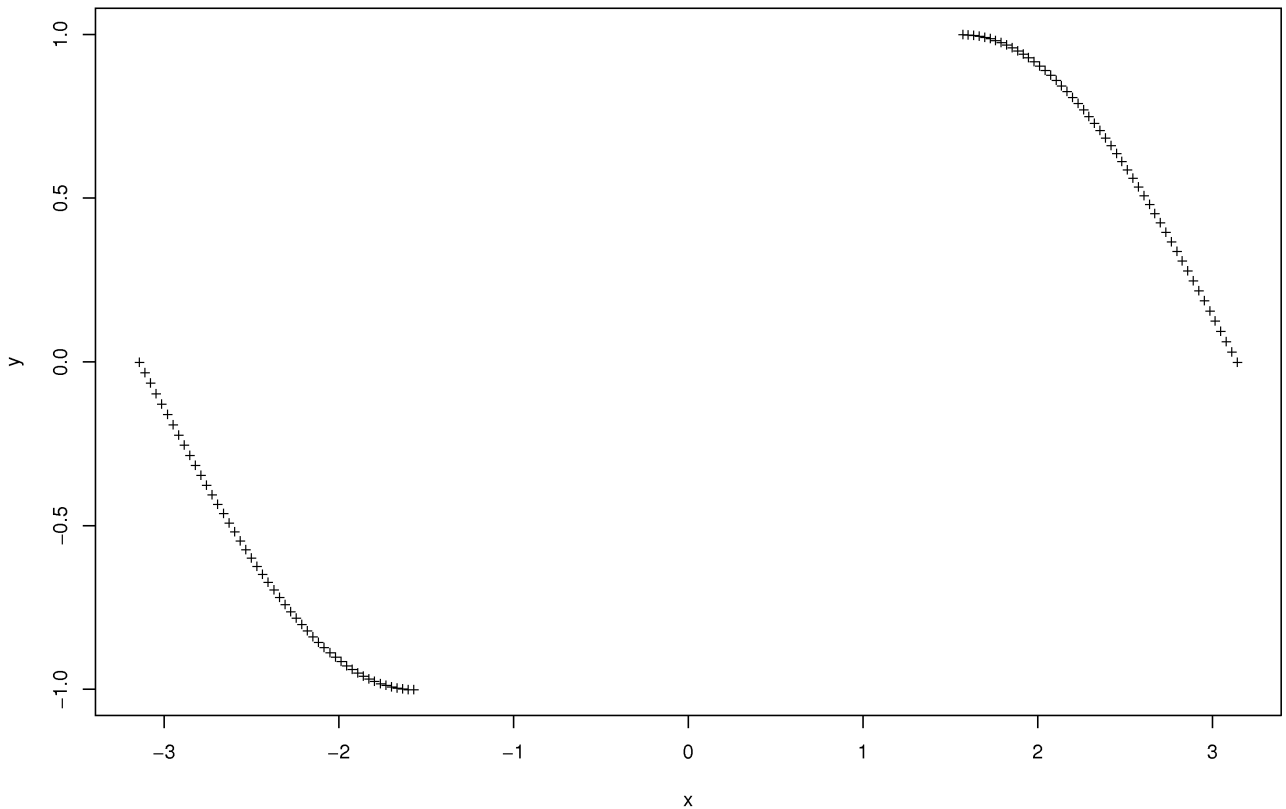
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 7: Kubische (unverrauschte) Funktion, $y = x^3 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r =$



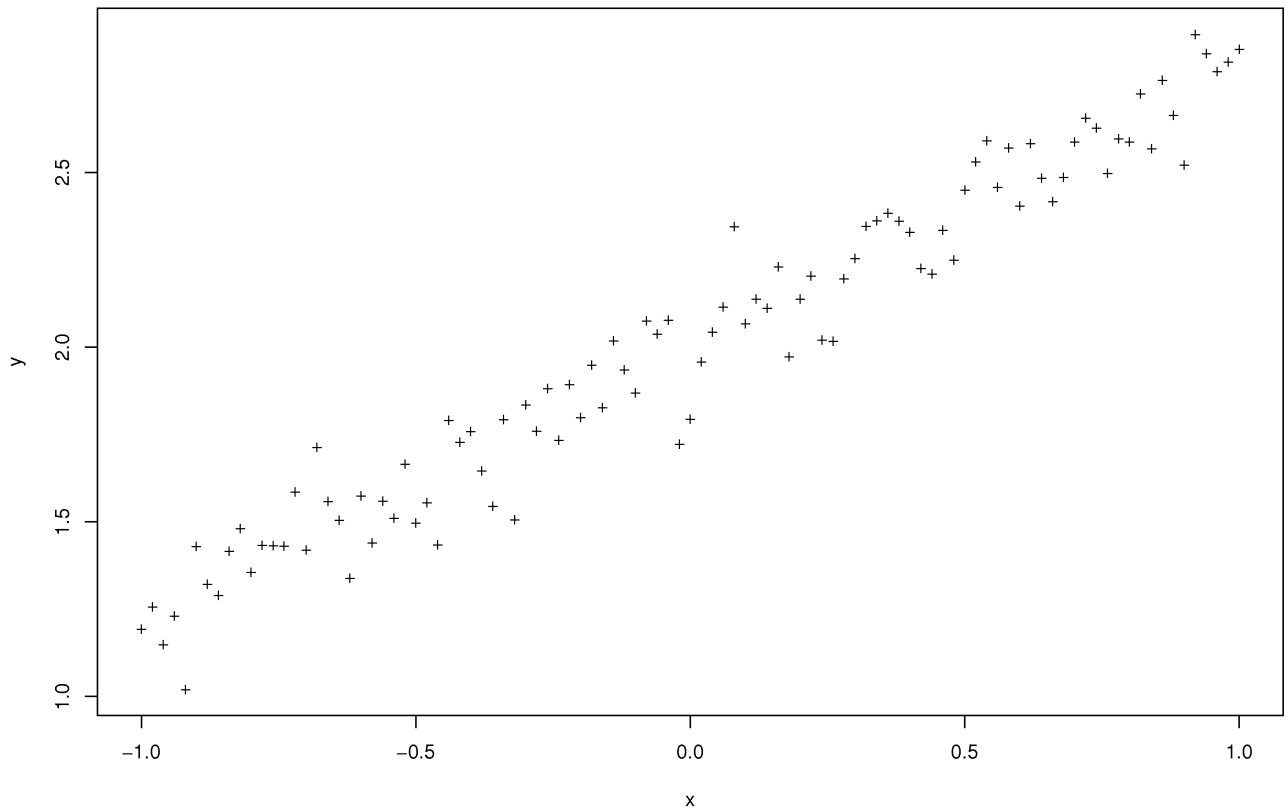
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 8: Abschnittweise definierte (unverrauschte) Funktion $y = \sin(x)$, 50 und 51 equidistante Stützstellen in den Intervallen $[-\pi, -\frac{\pi}{2}]$ und $[\frac{\pi}{2}, \pi]$, $r =$



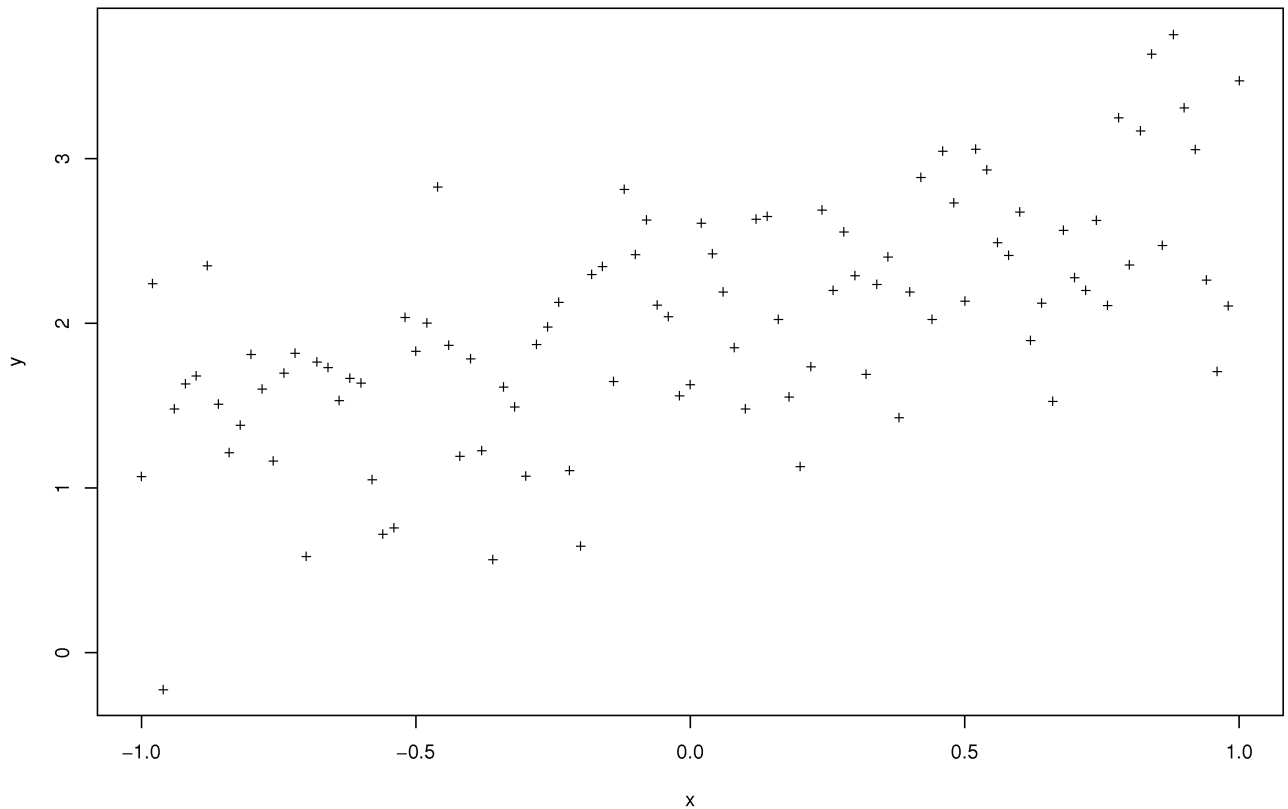
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 9: Lineare, schwach verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



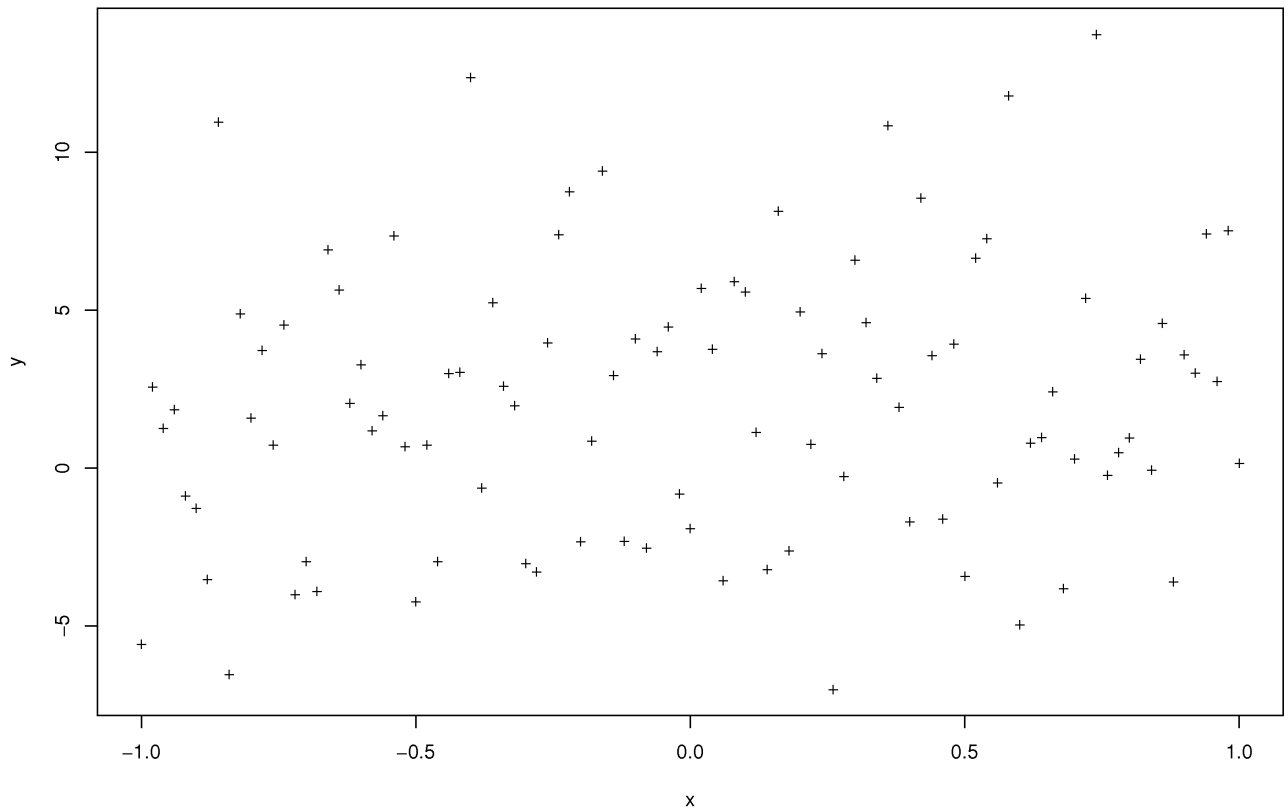
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 10: Lineare, stärker verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



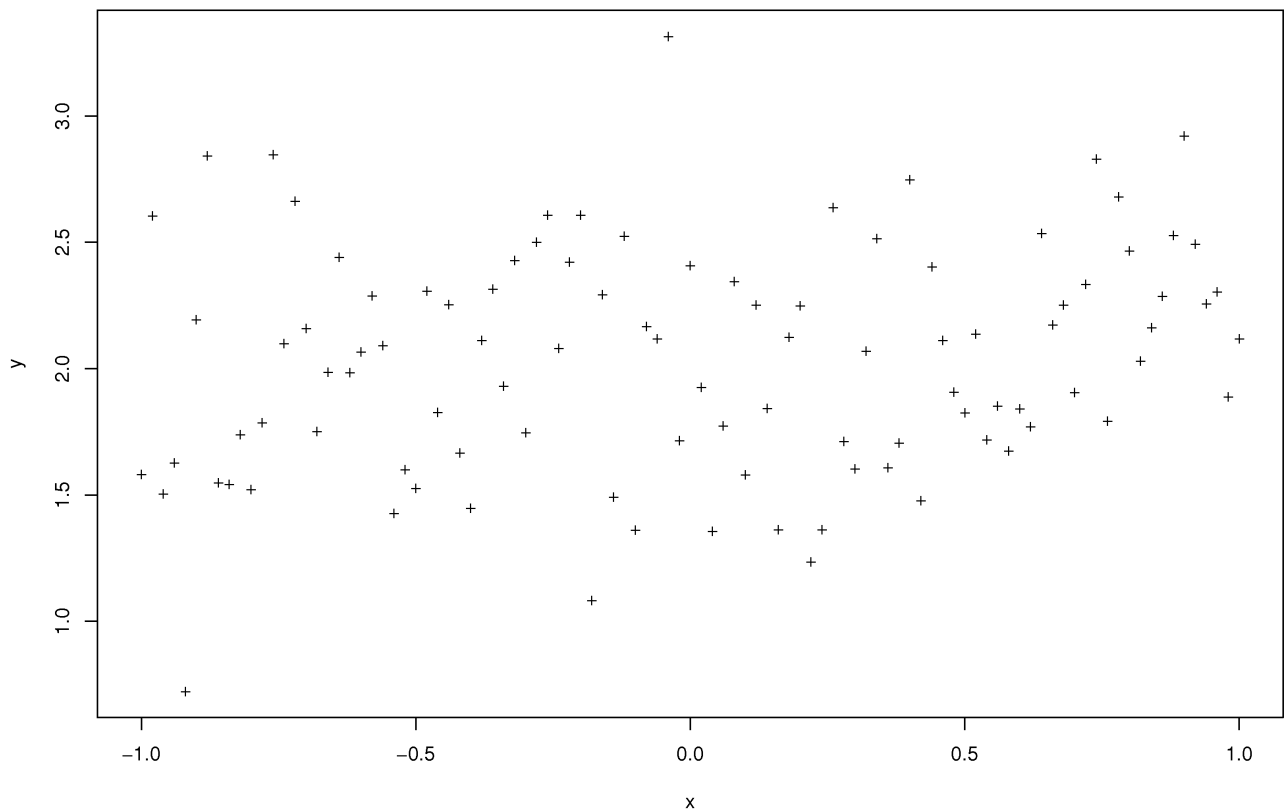
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 11: Lineare, stark verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 12: Lineare, stärker verrauschte Funktion, $y = 0.1x + 2.0 + N(0, 0.5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Lineare Transformationen

- Bei exakten lineare Zusammenhängen gilt:

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

- 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 $\tilde{Y}$

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

Vektor-Darstellung

Definiere die zentrierten Datenvektoren

$$x_Z = (x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_n - \bar{x})'$$
$$y_Z = (y_1 - \bar{y}, \ldots, y_i - \bar{y}, \ldots, y_n - \bar{y})'$$

$\Rightarrow r = \frac{x_Z' y_Z}{||x_Z||\,||\bar{y}_Z||}$, mit $||\,.\,||$ euklidische Norm.

Aus der Schwarz-Cauchy-Ungleichung folgt

$$|x_Z' y_Z| \leq ||x_Z||\,||y_Z||,$$

d.h. $-1 \leq r \leq +1$.

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: Urliste 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 Urliste.

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,

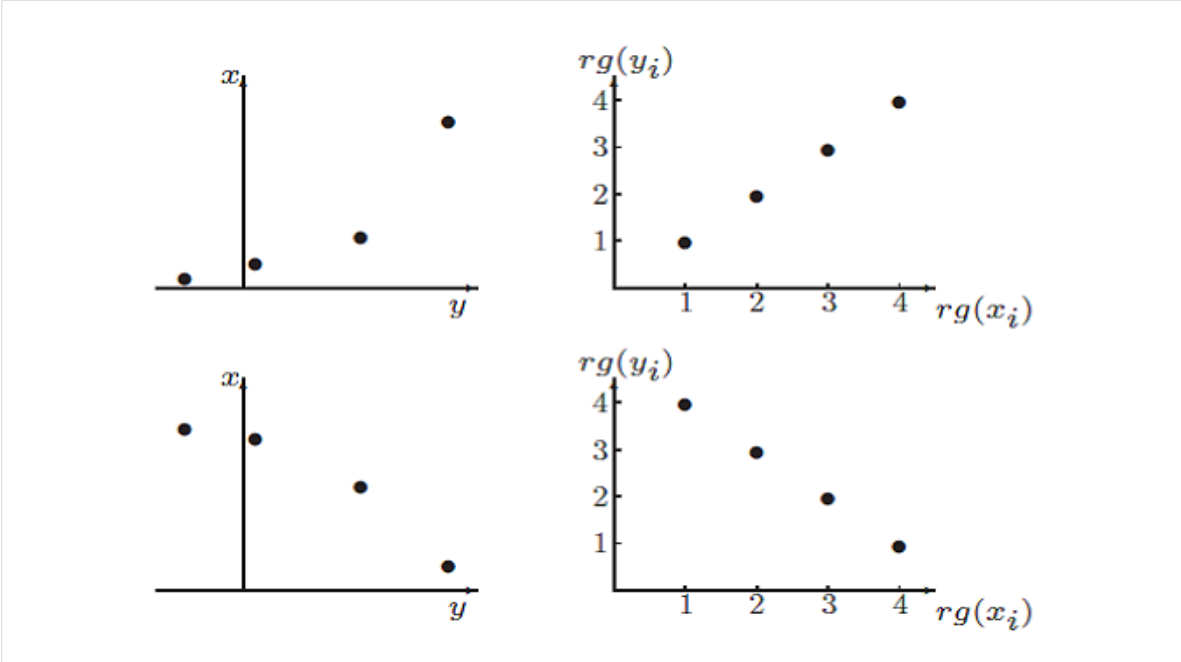
Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein

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

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

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

Extremfälle



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.

Kendall´s 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

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

Andere Varianten

- Goodman & Kruskal γ-Koeffizient

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

- Somers D wird typischerweise verwendet wenn Y binär ist
  T_x: Anzahl der Paare mit ungleichem y und gleichem x („Ties“ = Bindungen)

$$D_{xy} := \frac{N_C - N_D}{N_C + N_D + T_x} = \frac{N_C - N_D}{\text{Anzahl Paare mit ungleichem y}}$$

Kendall´s τ, Spearman´s r_{sp}

Beispiel:

					τ	r_{sp}
rg X	1	2	3	4	0.33	0.6
rg Y	2	1	4	3		
rg X	1	2	3	4	0.33	0.4
rg Y	1	3	4	2		

r_{sp} bestraft Abweichung stärker als τ

Unterschiede Kendall´s τ, Spearman´s ρ

- ρ verwendet Abstände auf der Rang-Skala
- τ orientiert sich an Paarvergleichen
- τ hat theoretische Entsprechung
- τ in der Regel kleiner als ρ

Dichotome und stetige Merkmale: Punktbiseriale Korrelation

Korrelations-Koeffizient zwischen dichotomen und metrischem Merkmal

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

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

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

Entspricht normiertem Abstand der Gruppenmittelwerte.

Dichotome und stetige Merkmale

- Beispiel 1 Kredit Scoring: Die Kreditwürdigkeit wird mit einem Scorewert gemessen (Schufa score)
  Dieser Scorewert soll auf seine Prognosegüte geprüft werden
  Variable : Y=1 (Eintrag nach 1.5 Jahren (Default) Y=0 kein Eintrag
- Beispiel 2: Blutserum Konzentration und stress-induzierte Herzinfarkte
  X: Marker für Herzinfarkt und
  Y: Infarkt während der WM (Gruppen)

ROC-Kurve

Jetzt Y dichotome Zielgröße und X metrische Einflussgröße:
$Y = 1 \longrightarrow$ Ausfall (krank)
$Y = 0 \longrightarrow$ kein Ausfall (gesund)
In der medizinischen Literatur ist das Testergebnis m:

$$\hat{Y}_i = 1 \Leftrightarrow x_i \geq c$$

Sensitivität und Spezifität

Richtig Positiv = Sensitivität:

$$f(\hat{Y} = 1|Y = 1) \quad = \quad f(x \geq c|Y = 1) = S_1(c)$$

$S_1(c)$ stellt die Survivorfunktion dar.

Richtig negativ = Spezifität:

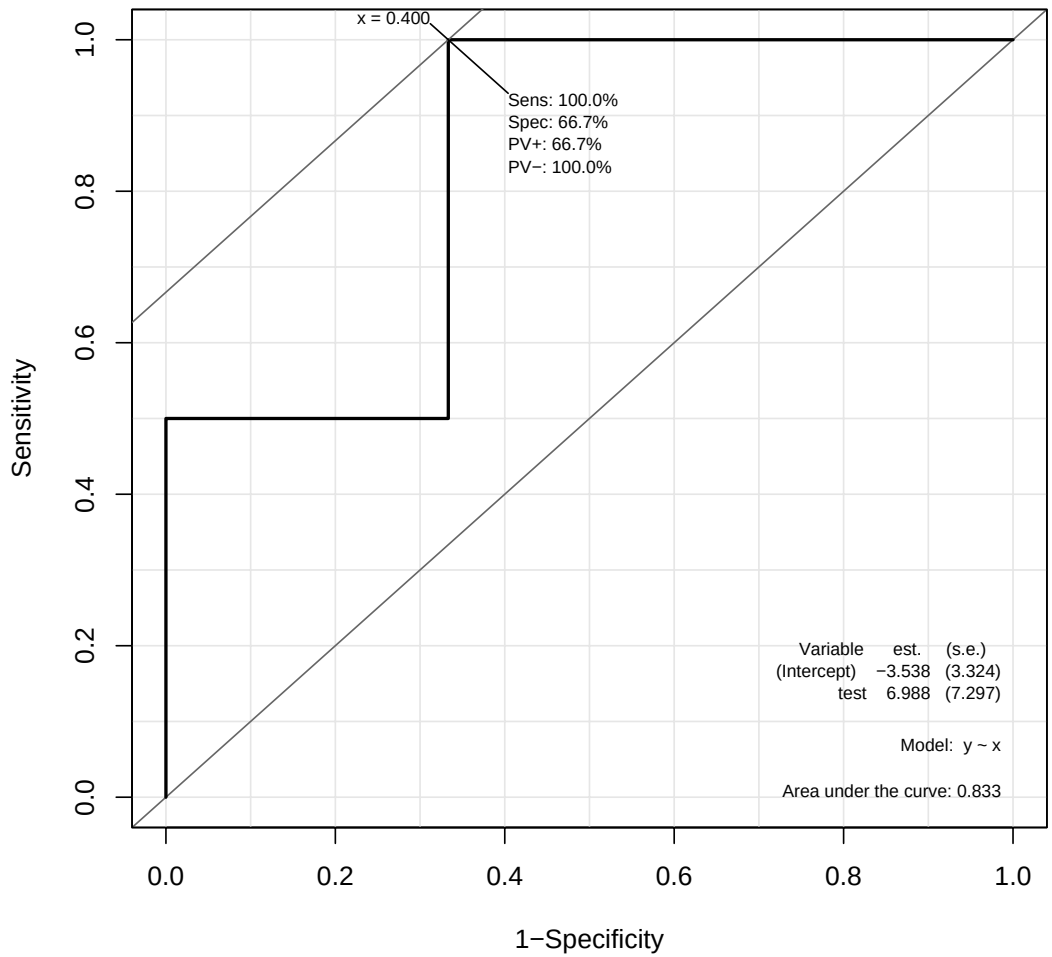
$$f(\hat{Y} = 0|Y = 0) = 1 - f(x \geq c|Y = 0) = 1 - S_0(c)$$

Falsch Positiv = 1- Spezifität:

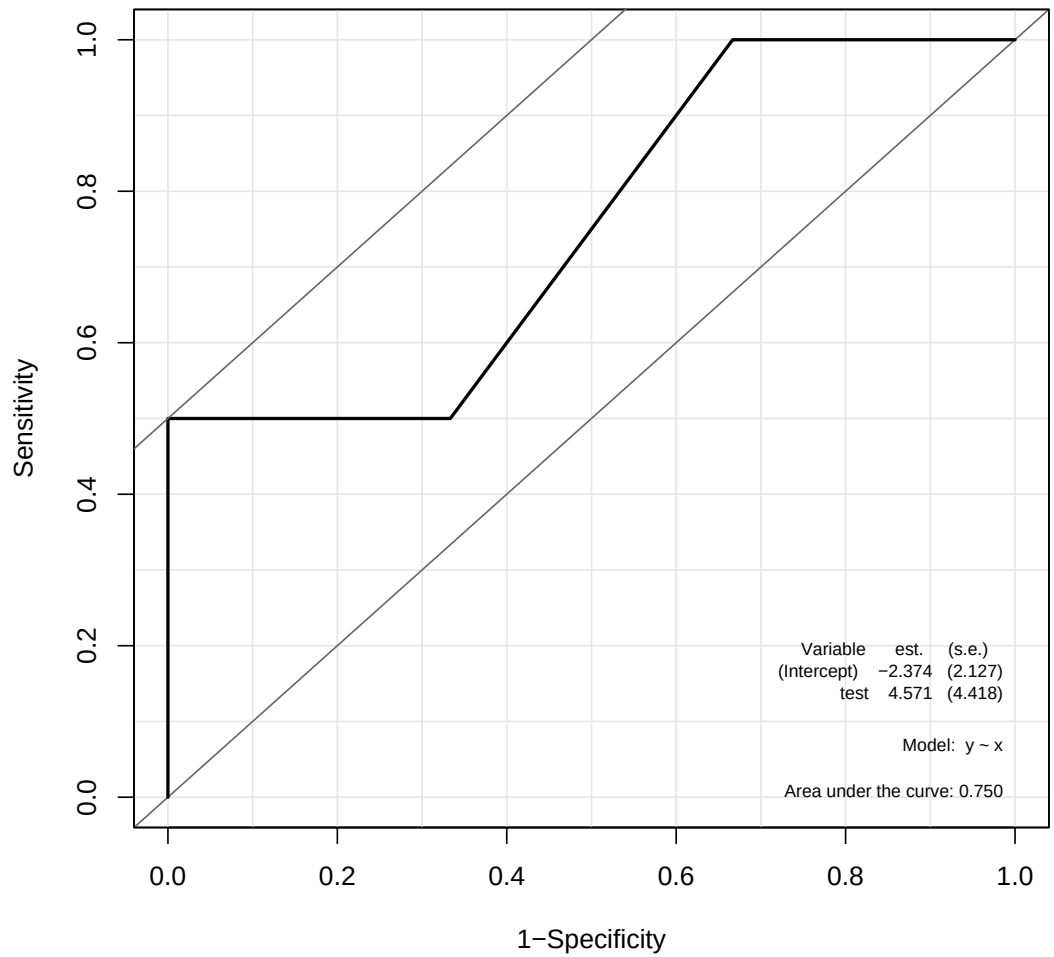
$$f(\hat{Y} = 1|Y = 0) = f(x \geq c|Y = 0) = S_0(c)$$

Die ROC-Kurve besteht aus den Punkten $(S_0(c), S_1(c))$

Beispiel für ROC-Kurve



Beispiel für ROC-Kurve mit Bindung



Maß zur Bewertung der Kurve: AUC

$$AUC = \int_{t=0}^{1} ROC(t)dt \tag{3.7}$$

Dies stellt die Fläche unter der Kurve dar.

Es gilt:

$$AUC = \frac{N_C + 0.5 * N_E}{N} \tag{3.8}$$

Dabei bezeichnet N_C die Anzahl der konkordanten Paare, N_E die Anzahl der identischen Paare, und N die Anzahl der Paare mit unterschiedlichem Y.

GINI- Koeffizient

Normierte Fläche zwischen Winkelhalbierender und ROC- Kurve

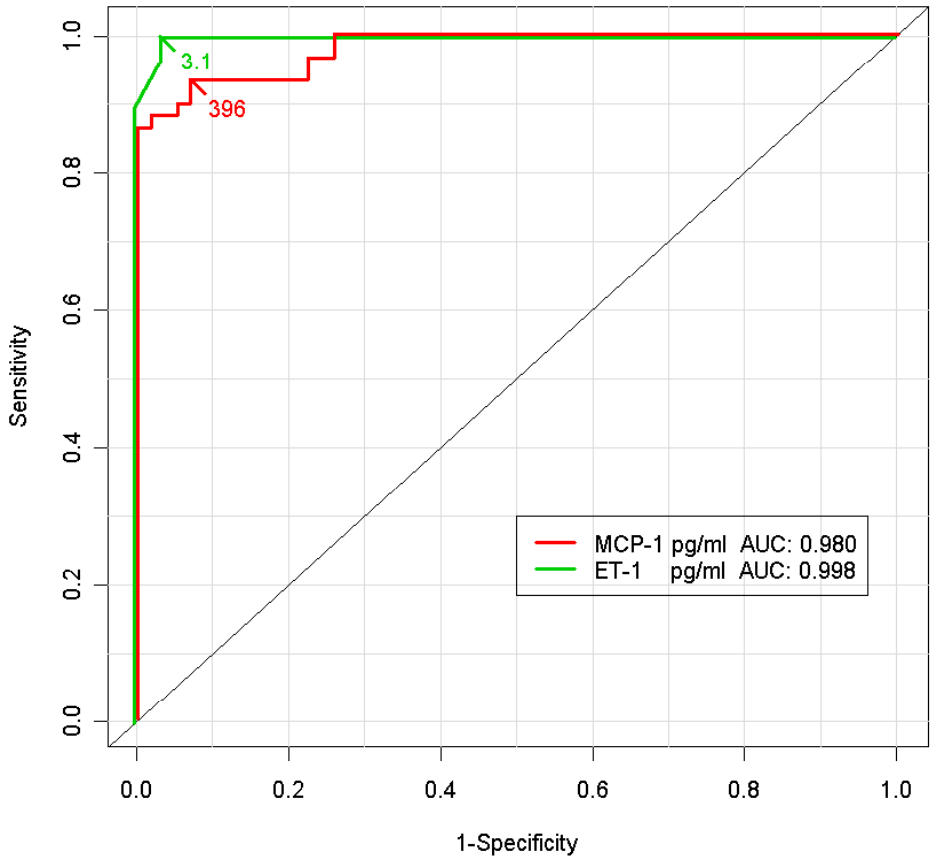
$$GINI = 2 \cdot (AUC - \frac{1}{2}) = 2 \cdot AUC - 1 \tag{3.9}$$

$$GINI = \frac{N_C - N_D}{N} \tag{3.10}$$

N_C: Anzahl der konkordanten Paare
N_D: Anzahl der diskordanten Paare
N: Anzahl der Paare mit ungleichem Y
N = $n_0 \cdot n_1$ mit n_i Anzahl der Daten mit Y=i.

Der GINI entspricht dem Somers D.

Beispiel: Stress induzierter Herzinfarkt



Korrelationsmatrix

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

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

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 $X = 0$

Bestimmung der Regressionsgerade

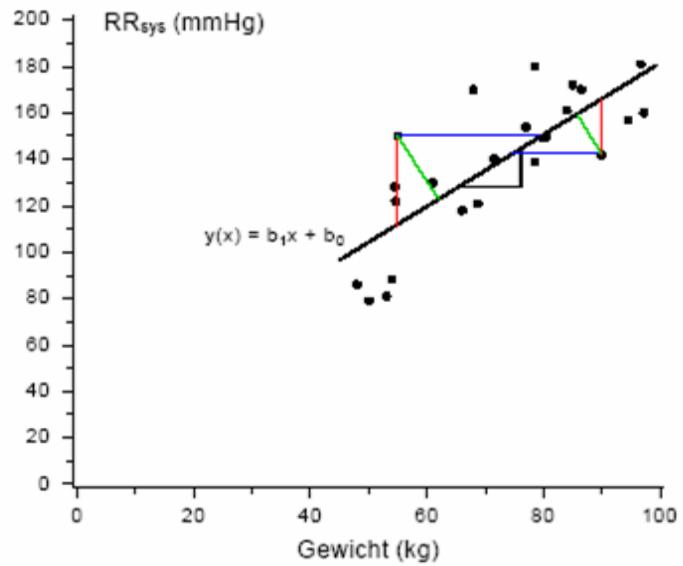
Welche Gerade ist die „Beste“?

- 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



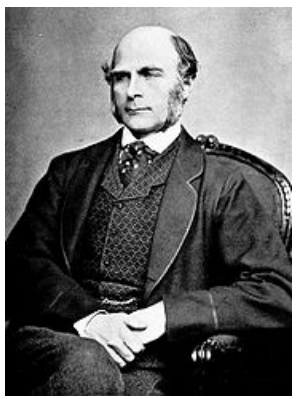
Methode der kleinsten Quadrate

- 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

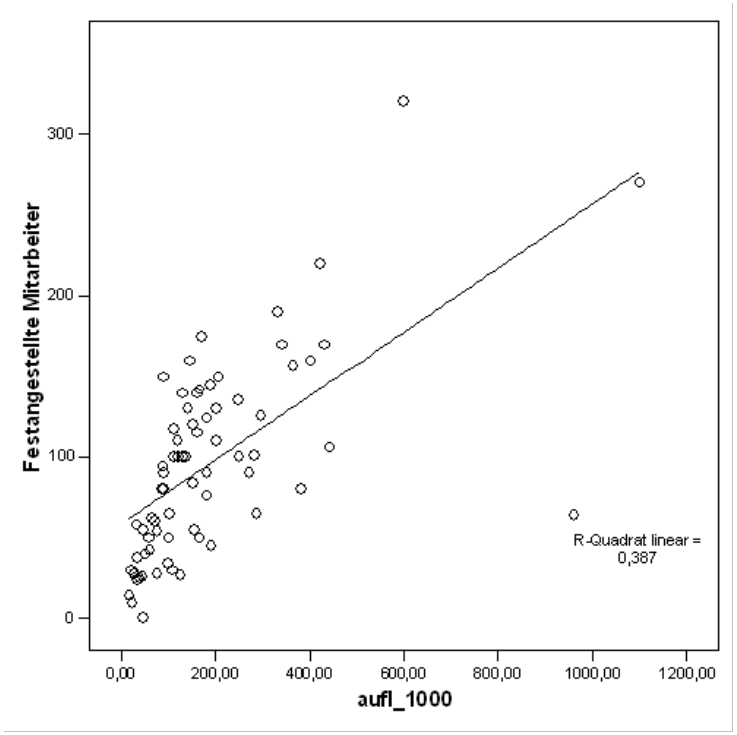
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{S_{xy}}{S_x^2}.$$

Die Residuen berechnen sich durch

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

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

Beispiel: Zahl der Mitarbeiter in Abhängigkeit von der Auflage



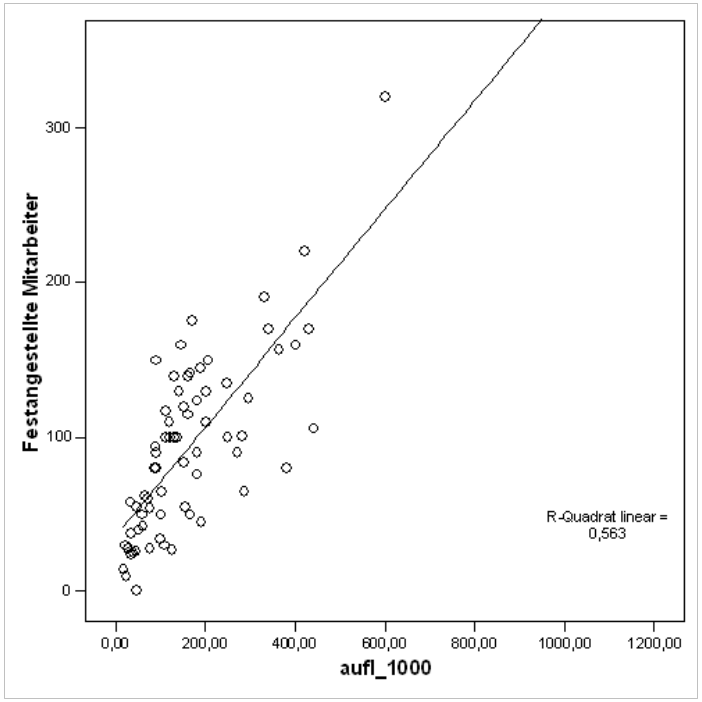
Interpretation:
Mit einer Auflagensteigerung von 1000 ist durchschnittlich die Einstellung von 0.199 Mitarbeitern verbunden.

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

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:

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

Umkehrregression I

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 II

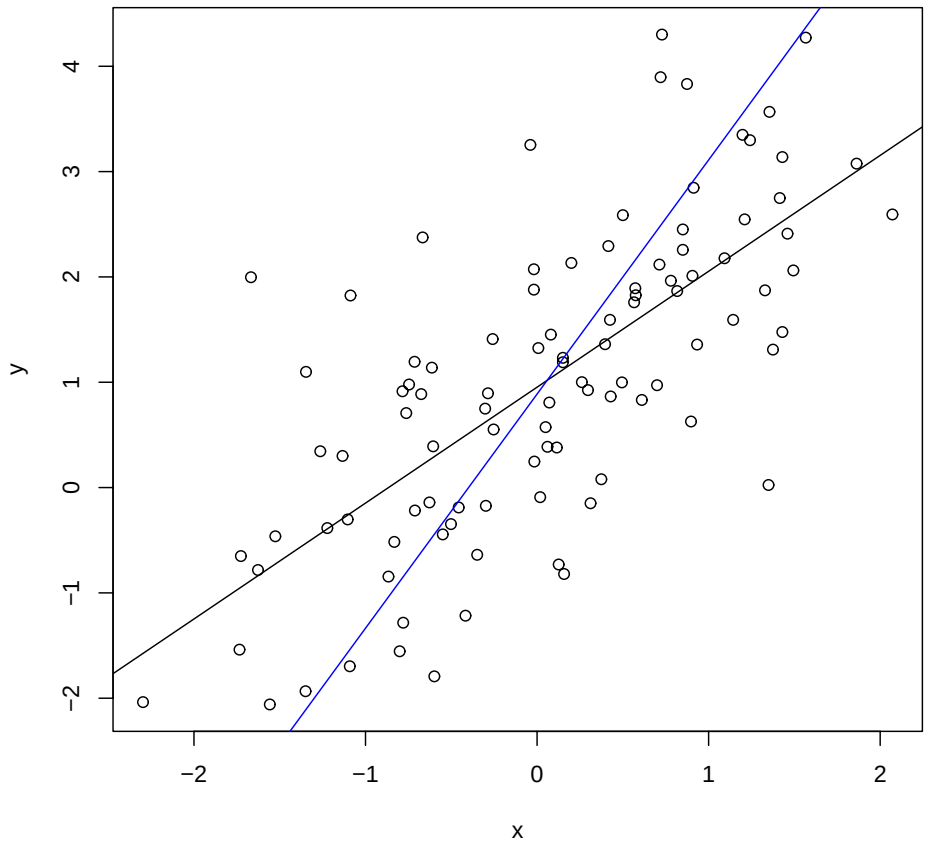
Es gilt:

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

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

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

Beispiel: Umkehrregression



Orthogonale Regression

Falls man die orthogonalen Abstände zur Gerade minimiert, erhält man eine Gerade zwischen Regression und Umkehrregression. Löse Minimierungsproblem in α, β

$$(\alpha_{ORR}, \beta_{ORR}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \underbrace{\frac{1}{1+\beta_i^2}(y_i - \alpha - \beta x_i)^2}_{\text{Orthogonaler Abstand}}$$

$$\begin{aligned} \hat{\beta}_{ORR} &= \frac{1}{2S_{XY}}\left[(S_Y^2 - S_X^2) + \sqrt{4S_{XY}^2 + (S_Y^2 - S_X^2)^2}\right] \\ \hat{\alpha}_{ORR} &= \bar{y} - \hat{\beta}_{ORR} \cdot \bar{x} \end{aligned}$$

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 b 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 b 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 b 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 b 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

Partielle Korrelation

Ziel:
Bestimmung der Korrelation zweier Merkmale unter „konstant halten“ eines dritten Merkmals

Beispiel:
Korrelation der Zahl der freien und festen Mitarbeiter unter konstanter Auflage

Idee:
Herausrechnen des Einflusses des dritten Merkmals durch lineare Regression

Partieller Korrelationskoeffizient (Definition)

Es soll der lineare Zusammenhang zwischen X und Y bei festen Z bestimmt werden. Betrachte lineare Regressionen

$$\begin{aligned} X &= \alpha_1 + \beta_1 Z + \varepsilon_1 \\ Y &= \alpha_2 + \beta_2 Z + \varepsilon_2. \end{aligned}$$

Aus den Daten (x_i, y_i, z_i) werden die Parameter nach der KQ-Methode geschätzt. Man erhält die bereinigten Variablen X^{BZ} und Y^{BZ} als Residuen der Regressionen:

$$\begin{aligned} X^{BZ} &= X - \hat{\alpha}_1 - \hat{\beta}_1 Z \\ Y^{BZ} &= Y - \hat{\alpha}_2 - \hat{\beta}_2 Z \end{aligned}$$

Dann heißt die Maßzahl

$$r_{XY|Z} = r_{X^{BZ}Y^{BZ}}$$

partieller Korrelationskoeffizient zwischen X und Y unter Z.

Berechnung der partiellen Korrelation

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.

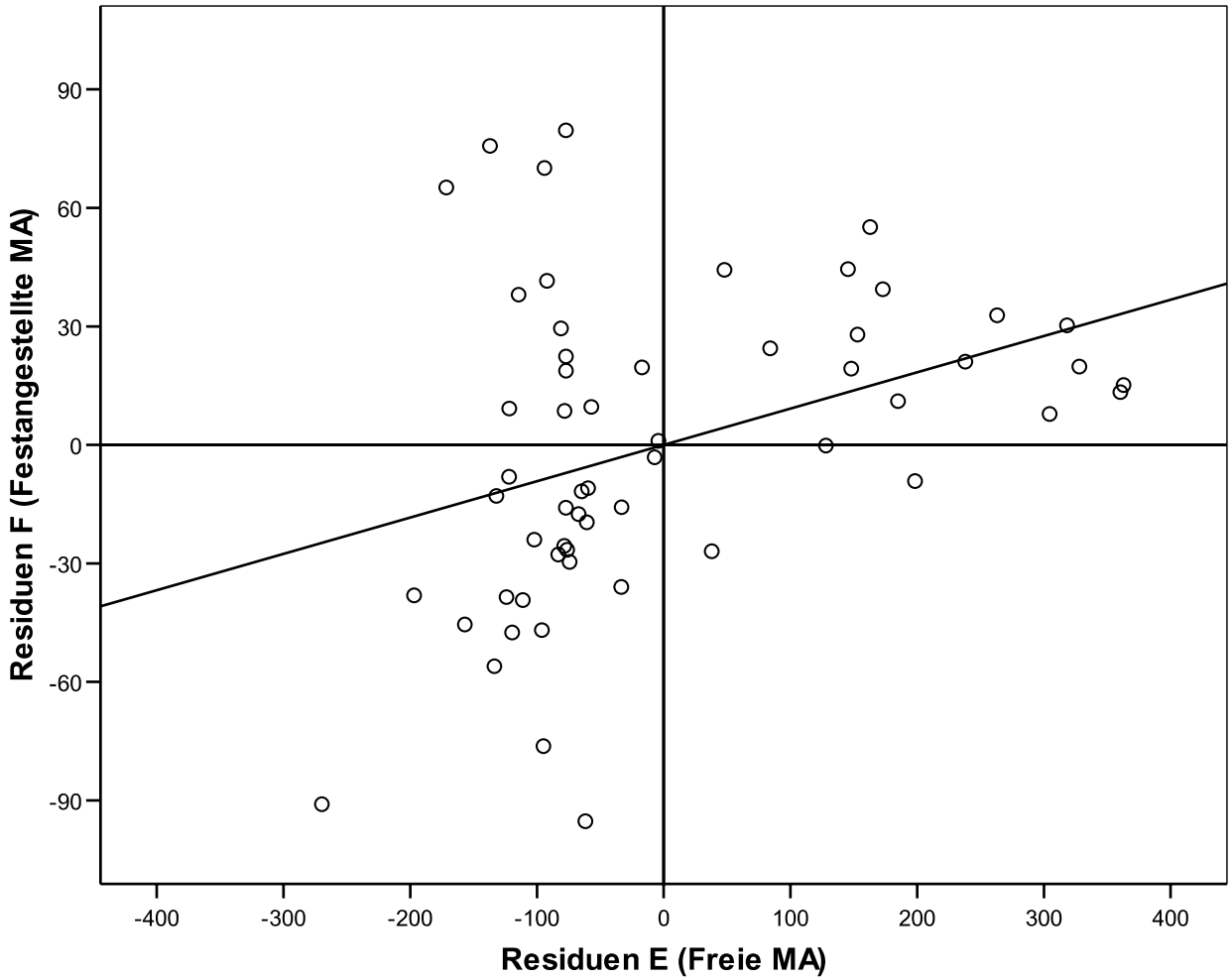
Einfache Korrelation

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)



Multiples Regressionsmodell

Gegeben sind die Zielgröße Y und die Einflussgrößen X_k

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_p \cdot x_p + \varepsilon$$

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

Beispiel: Festangestellte und Freie Mitarbeiter

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

$$
\begin{aligned}
FAM &= a + b_1 \cdot FM + b_2 \cdot AT + \varepsilon \\
FAM &= 31 + 0.092 \cdot FM + 0.32 \cdot AT + \varepsilon \\
FAM &= 67 + 0.17 \cdot FM + f
\end{aligned}
$$

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
- Herausrechnen von Störgrößen
- Auswahl von relevanten Einflussgrößen

Erweiterungen

Erweiterungen des Modells beinhalten:

- Nichtlineare Zusammenhänge
- Einbeziehung von nominalen Merkmalen als Einflussgrößen (z.B. Geschlecht, Nationalität, etc.)
- Binäre Zielgrößen (krank/gesund)

Das nichtlineare Regressionsmodell

Zusammenhang zwischen X und y: β kann Vektor sein

$$Y = f(X, \beta) + \varepsilon$$

KQ- Schätzer aus Daten Y_i, X_i:

$$\hat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

Berechnung oft nicht mit geschlossenen Formeln, aber numerisch möglich, z.B. mit Paket nls in R.
Beispiel: Wachstum mit Obergrenze β_1

$$y = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 * t)} + \varepsilon$$

Messfehler

Bei metrischen Größen: Es gibt einen wahren Wert X und eine Messung X^*
z.B.

$$X^* = X + \underbrace{U}_{\textit{Messfehler}}$$

$$\mathbb{E}(U) = 0$$

$\Rightarrow$ klassischer additiver Messfehler
Verhältnis von Streuung des Messfehlers zur Gesamtstreuung

$$r = \frac{S_X^2}{S_{X^*}^2} = \frac{S_X^2}{S_X^2 + S_U^2}$$

Zwei unabhängige Messungen (Messwiederholung): X_{i1}^* und X_{i2}^* liefern Reliabilität:

$$r = \rho(X_1^*, X_2^*)$$

2 Messungen

Messfehlerbestimmung durch Analyse der Differenz

$$X_{i1}^* - X_{i2}^*$$

Untersuchung auf

- Systematische Unterschiede
- Ausreißer
- Streuung des Messfehlers

Grafische Darstellung durch Bland - Altman Plot:
Scatterplot : Differenz der Messungen gegen den Mittelwert

Explorative Datenanalyse EDA

- Auffinden von Strukturen in Daten
- Hypothesen generieren
- Grafische Darstellungen

Tukey:
EDA must be considered as an open-ended, highly interactive, iterative process, whose actual steps are segments of a stubbily branching, tree-like pattern of possible actions.

Explorative Datenanalyse EDA

- Auffinden von Strukturen in Daten
- Hypothesen generieren
- Grafische Darstellungen

Tukey:
EDA must be considered as an open-ended, highly interactive, iterative process, whose actual steps are segments of a stubbily branching, tree-like pattern of possible actions.

Tukey and Tukey (1985): Since the aim of exploratory data analysis is to learn what seems to be, it should be no surprise that pictures play a vital role in doing it well. There is nothing better than a picture for making you think of questions you had forgotten to ask (even mentally).

Data Mining, Statistical learning

- Sehr große Datensätze
- Suche nach Mustern, Trends Clustern Ausreißern
- Supervised and unsupervised statistical learning

Literatur: Hastie, Tibshirani, Friedman: The Elements of Statistical Learning (2. ed 2008)
Neben statistischen Modellen und komplexen Verfahren sind komplexe Grafiken ein wichtige Aspekt

Grafische Darstellungen

Unterscheide nach Ziel:

- Darstellung von Daten (für Publikum)
- Werkzeuge zur Exploration (für ForscherIn)
- Information (Internetnutzer)

Strategien für Exploration und Präsentation

- Exploration
  - Zielt auf Erkenntnisgewinn
  - Hauptsächlich ein Nutzer
  - Kaum Skalen und Legenden
  - Hochgradig interaktiv und wenig persistent

Strategien für Exploration und Präsentation

- Exploration
  - Zielt auf Erkenntnisgewinn
  - Hauptsächlich ein Nutzer
  - Kaum Skalen und Legenden
  - Hochgradig interaktiv und wenig persistent
- Presentation
  - Präsentiert Ergebnisse
  - Optimiert für eine breite Leseschaft
  - Intensive Verwendung von Skalen und Legenden
  - Im statischen Druck ohne Interaktionen

Informationsgrafiken

Ursprünglich aus Kommunikationswissenschaft: Eigenständige journalistische Darstellungsform,
USA Today 1985, in Deutschland seit der Erstausgabe des Focus 1993

- Information in Grafikform
- Visualisierung abstrakter Vorgänge
- Darstellung gesammelter Daten, informativ und exakt
- Visuelle Repräsentation von Grundzusammenhängen
- Grafisch–bildhafte Darstellungen von Sachverhalten wie Ablaufdiagramme oder Erklärungsmodelle
- Kombination von Bild- und Textnachricht

Aufteilung in

1. Prinzipdarstellungen
2. Bildstatistiken
3. Kartografische Infografiken

Allgemeine Kriterien

- Ein eigenständiges Bild entwerfen
- Verständlichkeit
- sachliche/wertende Überschriften
- Erläuterungstexte mit Begriffen, Abkürzungen
- Quellenangabe

Inhalt

- Den Inhalt klar strukturieren
- Thema (Hauptaussage) visualisieren
- Den Eye-Catcher-Effekt erzielen
- Zusammenhänge sichtbar machen
- Mengen korrekt darstellen Proportionen müssen stimmen
- Leser (Benutzer) abholen, wo er steht
- Eher wenig Hintergrundwissen voraussetzen

Gestaltung

- Symbole
  - Sprechende Symbole verwenden
  - einfach
  - zeitlos, aber nicht antiquiert
  - Effekt der Wiedererkennung
- Farben
  - Farben systematisch einsetzen
  - Informationsvermittlungsfunktion
  - Differenzierbarkeit
  - Farbwertigkeit: hell oder dunkel – dominant oder zurückhaltend
  - Rote Farbe Achtung!

Interaktive Grafik

Als Kernfunktionen in einem interaktiven grafischen System fordern wir nach Theus/ Urbabnek (Interactive Graphics for Data analysis)

- Abfragen / Queries
  Wir brauchen Methoden um exakte, oder nicht sichtbare, Information in einer Grafik abzufragen
- Selektionen
  Um effiziente Gruppenvergleiche durchzuführen brauchen wir Werkzeuge um Daten auf vielfältige Art und Weise zu selektieren
- Highlighting
  Jede Selektion muss via Linking in alle Repräsentationen der Daten propagiert werden um einen Vergleich zu ermöglichen
- Modifikation von Grafikparametern
  Wir wollen die Eigenschaften von Grafiken schnell und effizient variieren können um immer die optimale Ansicht nutzen zu können

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension
- Parallel- Koordinaten -plots

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension
- Parallel- Koordinaten -plots
- Mosaik Plots

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension
- Parallel- Koordinaten -plots
- Mosaik Plots
- Gemeinsame Darstellung verschiedener Plots, z.B. Scatterplot Matrix

Mehrdimensionale Darstellungen

Möglichkeiten zum Umgang mit Dimension > 2:

- 3-Dimensionale plots
- Höherdimensionale Plots mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension
- Parallel- Koordinaten -plots
- Mosaik Plots
- Gemeinsame Darstellung verschiedener Plots, z.B. Scatterplot Matrix
- interaktive Verbindung von Plots

3D Plots

- 3D - Effekte nur bei tatsächlich 3 Dimensionen
- Verbesserung der Visualisierung durch Netz, Farben und Drehen, Höhenlinien
- Wahrnehmung oft problematisch

Scatterplot Matrix

Die paarweisen Scatterplots werden in Form einer Matrix angeordnet.

Für Übersicht sinnvoll, aber im Kern 2-dimensional
nicht sehr effizient wegen Symmetrie

3. Dimension durch Farbe: Heat Plot

- Bekannt von Landkarten
- Diskret und Stetig möglich
- Bei nominalen Merkmalen: Algorithmen zur Anordnung
- Bei großen Datensätzen geeignet

Parallel- Koordinaten -Plots

Parallel Coordinate Plots (Inselberg, 1985)
Grundidee: Werte von Variablen werden auf parallelen Achsen eingezeichnet und
die Werte zu der gleichen Variablen miteinander verbunden
Vorteile:

- Geeignet für hochdimensionale Daten
- Erster Überblick
- Mit Highlighting können Strukturen sichtbar werden
- Kompakter als z.B. Scatterplot Matrix

Nachteile

- Eindruck abhängig von Anordnung der Variablen aber Sortierung nach verschiedenen Kriterien möglich
- Mehrdimensionale Ausreißer werden nicht erkannt
- Strukturen zunächst schwer interpretierbar

Trellis- Displays (Lattice Graphics)

Gitter (Netz von Plots)
Grundidee: Zeichne gleiche Plots bedingt auf eine oder mehrere (kategorielle) Variable
Beispiel: Boxplots für eine Variable für verschiedene Länder
Mosaik Plot kann so gesehen werden

Plot Ensembles (Menge von Plots)

Verschiedene (eindimensionale Plots, die verbunden sind und interaktiv analysiert werden
Beispiel: Histogramm der Zielvariablen, Boxplots weitere metrischer Größen, Barplots von kategorialen Daten.
Markierung innerhalb einzelner Plots liefert Hinweise auf Strukturen

Fallstudie: Staub Exposition und Chronische Bronchitis

Studie zur Bestimmung von MAK-Werten (Maximale Arbeitsplatzkonzentration) Die Daten wurden in den Jahren 1960 bis 1977 unter den Mitarbeitern einer Münchner Fabrik (1246 Mitarbeiter) erhoben.
Daten sind zugänglich unter
`http://www.stat.uni-muenchen.de/service/datenarchiv/dust/dust.html`

Datenstruktur

4 Merkmale:
CBR: Auftreten von chronischer Bronchitis
1 : Ja
0 : Nein
dust: Staubbelastung am Arbeitsplatz (in mg/m^3)
smoking: Ist der Mitarbeiter Raucher?
1 : Ja
0 : Nein
expo: Dauer der Belastung in Jahren