

Statistik I

Descriptive Statistik & Explorative Datenanalyse

LMU München

Termine

Vorlesung: *Fabian Scheipl*

Montag & Donnerstag 14-16

Übung: *Patrick Schenk*

Beginn am **21.10.**

Montag 10-12

Donnerstag 12-14

Tutorium: *Michael Kobl*

Beginn am **22.10.**

Dienstag 18-20

Korrekturen: *Felix Schweikl, Simon Schreiner*

Formales

- ▶ 2 Module: Vorlesung (6 ECTS) + Übung (3 ECTS)
- ▶ Prüfungsleistung für Statistik PO 2021:
 - ▶ benotete Hausaufgaben während dem Semester (Übungsmodul)
 - ▶ GOP (Statistik 1 + 2, 150 min) im September (Vorlesungsmodul)
- ▶ Prüfungsleistung für alle anderen: Klausur im Februar, kein Übungsmodul
- ▶ Benotete Hausaufgaben (nur HF Statistik PO 2021):
 - ▶ mindestens alle 2 Wochen, markierte Aufgaben auf den Übungsblättern
 - ▶ Abgaben immer bis Montag 10:00 der folgenden Woche über moodle
 - ▶ keine Gruppenarbeit – nur Einzelabgaben
 - ▶ Gesamtpunktzahl über das Semester ergibt Note für das Übungsmodul
- ▶ Klausur (nur NF und alte Statistik PO 2010):
 - ▶ Termin am Ende des Semesters
 - ▶ 90 min
 - ▶ *open book* – alle eigenen Unterlagen & Bücher zugelassen
 - ▶ Zweitklausur Ende März/Anfang April: selber Stoff, selbe Schwierigkeit

Moodle

- ▶ Online-Learning-Plattform der LMU
- ▶ Login mit Ihrer @campus.lmu.de-Kennung
- ▶ **Alle** Teilnehmenden müssen sich in die Veranstaltung einschreiben.
Einschreibeschlüssel ist **dskrpt**
Nur wer eingeschrieben ist kann auf Material zugreifen, Hausaufgaben abgeben, etc.

Moodle URL: moodle.lmu.de/course/view.php?id=35516

Programmierung

Programmpaket R

- ▶ Frei verfügbar, *open source*: r-project.org
- ▶ *lingua franca* der Statistik
 - ▶ (fast) jede denkbare statistische Methode implementiert
 - ▶ aktuelle Forschung geht in Form von Zusatzpaketen (“packages”) ein
 - ▶ immer wichtiger / etablierter auch in Wirtschaft & Verwaltung
- ▶ Obacht, steile Lernkurve: Kommandozeile, kein grafisches *interface* – tippen statt klicken&wischen!
- ▶ Editor **RStudio** (rstudio.com, frei verfügbar)
- ▶ Interaktive R-Kursprogramme zum Selbststudium: [swirl](#), DataCamp, Software Carpentry, Code Academy,

Programmierung

Äußerst wichtige Schwesterveranstaltung für HF Statistik:
Einführung in die statistische Software

⇒ praktische Umsetzung der hier vermittelten Inhalte mit
Programmiersprache R



Wir bieten Ihnen kostenfreie, zeitlich limitierte Datacamp-Accounts fürs Selbststudium.

DataCamp

Sign-Up Link:



Anmeldung mit Ihrer @campus.lmu.de-, @lmu.de oder @stat.uni-muenchen.de-Adresse möglich.

Literatur

L. Meier:

Wahrscheinlichkeitsrechnung und Statistik - Eine Einführung für Verständnis, Intuition und Überblick

Springer-Verlag, 1. Auflage, 2020

L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz:

Statistik - Der Weg zur Datenanalyse

Springer-Verlag, 8. Auflage, 2016

Nutzen Sie auch das “Arbeitsbuch Statistik” mit Übungsaufgaben dazu.

Mehr Lektüre & Material zu speziellen Themen auf der Moodle-Seite und verlinkt auf den Folien.

Inhalt

Einführung

Datenerhebung & Messung

**Wahrscheinlichkeit: Grundlagen &
Definitionen**

Zusammenhangsmaße für diskrete Merkmale

Zufallsvariablen, Verteilungen & Häufigkeiten

Statistische Grafiken

Kennwerte & Verteilungseigenschaften

Wichtige parametrische Verteilungen

Zufallsvektoren & multivariate Verteilungen

Schätzung & Grenzwertsätze

**Zusammenhangsmaße für metrische
Merkmale**

Korrelation & Kausalität

Einführung

Beispiele

Statistik: Was - Wie - Warum

Gegenwart & Zukunft der Statistik

Theorie-Empirie-Statistik

Einführung

Beispiele

Statistik: Was - Wie - Warum

Gegenwart & Zukunft der Statistik

Theorie-Empirie-Statistik

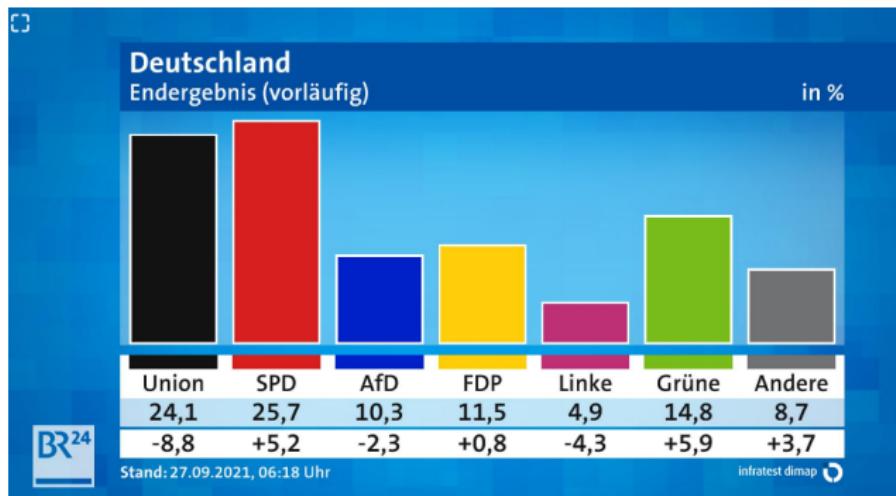
Bundestagswahl 2021

Prognose 18:00 Infratest Dimap (ARD):

Partei:	CDU/CSU	SPD	AfD	FDP	Linke	Grüne
Stimmanteil (%):	25	25	11	11	5	15

Basis: Nachwahlbefragung, ca. 33 000 Wahlberechtigte

Ergebnis:



bundestagswahl.br.de/public/ec/ergebnis-bundestagswahl-2021-bayern-deutschland.html

Bundestagswahl 2021

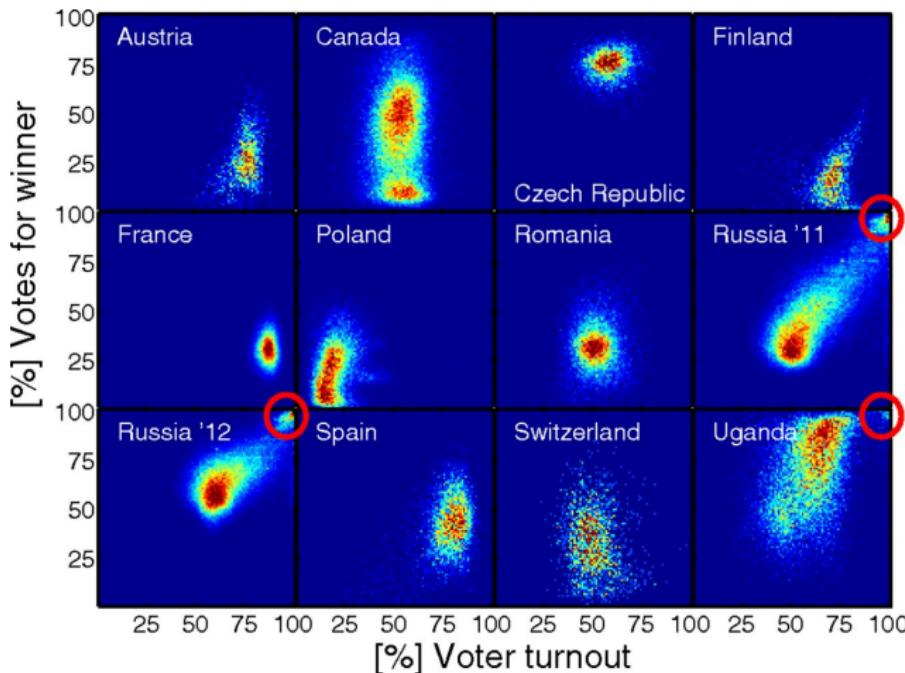
Ziele:

- ▶ Schluss von der Befragung einer Stichprobe von Wähler:innen auf Gesamtergebnis
- ▶ Analyse von Wahlverhalten durch weitere Fragen (Wechselwähler:innen etc.)

Wahlfälschung

Idee: Untersuche Zusammenhang zwischen Wahlergebnis (Stimmenanteil der Sieger) und Wahlbeteiligung.

ballot stuffing treibt beides in die Höhe!



P. Klimek, Y. Yegorov, R. Hanel, S. Thurner (2012). Statistical detection of systematic election irregularities.
PNAS 109(41):16469–16473.

KOALA - Koalitions Analyse

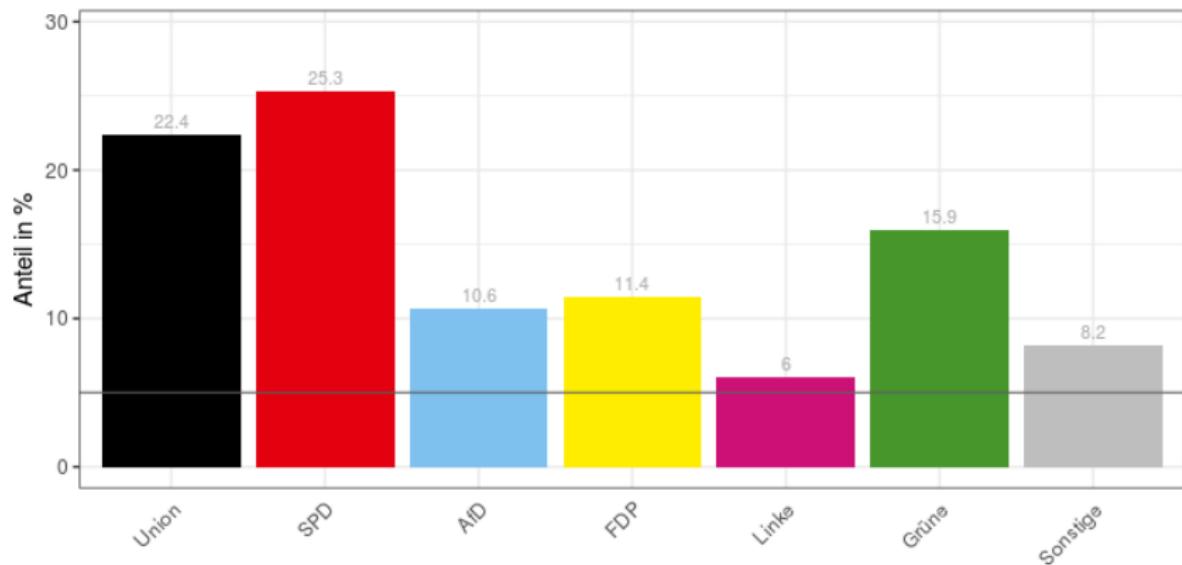
StaBLab-Projekt mit H. Küchenhoff, A. Bender, A. Bauer

- ▶ Ziel: Bessere Vermittlung der den Wahlumfragen zugrunde liegenden Unsicherheiten
- ▶ Kleinere Änderungen in Umfrageanteilen der Parteien vor Wahlen nicht immer relevant, teilweise überbewertet
- ▶ Tatsächlich relevant sind Änderungen in der Wahrscheinlichkeit bestimmter Ereignisse nach Wahlen (Wahrscheinlichkeit für Zustandekommen von Mehrheiten für Koalitionen)
- ▶ Basis: Simulation vieler möglicher Wahlergebnisse auf Basis (mehrerer) aktueller Umfragen
- ▶ Wahrscheinlichkeit: Anteil der Simulationen in welchen Ereignis eintritt

Website: koala.stat.uni-muenchen.de

KOALA - Koalitions Analyse

Gepoolte Umfrage (10.09 - 24.09.2021)

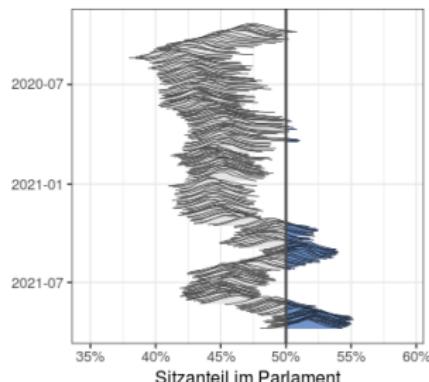
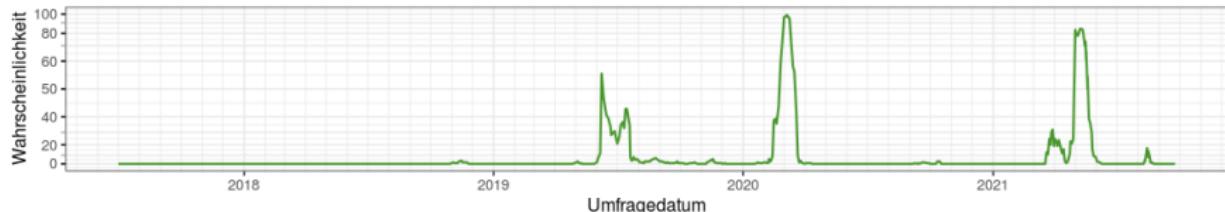


KOALA - Koalitions Analyse

Zeitverlauf bzgl. Koalitionen

Betrachtete Koalitionen

Grüne-SPD-Linke



Sitzmehrheit nein ja

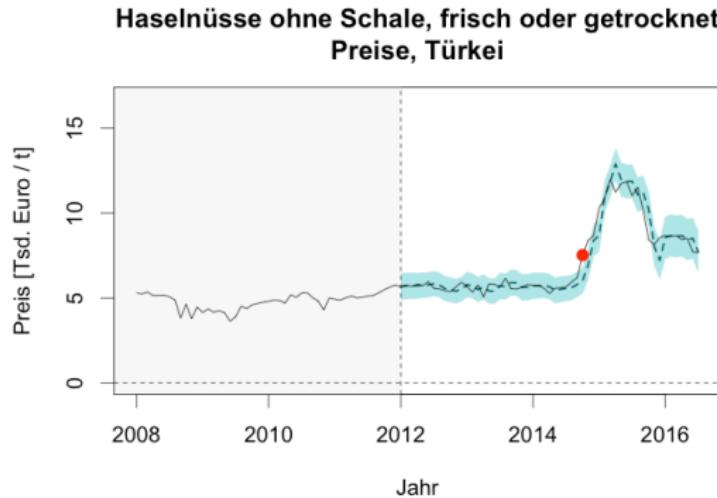
ISAR - Lebensmittelimportscreening

StaBLab-Projekt mit H. Küchenhoff, A. Bauer und F. Günther in Zusammenarbeit mit Bayrischem Landesamt für Gesundheit und Lebensmittelsicherheit, sowie Bundesamt für Verbraucherschutz und Lebensmittelsicherheit

- ▶ Ziel: Früherkennung von Risiken im Lebensmittelhandel
- ▶ Ansatz: Dauerhaftes Screening von produkt- und länderspezifischen Daten aller deutscher Lebensmittelimporte
- ▶ Auffälligkeiten in Mengen oder Preisen können Hinweise zu Risiken und Betrugspotentialen liefern
- ▶ Ansatz: Verwendung statistischer Zeitreihen-Modelle
 - ▶ Vergleich der Beobachtungen mit Vorhersagen der Modelle
 - ▶ Definition von Auffälligkeiten von Interesse
 - ▶ Jeden Monat erhalten Lebensmittelkontrolleure Liste mit auffälligen Zeitreihen
- ▶ Programmierung von Tool zur interaktiven Darstellung der Daten in enger Zusammenarbeit mit Anwendern

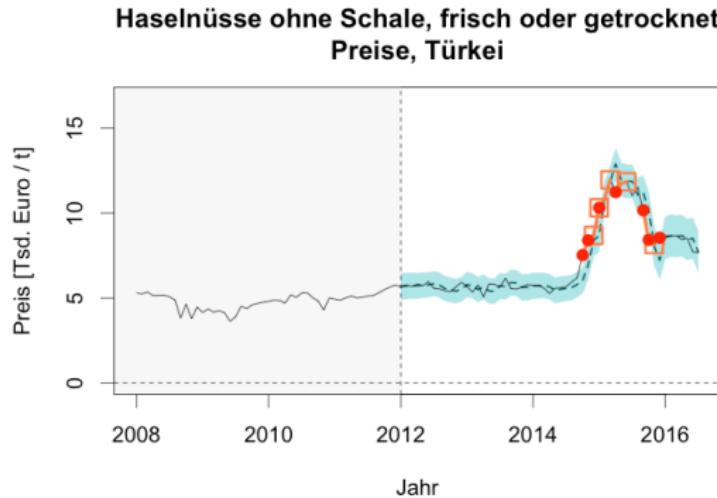
ISAR: Beispiel Haselnüsse Türkei

- ▶ 2014 Ernteeinbruch von Haselnüssen in der Türkei
- ▶ Darauf folgend starker Preisanstieg
- ▶ Es wurden verstärkt Kontrollen der nach Deutschland importierten Haselnussprodukte durchgeführt
- ▶ In verarbeiteten Produkten wurde Verfälschungen, u.a. durch Cashews/Mandeln gefunden, hohes Gesundheitsrisiko (Allergie)
- ▶ In ISAR erste Signale für auffällige Preisentwicklung im September 2015



ISAR: Beispiel Haselnüsse Türkei

- ▶ 2014 Ernteeinbruch von Haselnüssen in der Türkei
- ▶ Darauf folgend starker Preisanstieg
- ▶ Es wurden verstärkt Kontrollen der nach Deutschland importierten Haselnussprodukte durchgeführt
- ▶ In verarbeiteten Produkten wurde Verfälschungen, u.a. durch Cashews/Mandeln gefunden, hohes Gesundheitsrisiko (Allergie)
- ▶ In ISAR erste Signale für auffällige Preisentwicklung im September 2015



Mineralwasserstudie

Studie in Zusammenarbeit mit Prof. O. 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 → 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

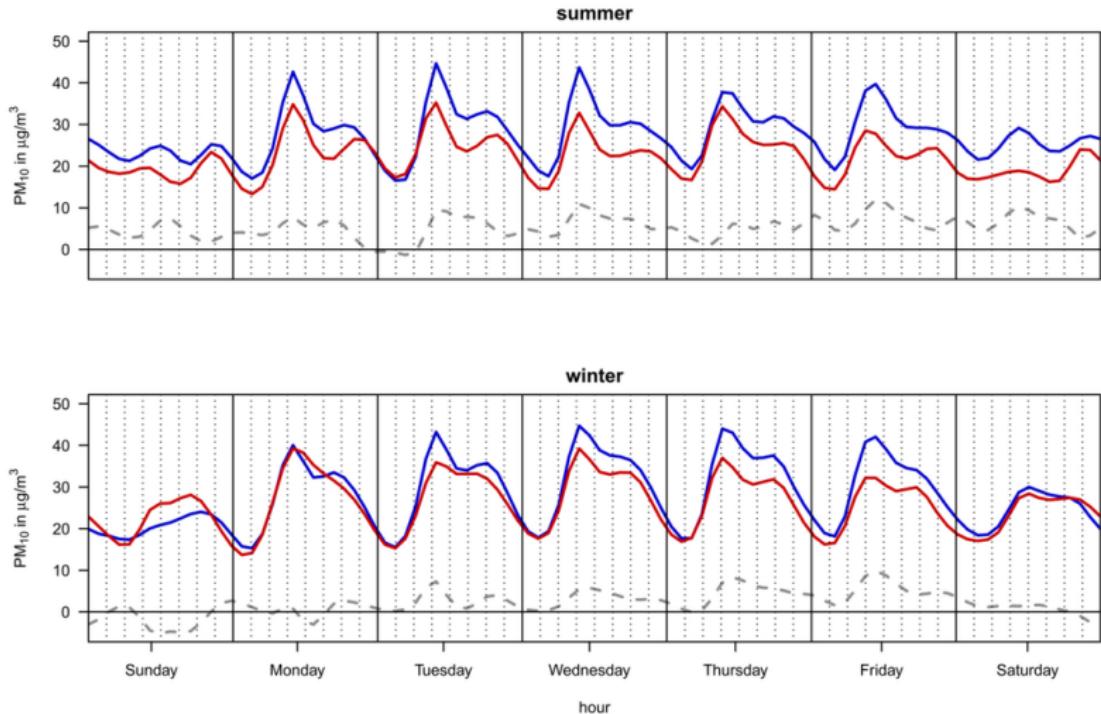
Probleme:

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

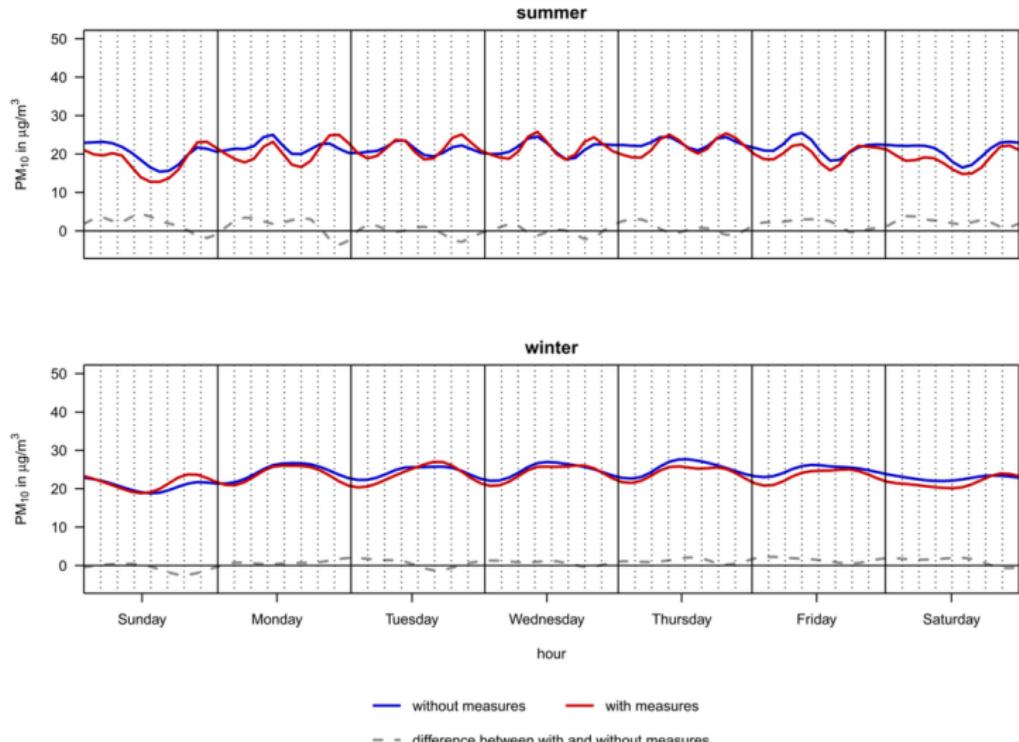
Daher: Regressionsmodell mit Referenzstation, Wetter, Tagesverlauf

V. Fensterer, H. Küchenhoff, V. Maier, H.-E. Wichmann, S. Breitner, A. Peters, J. Gu, and J. Cyrys (2014). 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.

Wirkung der Umweltzone: Prinzregentenstrasse



Wirkung der Umweltzone: Lothstrasse



Weitere Beispiele

- ▶ Klinische Studien
- ▶ Epidemiologische Studien
- ▶ Qualitätskontrolle
- ▶ Marktforschung
 - ▶ Einschaltquoten
 - ▶ Bewertung und Vergleich von Produkten gleichen Typs aber verschiedener Produzenten durch Verbraucher (Waschmittel, Kaffee, Schokolade, usw.)
 - ▶ Online-Tracking-Daten: Cookies, Userprofile, Websitenutzung, ...
 - ▶ A-B-Testing von Websitedesigns, User Interfaces
- ▶ Sportstatistik
- ▶ Analyse von Genexpressions- oder -sequenzdaten
- ▶ Netzwerkanalysen
- ▶ Mustererkennung (“‘Pattern recognition’”): Spamfilter, Customer Churn, Kundensegmentierung, ...

Einführung

Beispiele

Statistik: Was - Wie - Warum

Gegenwart & Zukunft der Statistik

Theorie-Empirie-Statistik

Was ist 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:
Beschreibung, Zusammenfassung, Visualisierung von Daten.
Techniken: Grafiken, Tabellen, Kennzahlen
- ▶ **Explorative** Statistik:
Suche nach Strukturen in Daten. Benutzt deskriptive und induktive Techniken interaktiv und iterativ.
- ▶ **Induktive** Statistik: Schließt von beobachteten Daten auf zugrundeliegende Strukturen.
Angewandte Wahrscheinlichkeitsmathematik.
Techniken: Statistische Modellierung, statistische Tests.

Deskriptive Statistik

- ▶ *Data is merely the raw material of knowledge.*
Charles Wheelan (* 1966, Autor)

Ziel: Wahrhaftige Beschreibung von Daten mit möglichst geringem Informationsverlust

- ▶ Eigenschaften und Strukturen erkennbar machen
- ▶ Grafisch und durch Kennwerte
- ▶ Eindimensional und mehrdimensional (“uni-/multivariat”)
- ▶ keine Rückschlüsse von beobachteter Stichprobe auf die Grundgesamtheit oder allgemeine Phänomene (nur *Deskription*, nicht *Inferenz*)

Was ist Statistik?

- ▶ *Maths is a language that you use to describe statistics, but really it's about collecting information and putting it in an order that makes sense.*
Lauren Stamile (* 1976, Schauspielerin)
- ▶ *Statistics is the grammar of science.*
Karl Pearson (1857 - 1936, Mathematiker)
- ▶ *Statistics is the science of learning from experience.*
Bradley Efron (* 1938, Statistiker)
- ▶ ⇒ **Statistische Methodik als unabdingbares Werkzeug jeder empirischen Wissenschaft.**

Was ist Statistik?

- ▶ *Maths is a language that you use to describe statistics, but really it's about collecting information and putting it in an order that makes sense.*
Lauren Stamile (* 1976, Schauspielerin)
- ▶ *Statistics is the grammar of science.*
Karl Pearson (1857 - 1936, Mathematiker)
- ▶ *Statistics is the science of learning from experience.*
Bradley Efron (* 1938, Statistiker)
- ▶ ⇒ **Statistische Methodik als unabdingbares Werkzeug jeder empirischen Wissenschaft.**

Was ist Statistik?

- ▶ *Maths is a language that you use to describe statistics, but really it's about collecting information and putting it in an order that makes sense.*
Lauren Stamile (* 1976, Schauspielerin)
- ▶ *Statistics is the grammar of science.*
Karl Pearson (1857 - 1936, Mathematiker)
- ▶ *Statistics is the science of learning from experience.*
Bradley Efron (* 1938, Statistiker)
- ▶ ⇒ **Statistische Methodik als unabdingbares Werkzeug jeder empirischen Wissenschaft.**

Was ist Statistik?

- ▶ *Maths is a language that you use to describe statistics, but really it's about collecting information and putting it in an order that makes sense.*
Lauren Stamile (* 1976, Schauspielerin)
- ▶ *Statistics is the grammar of science.*
Karl Pearson (1857 - 1936, Mathematiker)
- ▶ *Statistics is the science of learning from experience.*
Bradley Efron (* 1938, Statistiker)
- ▶ **⇒ Statistische Methodik als unabdingbares Werkzeug jeder empirischen Wissenschaft.**

Warum Statistik?

- ▶ *Statistics is a body of methods for making wise decisions in the face of uncertainty.*
W. Allen Wallis (1912-1998, Statistiker & Ökonom)
- ▶ *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*
Steven Pinker (* 1954, Psychologe & Linguist)
- ▶ *It is the mark of a truly intelligent person to be moved by statistics.*
George Bernard Shaw (1856 - 1950, Dramatiker)
- ▶ *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.*
Elisabeth Noelle-Neumann (1916-2010, Demoskopin)
- ▶ ⇒ **Statistische (Aus-)Bildung um irrationales Handeln zu vermeiden und mit Unsicherheit vernünftig umzugehen.**

Warum Statistik?

- ▶ *Statistics is a body of methods for making wise decisions in the face of uncertainty.*
W. Allen Wallis (1912-1998, Statistiker & Ökonom)
- ▶ *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*
Steven Pinker (* 1954, Psychologe & Linguist)
- ▶ *It is the mark of a truly intelligent person to be moved by statistics.*
George Bernard Shaw (1856 - 1950, Dramatiker)
- ▶ *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.*
Elisabeth Noelle-Neumann (1916-2010, Demoskopin)
- ▶ ⇒ **Statistische (Aus-)Bildung um irrationales Handeln zu vermeiden und mit Unsicherheit vernünftig umzugehen.**

Warum Statistik?

- ▶ *Statistics is a body of methods for making wise decisions in the face of uncertainty.*
W. Allen Wallis (1912-1998, Statistiker & Ökonom)
- ▶ *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*
Steven Pinker (* 1954, Psychologe & Linguist)
- ▶ *It is the mark of a truly intelligent person to be moved by statistics.*
George Bernard Shaw (1856 - 1950, Dramatiker)
- ▶ *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.*
Elisabeth Noelle-Neumann (1916-2010, Demoskopin)
- ▶ ⇒ **Statistische (Aus-)Bildung um irrationales Handeln zu vermeiden und mit Unsicherheit vernünftig umzugehen.**

Warum Statistik?

- ▶ *Statistics is a body of methods for making wise decisions in the face of uncertainty.*
W. Allen Wallis (1912-1998, Statistiker & Ökonom)
- ▶ *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*
Steven Pinker (* 1954, Psychologe & Linguist)
- ▶ *It is the mark of a truly intelligent person to be moved by statistics.*
George Bernard Shaw (1856 - 1950, Dramatiker)
- ▶ *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.*
Elisabeth Noelle-Neumann (1916-2010, Demoskopin)
- ▶ ⇒ **Statistische (Aus-)Bildung um irrationales Handeln zu vermeiden und mit Unsicherheit vernünftig umzugehen.**

Warum Statistik?

- ▶ *Statistics is a body of methods for making wise decisions in the face of uncertainty.*
W. Allen Wallis (1912-1998, Statistiker & Ökonom)
- ▶ *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*
Steven Pinker (* 1954, Psychologe & Linguist)
- ▶ *It is the mark of a truly intelligent person to be moved by statistics.*
George Bernard Shaw (1856 - 1950, Dramatiker)
- ▶ *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.*
Elisabeth Noelle-Neumann (1916-2010, Demoskopin)
- ▶ **⇒ Statistische (Aus-)Bildung um irrationales Handeln zu vermeiden und mit Unsicherheit vernünftig umzugehen.**

Warum (nicht) Statistik?

- ▶ *All models are wrong, but some are useful.*
George E. P. Box (1919 - 2013, Statistiker)
- ▶ *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*
John Tukey (1915- 2000, Mathematiker)
- ▶ *If you torture the data enough, nature will always confess.*
Ronald Coase (1910 - 2013, Ökonom)
- ▶ *There are no routine statistical questions, only questionable statistical routines.*
David R. Cox (1924 - 2022, Statistiker)
- ▶ → **Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

Warum (nicht) Statistik?

- ▶ *All models are wrong, but some are useful.*
George E. P. Box (1919 - 2013, Statistiker)
- ▶ *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*
John Tukey (1915- 2000, Mathematiker)
- ▶ *If you torture the data enough, nature will always confess.*
Ronald Coase (1910 - 2013, Ökonom)
- ▶ *There are no routine statistical questions, only questionable statistical routines.*
David R. Cox (1924 - 2022, Statistiker)
- ▶ → **Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

Warum (nicht) Statistik?

- ▶ *All models are wrong, but some are useful.*
George E. P. Box (1919 - 2013, Statistiker)
- ▶ *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*
John Tukey (1915- 2000, Mathematiker)
- ▶ *If you torture the data enough, nature will always confess.*
Ronald Coase (1910 - 2013, Ökonom)
- ▶ *There are no routine statistical questions, only questionable statistical routines.*
David R. Cox (1924 - 2022, Statistiker)
- ▶ → **Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

Warum (nicht) Statistik?

- ▶ *All models are wrong, but some are useful.*
George E. P. Box (1919 - 2013, Statistiker)
- ▶ *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*
John Tukey (1915- 2000, Mathematiker)
- ▶ *If you torture the data enough, nature will always confess.*
Ronald Coase (1910 - 2013, Ökonom)
- ▶ *There are no routine statistical questions, only questionable statistical routines.*
David R. Cox (1924 - 2022, Statistiker)
- ▶ → **Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

Warum (nicht) Statistik?

- ▶ *All models are wrong, but some are useful.*
George E. P. Box (1919 - 2013, Statistiker)
- ▶ *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*
John Tukey (1915- 2000, Mathematiker)
- ▶ *If you torture the data enough, nature will always confess.*
Ronald Coase (1910 - 2013, Ökonom)
- ▶ *There are no routine statistical questions, only questionable statistical routines.*
David R. Cox (1924 - 2022, Statistiker)
- ▶ **⇒ Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

Warum nicht Statistik?

- ▶ *Not everything that counts can be counted, and not everything that can be counted counts.*

William Bruce Cameron (1904 - 1988, Soziologe)

- ▶ *Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?*

Dennis Prager (* 1948, Radiomoderator & Publizist)

- ▶ *Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.*

Hilaire Belloc (1870 - 1953, Schriftsteller & Historiker)

- ▶ ⇒ **Statistik nur dort sinnvoll, wo Messbarkeit und Quantifizierbarkeit gegeben sind.**

Warum nicht Statistik?

- ▶ *Not everything that counts can be counted, and not everything that can be counted counts.*

William Bruce Cameron (1904 - 1988, Soziologe)

- ▶ *Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?*

Dennis Prager (* 1948, Radiomoderator & Publizist)

- ▶ *Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.*

Hilaire Belloc (1870 - 1953, Schriftsteller & Historiker)

- ▶ ⇒ **Statistik nur dort sinnvoll, wo Messbarkeit und Quantifizierbarkeit gegeben sind.**

Warum nicht Statistik?

- ▶ *Not everything that counts can be counted, and not everything that can be counted counts.*

William Bruce Cameron (1904 - 1988, Soziologe)

- ▶ *Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?*

Dennis Prager (* 1948, Radiomoderator & Publizist)

- ▶ *Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.*

Hilaire Belloc (1870 - 1953, Schriftsteller & Historiker)

- ▶ ⇒ **Statistik nur dort sinnvoll, wo Messbarkeit und Quantifizierbarkeit gegeben sind.**

Warum nicht Statistik?

- ▶ *Not everything that counts can be counted, and not everything that can be counted counts.*
William Bruce Cameron (1904 - 1988, Soziologe)
- ▶ *Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?*
Dennis Prager (* 1948, Radiomoderator & Publizist)
- ▶ *Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.*
Hilaire Belloc (1870 - 1953, Schriftsteller & Historiker)
- ▶ **⇒ Statistik nur dort sinnvoll, wo Messbarkeit und Quantifizierbarkeit gegeben sind.**

... und außerdem:

- ▶ Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.
Goodhart's Law, auch:
“When a measure becomes a target, it ceases to be a good measure.”
oder
“The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures.”
- ▶ Ähnlich auch: Präventions-Paradox, self-fulfilling/-destroying prophecies
- ▶ → Verlässliche Messungen & Vorhersagen in Systemen, die selbst Messergebnissen beeinflussen können & von Vorhersagen beeinflusst werden, sind problematisch.

... und außerdem:

- ▶ Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.
Goodhart's Law, auch:
“When a measure becomes a target, it ceases to be a good measure.”
oder
“The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures.”
- ▶ Ähnlich auch: Präventions-Paradox, self-fulfilling/-destroying prophecies
- ▶ **⇒ Verlässliche Messungen & Vorhersagen in Systemen, die selbst Messergebnissen beeinflussen können & von Vorhersagen beeinflusst werden, sind problematisch.**

Einführung

Beispiele

Statistik: Was - Wie - Warum

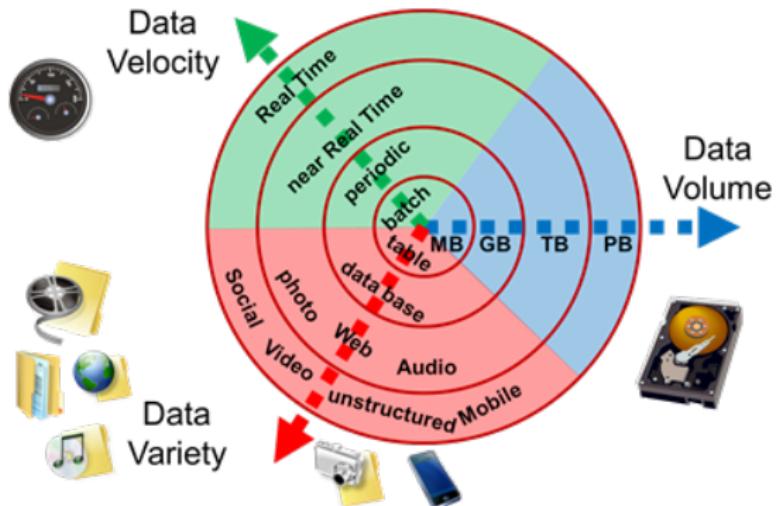
Gegenwart & Zukunft der Statistik

Theorie-Empirie-Statistik

“Big Data”

- ▶ Analyse und Verarbeitung großer Datenmengen
- ▶ Drei Vs: Volume, Velocity, Variety
- ▶ Häufig rein heuristische/algorithmmische Herangehensweise ohne probabilistische Modelle oder kausale Theorien
- ▶ Große Herausforderungen für Statistik und Informatik

“Big Data”



Big Data, Machine Learning, KI

- ▶ Induktive Statistik umfasst "Machine Learning" & heutige "Künstliche Intelligenz" :≈ Methoden für Mustererkennung & Vorhersagen basierend auf Daten
- ▶ Alltagserfahrungen zunehmend durch datengetriebene Algorithmen bestimmt – Beispiele:
 - ▶ Medien: "filter bubble", "engagement maximization" durch Empfehlungsalgorithmen
 - ▶ Finanziell: z.B. werden verfügbare Zahlungsmethoden bei Online-Käufen durch Creditscores bestimmt
- ▶ Entscheidungen großer Tragweite zunehmend durch datengetriebene Algorithmen – Beispiele:
 - ▶ Kreditvergabe via Creditscoreing
 - ▶ Risikoscoreing der Corona-Warn-App
 - ▶ Medizinische Diagnostik und Therapievorschläge (z.B *Watson for Oncology*)
 - ▶ "Predictive Policing"
 - ▶ Haftentlassung auf Bewährung - z.B. Berk, R. (2017) *Impact assessment of machine learning risk forecasts on parole board decisions and recidivism.*
 - ▶ Militärische KI: US Algorithmic Warfare Initiative, Project Maven,

Big Data & ML: Politische & soziale Folgen

- ▶ *Big data knows and can deduce more about you than Big Brother ever could.*
Toomas Hendrik Ilves (*1953, Diplomat & Politiker)
- ▶ *Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say. [...] Privacy isn't about something to hide. Privacy is about something to protect. And that's who you are. That's what you believe in. That's who you want to become. Privacy is the right to the self. Privacy is what gives you the ability to share with the world who you are on your own terms.*
Edward Snowden (* 1983, Dissident)
- ▶ *No system of mass surveillance has existed in any society, that we know of to this point, that has not been abused.*
Edward Snowden (* 1983, Dissident)
- ▶ ⇒ **Bedeutet vollautomatische, allgegenwärtige, pausenlose Datenerfassung das Ende der Privatsphäre?**

Big Data & ML: Politische & soziale Folgen

- ▶ *Big data knows and can deduce more about you than Big Brother ever could.*
Toomas Hendrik Ilves (*1953, Diplomat & Politiker)
- ▶ *Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say. [...]*
Privacy isn't about something to hide. Privacy is about something to protect. And that's who you are. That's what you believe in. That's who you want to become. Privacy is the right to the self. Privacy is what gives you the ability to share with the world who you are on your own terms.
Edward Snowden (* 1983, Dissident)
- ▶ *No system of mass surveillance has existed in any society, that we know of to this point, that has not been abused.*
Edward Snowden (* 1983, Dissident)
- ▶ ⇒ **Bedeutet vollautomatische, allgegenwärtige, pausenlose Datenerfassung das Ende der Privatsphäre?**

Big Data & ML: Politische & soziale Folgen

- ▶ *Big data knows and can deduce more about you than Big Brother ever could.*
Toomas Hendrik Ilves (*1953, Diplomat & Politiker)
- ▶ *Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say. [...]*
Privacy isn't about something to hide. Privacy is about something to protect. And that's who you are. That's what you believe in. That's who you want to become. Privacy is the right to the self. Privacy is what gives you the ability to share with the world who you are on your own terms.
Edward Snowden (* 1983, Dissident)
- ▶ *No system of mass surveillance has existed in any society, that we know of to this point, that has not been abused.*
Edward Snowden (* 1983, Dissident)
- ▶ → *Bedeutet vollautomatische, allgegenwärtige, pausenlose Datenerfassung das Ende der Privatsphäre?*

Big Data & ML: Politische & soziale Folgen

- ▶ *Big data knows and can deduce more about you than Big Brother ever could.*
Toomas Hendrik Ilves (*1953, Diplomat & Politiker)
- ▶ *Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say. [...]*
Privacy isn't about something to hide. Privacy is about something to protect. And that's who you are. That's what you believe in. That's who you want to become. Privacy is the right to the self. Privacy is what gives you the ability to share with the world who you are on your own terms.
Edward Snowden (* 1983, Dissident)
- ▶ *No system of mass surveillance has existed in any society, that we know of to this point, that has not been abused.*
Edward Snowden (* 1983, Dissident)
- ▶ **⇒ Bedeutet vollautomatische, allgegenwärtige, pausenlose Datenerfassung das Ende der Privatsphäre?**

Big Data & ML: Politische & soziale Folgen

- ▶ **Predictive models are, increasingly, the tools we will be relying on to run our institutions, deploy our resources, and manage our lives.** But [...] these models are constructed not just from data but from the choices we make about which data to pay attention to – and which to leave out. Those choices are not just about logistics, profits, and efficiency. They are fundamentally moral. If we back away from them and treat mathematical models as a neutral and inevitable force, like the weather or the tides, we abdicate our responsibility. We must come together to police [...], tame and disarm them. My hope is that they'll be remembered, like the deadly coal mines of a century ago, as relics of the early days of this new revolution, before we learned how to bring fairness and accountability to the age of data.
Cathy O'Neil (Statistikerin, Zitat aus "Weapons of Math Destruction" (Crown, 2016))

Big Data & ML: Politische & soziale Folgen

- ▶ [... M]any of these models encode human prejudice, misunderstanding, and bias into the software systems that increasingly manage our lives. Like gods, these mathematical models are opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, are beyond dispute or appeal. And they tend to punish the poor and the oppressed in our society, while making the rich richer.
Cathy O'Neil (Statistikerin, Zitat aus "Weapons of Math Destruction" (Crown, 2016))
- ▶ Hidden biases in both the collection and analysis stages present considerable risks and are as important to the big-data equation as the numbers themselves. [...] The fear isn't that big data discriminates. We already know that it does. It's that you don't know if you've been discriminated against.
Kate Crawford (Informatikerin)
- ▶ ⇒ Große Herausforderung der nächsten Jahrzehnte: Humane & faire Verwirklichung der positiven Potentiale, Vermeidung der dystopischen Aspekte.

Big Data & ML: Politische & soziale Folgen

- ▶ [... M]any of these models encode human prejudice, misunderstanding, and bias into the software systems that increasingly manage our lives. Like gods, these mathematical models are opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, are beyond dispute or appeal. And they tend to punish the poor and the oppressed in our society, while making the rich richer.
Cathy O’Neil (Statistikerin, Zitat aus “Weapons of Math Destruction” (Crown, 2016))
- ▶ Hidden biases in both the collection and analysis stages present considerable risks and are as important to the big-data equation as the numbers themselves. [...] The fear isn’t that big data discriminates. We already know that it does. It’s that you don’t know if you’ve been discriminated against.
Kate Crawford (Informatikerin)
- ▶ ⇒ Große Herausforderung der nächsten Jahrzehnte: Humane & faire Verwirklichung der positiven Potentiale, Vermeidung der dystopischen Aspekte.

Big Data & ML: Politische & soziale Folgen

- ▶ [... M]any of these models encode human prejudice, misunderstanding, and bias into the software systems that increasingly manage our lives. Like gods, these mathematical models are opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, are beyond dispute or appeal. And they tend to punish the poor and the oppressed in our society, while making the rich richer.
Cathy O'Neil (Statistikerin, Zitat aus "Weapons of Math Destruction" (Crown, 2016))
- ▶ Hidden biases in both the collection and analysis stages present considerable risks and are as important to the big-data equation as the numbers themselves. [...] The fear isn't that big data discriminates. We already know that it does. It's that you don't know if you've been discriminated against.
Kate Crawford (Informatikerin)
- ▶ ⇒ Große Herausforderung der nächsten Jahrzehnte: Humane & faire Verwirklichung der positiven Potentiale, Vermeidung der dystopischen Aspekte.

Gegenwart & Zukunft der Statistik

- ▶ Enorm gewachsener Bedarf durch Allgegenwart automatisiert erhobener Daten & Digitalisierung
- ▶ Falsch/missbräuchlich angewendete Statistik führt zu Reproduzierbarkeitskrisen in (Sozial-)psychologie, vielen Bereichen der Medizin, Ernährungswissenschaften
Gegenwärtiger Paradigmenwechsel in der angewandten Forschung
(Stichworte: *p-hacking, replication crisis*)
- ▶ “Künstliche Intelligenz” / Maschinelles Lernen & Big Data:
Hochkomplexe, extrem rechenintensive Algorithmen in Kombination mit riesigen Datenmengen ermöglichen ganz neue Anwendungen von Datenanalyse
Beitrag der Statistik zur Beherrschbarmachung / Interpretation / Risikoabschätzung dieser Techniken?

Einführung

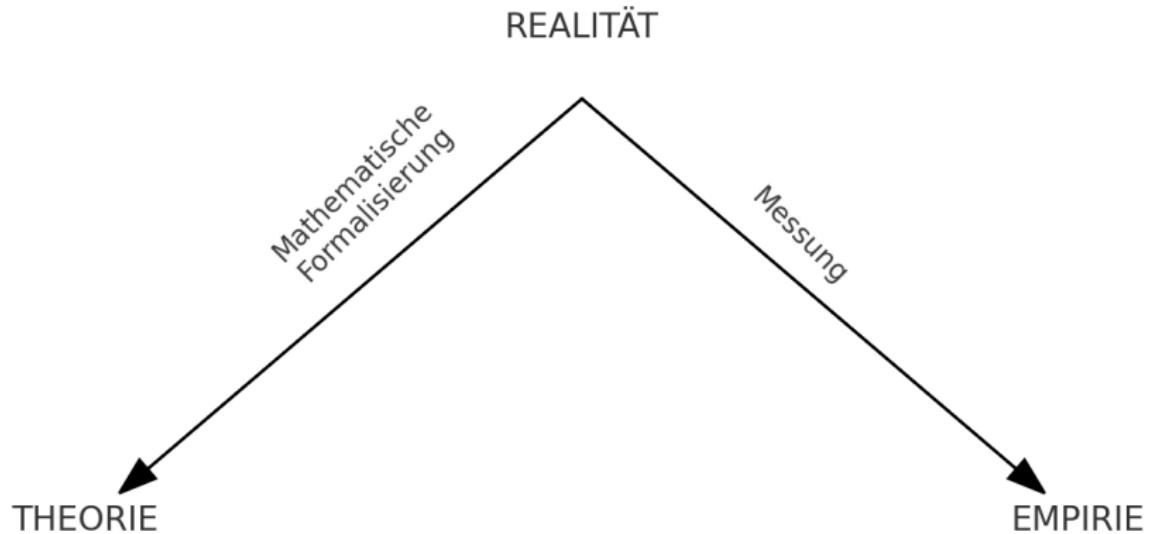
Beispiele

Statistik: Was - Wie - Warum

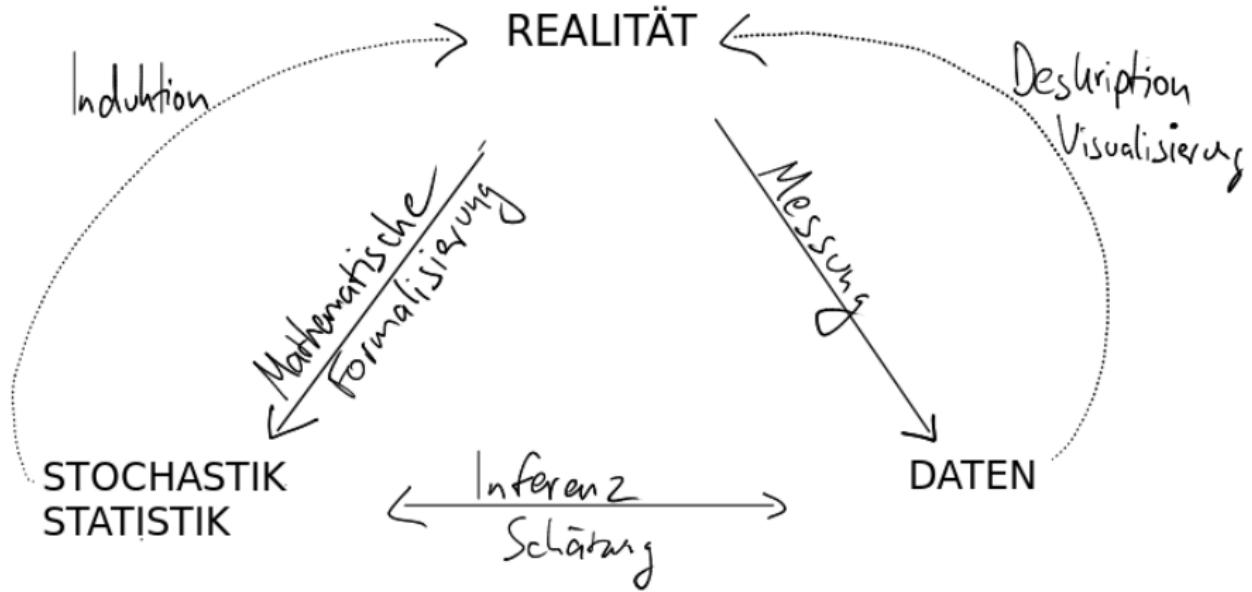
Gegenwart & Zukunft der Statistik

Theorie-Empirie-Statistik

Theorie & Empirie



Theorie & Empirie



Theorie & Empirie

- ▶ Modellbildung aus mathemat. Formalisierung liefert vereinfachte, grobe Abbildung eines Ausschnitts der Realität in Symbolen.
- ▶ Daten aus Messung liefern vereinfachte, grobe Abbildung eines Ausschnitts der Realität in Zahlen.
- ▶ Deskriptive Statistik fasst Aspekte der gemessenen Daten in Kennzahlen, Grafiken, Tabellen für Menschen lesbar zusammen.
- ▶ Statistische Inferenz (Schätzungen & Tests) liefert auf Basis von Daten quantitative Aussagen über Modell(komponenten).
Obacht: Modell ≠ Realität!
- ▶ Messung selbst bzw. Definition von Messverfahren oft “modellbasiert”.
- ▶ Expertise für substantielle Fragestellung *und* Mathe nötig um Realitätsnähe & Tauglichkeit einer mathematischen Formalisierung beurteilen zu können.
⇒ sinnvolle angewandte Statistik ist **immer interdisziplinär!**

Theorie & Empirie

Mathematischer Formalismus

Wahrscheinlichkeit

verursacht(?)
→

Wahrscheinlichkeit

←
repräsentiert(?)

Zufallsvariable

entspricht
↔

Erwartungswert

←
Schätzung für

Stochastische Grundgesamtheit Ω

↔
?

Empirie / Daten

Beobachtete Häufigkeit

Beobachtete Häufigkeit

Statistisches Merkmal

Arithmetisches Mittel

Statistische Grundgesamtheit

usw.

Datenerhebung & Messung

Messung

Skalenniveaus

Datenerhebung

Grundbegriffe

Def.: Statistische Einheit, Untersuchungseinheit (UE)

Objekte, an denen interessierende Größen erfasst werden.
“Merksalsträger”.

Def.: Grundgesamtheit, Population (GG)

Menge aller für eine Fragestellung relevanten statistischen Einheiten

Def.: Teilgesamtheit, Stichprobe

Eine Teilmenge der GG, speziell: die Teilmenge an UE, für die Daten vorliegen.

Grundbegriffe

Def.: Merkmal, (Ko-)Variable

Eine messbare Eigenschaft der Untersuchungseinheiten

Def.: Merkmalsausprägung

Konkreter Wert eines Merkmals für eine bestimmte UE.

Def.: Merkmalsraum (Zustandsraum)

Menge aller *möglichen* Merkmalsausprägungen eines Merkmals

Beachte: *beobachtete* Ausprägungen \subseteq Merkmalsraum

Def.: Beobachtung

Die Gesamtheit der beobachteten Merkmalsausprägungen der gemessenen Merkmale einer Untersuchungseinheit zu einem bestimmten Messzeitpunkt.

Datenerhebung & Messung

Messung

Skalenniveaus

Datenerhebung

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 to rules.*
Stanley S. Stevens (1951)

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 to rules.*
Stanley S. Stevens (1951)

Messen

Messen bedeutet: Zuordnung von Zahlen/Symbolen zu Ausprägungen von Merkmalen an Untersuchungseinheiten.

- ▶ Physikalische Messungen:
Länge, Blutdruck, Temperatur, ...
- ▶ Psychometrische Beispiele:
Gewaltbereitschaft, Schwere der Depression, ...
- ▶ Wirtschaftswissenschaftliche Beispiele:
Inflation, Bruttonsozialprodukt, Arbeitslosenquote, ...

Messen

Amir → 1.84

Liz → 1.61

Feihong → 1.72

Jedes Merkmal definiert bestimmte **Relationen** zwischen den UE

Hier z.B.: "Amir größer als Feihong größer als Liz"

Jede gültige Messung ist eine **strukturerhaltende Abbildung**

(Homomorphismus) bezüglich dieser Relationen.

Hier z.B.: "Liz ist kleiner als Amir" \iff "1.61 < 1.84"

Typen von Messungen

1. Messung hat reales (physikalisches) Relativ;
direkte Messung ("representational measurement").
z.B. Länge, Gewicht, Anzahl, Blutzuckerkonzentration
2. Messung besitzt durch *Operationalisierung* definiertes Relativ;
indirekte/operationale Messung ("pragmatic measurement").
z.B. Intelligenz, Schweregrad einer Krankheit, Arbeitslosenquote

Operationalisierung bedeutet "Messbarmachung", definiert genaue
Messvorschrift, Messinstrumente, etc.

Datenerhebung & Messung

Messung

Skalenniveaus

Datenerhebung

Skalen

Messungen als strukturerhaltende Abbildungen
also: empirisches Relativ \cong numerisches Relativ

Existenz:

Ist die Struktur der Objekte so, dass eine strukturerhaltende Abbildung existiert?

- \Rightarrow Axiome von Repräsentationstheoremen müssen erfüllt sein (z.B. Transitivität)
- \Rightarrow Verletzt z.B. oft durch unerfüllte Annahme von Eindimensionalität

Eindeutigkeit:

Gibt es mehrere zulässige Skalen?
(z.B. Fläche in km^2 & ha ; Temperatur in $^\circ\text{C}$, $^\circ\text{F}$ & $^\circ\text{K}$)

- \Rightarrow gibt es zulässige (strukturerhaltende) Transformationen?

Transformationen und Operationen

Def.: Transformation einer Skala

Funktion, die Ausprägungen eines Merkmals auf neue Ausprägungen abbildet.

in etwa: "Skalenwechsel", z.B:

- Umwandlung in andere physikalische Einheit (Temperatur: $^{\circ}\text{F} \rightarrow ^{\circ}\text{C}$)
- Relabeling (Beruf: "Putzkraft" \rightarrow "Raumpfleger*in")
- Gruppierung (Haarfarbe: {"braun", "schwarz"} \rightarrow {"dunkel"})

Def.: Operation auf einer Skala

Funktion, die Ausprägungen eines Merkmals auf der selben Skala miteinander in Bezug setzt.

z.B " $= / \neq$ ", " $> / = / <$ ", Differenzen, Quotienten etc.

Skalenniveaus

Zulässige Skalentransformationen erhalten die Struktur der empirischen Relative, die durch Messungen repräsentiert wird.

Die Menge zulässiger Transformationen bestimmt das Skalenniveau eines Merkmals.

Höhere Skalenniveaus erlauben eine *größere* Menge von sinnvollen *Operationen auf Werten der Skala*, aber eine *kleinere* Menge an zulässigen strukturerhaltenden *Transformationen der Skala* an sich.

Nominalskala

- ▶ Beispiele: Religionszugehörigkeit, Wohnort, Sockenfarbe
- ▶ Struktur: keine
- ▶ Sinnvolle Operationen: gleich/ungleich
- ▶ Erlaubte Transformationen: alle eineindeutigen Abbildungen f , da für sie gilt:

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

Ordinal- oder Rangskala

- ▶ Beispiele: Schulbildung, soziale Schicht, Schweregrad einer Erkrankung
- ▶ Struktur: **lineare Ordnung**
- ▶ Sinnvolle Operationen: gleich/ungleich, größer/kleiner
- ▶ Erlaubte Transformationen: alle streng monoton steigenden Abbildungen f , da für sie gilt:

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

Intervallskala

- ▶ Beispiele: Temperatur, Jahreszahlen, IQ
- ▶ Struktur: **Abstände** quantifizierbar
- ▶ Sinnvolle Operationen: gleich/ungleich, größer/kleiner,
Differenzbildung
- ▶ Erlaubte Transformationen: alle linearen Transformationen
 $f(x) = ax + b, a > 0$, da für sie gilt:

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

Verhältnisskala

Intervallskala mit (natürlichem) Nullpunkt

- ▶ Beispiele: Zeitdauer, Preise, Längen, Gewichte
- ▶ Struktur: Abstände quantifizierbar und **Nullpunkt** eindeutig festgelegt
- ▶ Sinnvolle Operationen: gleich/ungleich, größer/kleiner, Differenzen, **Verhältnisse**
- ▶ Erlaubte Transformationen: Multiplikation/Reskalierung mit $f(x) = ax$, $a > 0$, da für sie gilt:

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

Absolutskala

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

Skalenniveau

Beachte:

- ▶ Je höher das Skalenniveau, desto mehr Rechenoperationen können mit den beobachteten Werten sinnvoll durchgeführt werden.
- ▶ Nur Rechenoperationen, deren inhaltliche Ergebnisse nicht von den zulässigen Transformationen der Skala beeinflusst werden, sind *sinnvoll* interpretierbar.

Zusammenfassung:

Überbegriff	Skalenniveau	auszählen	ordnen	Differenzen	Quotienten
qualitativ	Nominal	✓	X	X	X
	Ordinal	✓	✓	X	X
quantitativ/ metrisch	Intervall	✓	✓	✓	X
	Verhältnis	✓	✓	✓	✓
	Absolut	✓	✓	✓	✓

Sonderfall: dichotome, binär codierte ("0"/"1") Merkmale sind nominalskaliert, aber z.B. ihr Mittelwert ist sinnvoll interpretierbar.

Skalentransformationen

Beispiel: Konzentration von Bakterien

$$\begin{array}{ll} 0.003 & \log_{10}(3.0 \cdot 10^{-3}) \approx -2.5 \\ 0.0003 \quad \text{oder} & \log_{10}(3.0 \cdot 10^{-4}) \approx -3.5 \\ 0.00003 & \log_{10}(3.0 \cdot 10^{-5}) \approx -4.5 \end{array}$$

Skalenwahl \Leftrightarrow Interpretation der Differenz

Also: Inhaltlich sinnvoll, theoretisch nicht zulässig!¹

Bei log-Skala: Differenz = log(Faktor der Veränderung)

Oft: Verwende log zur Basis 10 oder 2

¹Willkommen in der wunderbaren Welt der angewandten Statistik...

Indexbildung

Zusammenfassung verschiedener Merkmale (*Items*) zu einem aggregierten Merkmal (*Score, Index*).

Häufig: Bildung von (gewichteten) Summen

Beispiel: Wahl-O-Mat

Übereinstimmung SPD = $w_1 \cdot Q(\text{Tempolimit}) + w_2 \cdot Q(\text{Verteidigungsausgaben}) + \dots$

Indexbildung meist theoriegeleitet bzw. nach fachspezifischen Überlegungen

Häufige Problematik:

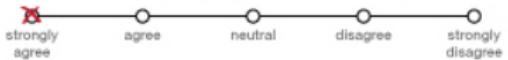
- ▶ Gewichtung der Items?
- ▶ Skalenniveau des Index?

Indexbildung: Likert-Skalen

1. The website has a user friendly interface



2. The website is usually my first choice for research.



3. The website has a good selection of images.



4. It is easy to upload new images to the website.



5. The website has a pleasing color scheme.



{Bild: Nicholas Smith, CC BY-SA 3.0, via Wikimedia Commons}

Indexbildung: Likert-Skalen

Häufig verwendetes Verfahren in der Psychometrie:

- ▶ Items sind Ratingskalen die Zustimmung/Ablehnung zu bestimmten Aussagen zum selben Thema abfragen
- ▶ Skalenwert der Likert-Skala als Summen- oder Durchschnittsscore der (ordinalen!) Ratings der verschiedenen Items

Merkmalstypen: Stetige und diskrete Merkmale

- ▶ **Diskretes** Merkmal:
nur endlich oder abzählbar unendlich viele verschiedene Werte möglich
z.B. Geschlecht, Kinderanzahl
- ▶ **Stetiges** Merkmal:
alle Werte in einem Kontinuum möglich
z.B. Zeitdauer, Größe, Gewicht

Wichtige Unterscheidung für Wahl geeigneter grafischer Darstellungsformen & numerischer Zusammenfassungen.

Weitere Klassen

- ▶ **Quasi-stetiges** Merkmal:
diskret, sehr kleine Einheiten, "praktisch" stetig.
Beispiel: Monetäre Größen in Cent
(Real existierende Messungen immer quasi-stetig wegen beschränkter Auflösung des Messinstruments & der Gleitkommazahlendarstellung im Rechner.)
- ▶ **Gruppierte** Daten, **Häufigkeits**daten:
Wertebereich eines (quasi-)stetigen Merkmals wird in Gruppen (Klassen, Kategorien) eingeteilt.
Beispiele: Gehalt in Gehaltsklassen, Alter in Altersklassen
Bemerkung: Gruppierung dient auch dem Datenschutz!

Datenerhebung & Messung

Messung

Skalenniveaus

Datenerhebung

Erhebungsarten: Umfang

Def: Vollerhebung

Alle statistischen Einheiten einer Grundgesamtheit werden untersucht ("erhoben").

Def: Stichprobe

Auch *Teilerhebung*. Ein Teil der Untersuchungseinheiten in einer Grundgesamtheit wird untersucht. Ist die Auswahl dieser UE zufällig, spricht man von einer **Zufallsstichprobe**.

Induktive Statistik ist in der Regel nur auf Basis geeigneter Zufallsstichproben zulässig!

Erhebungsarten: Datenform

Def.: Querschnittsdaten:

Ein oder mehrere verschiedene *Merkmale* werden an mehreren *Untersuchungseinheit* einmal erhoben (zu einem bestimmten Zeitpunkt oder über einen bestimmten Zeitraum).

Also: **Eine Beobachtung pro Untersuchungseinheit.**

Def.: Zeitreihe:

Ein bestimmtes *Merkmal* wird wiederholt für die selbe *Untersuchungseinheit* erhoben (üblicherweise in regelmäßigen Abständen).
Also: **Mehrere Beobachtungen einer Untersuchungseinheit.**

\vskip -.5em z.B. Aktienkurs eines Unternehmens, Corona-Inzidenz in einem Landkreis

Def.: Längsschnittdaten (auch: Longitudinal-, Paneldaten):

Die selben *Merkmale* werden mehrmals zu verschiedenen Zeitpunkten an mehreren *Untersuchungseinheiten* erhoben.

Also: Jeweils **mehrere Beobachtungen mehrerer Untersuchungseinheiten.**

Erhebungsarten: Methoden operationaler Messungen

- ▶ Beobachtung:
 - ▶ verdeckt oder teilnehmend
 - ▶ systematisch mit Beobachtungsprotokoll
- ▶ Befragung:
 - ▶ (fern)mündlich mit/ohne Interviewer, oder schriftlich/online
 - ▶ Fragebogen
- ▶ Experiment
 - ▶ kontrollierte Situation, evtl. Erhebung durch Beobachtung oder Befragung

Experimente

Es werden in der Regel verschiedene “Behandlungen” verglichen.

Gemeinsamkeit: Experimenteller Eingriff ins Geschehen

- ▶ **Randomisierte klinische Studie** : Zuordnung von Behandlungen zu Personen erfolgt durch Losverfahren (Randomisierung)
- ▶ Randomisierte Experimente (Produktion, Landwirtschaft)
Vorlesung: *Versuchplanung*
- ▶ Experimente in Medizin und Biologie
- ▶ Naturwissenschaftliche Experimente mit zufälligen Komponenten
- ▶ “A-B-Testing” in Web Development oder UX Design

Epidemiologische Studien

- ▶ **Kohortenstudien:**

Synonym für Längsschnittdaten (*retrospektiv oder prospektiv*)

Beispiel: EPIC Studie (European Prospective Investigation into Cancer)
400.000 Personen in neun europäischen Ländern

- ▶ **Fall-Kontroll-Studien:**

Erhebung von Erkrankten (Fälle) und dann dazu “passenden” Gesunden (Kontrollen)

Beispiel: Deutsche Radon Studie

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Wahrscheinlichkeit

Grundbegriff der Stochastik:

Wahrscheinlichkeit $P(A)$ für das Auftreten eines bestimmten Ereignisses A

$$P(A) = 1$$

A tritt mit Sicherheit ein

$$P(A) = 0$$

A tritt mit Sicherheit *nicht* ein

$$P(A) = p \in (0, 1)$$

A tritt mit Wahrscheinlichkeit p ein

Interpretation?

Subjektivistische Interpretation

Wahrscheinlichkeit aus Wetteinsatz:

“Wie sicher bist *Du*, dass das Ereignis A eintreten wird?”

→ “Wie viel Einsatz e bist *Du* maximal bereit zu setzen, wenn beim Eintreten von A ein Gewinn g ausgezahlt wird?” (unter Risikoneutralität)

$$\implies P(A) = \frac{e}{g}$$

⇒ Wahrscheinlichkeit als Maß für **individuelle/subjektive Unsicherheit**.

Beispiel: Sportwette

Nick zahlt Yolanda den doppelten Einsatz als Gewinn aus, falls Yolanda's Lieblingsteam die Meisterschaft gewinnt (Ereignis M).

- ▶ Nur falls Yolanda glaubt, dass $P(M) \geq \frac{1}{2}$, lässt sie sich auf die Wette ein.
- ▶ Nick glaubt $P(M) \leq \frac{1}{2}$, sonst würde er die Wette nicht anbieten.

Frequentistische Interpretation

Wahrscheinlichkeit als **Häufigkeit**:

Wenn der zufällige Vorgang beliebig oft wiederholt werden würde, dann würde die **relative Häufigkeit** des Eintretens des Ereignisses A gegen die Wahrscheinlichkeit $P(A)$ konvergieren.

Klassisches Beispiel:

Wiederholtes Werfen eines “fairen” Würfels

Interpretation der Interpretationen

- ▶ Für mathematische Theoriebildung weitestgehend irrelevant was " $P(A)$ " *inhaltlich* bedeutet.
(Für Kommunikation und Interpretation statistischer Analysen aber *höchst relevant!*)
- ▶ Beide Interpretationen jeweils (nicht) sinnvoll für bestimmte Arten von Zufallsvorgängen:
 - ▶ Frequentistische Interpretation der W.keit von *einzigartigen* Ereignissen?
... von bereits eingetretenen, aber *unbeobachteten* Ereignissen?
 - ▶ Subjektivistische Interpretation von W.keiten für einfache & wiederholbare physikalische Prozesse wie Würfelwurf? ...
- ▶ Anwendungen des subjektivistischen W.keitsbegriff:
→ **Bayesianische Statistik**
- ▶ Anwendungen des frequentistischen W.keitsbegriff:
→ **Klassische/Frequentistische Statistik**

Grundbegriffe der Stochastik

Def.: Ergebnisse ω

ω : mögliches **Ergebnis** eines Zufallsexperiments.

Def.: Stochastische Grundgesamtheit Ω

Als **stochastische Grundgesamtheit** Ω bezeichnet man die Menge aller möglichen Ergebnisse ω eines Zufallsexperiments.

Auch: "Grundraum", "Basismenge", "Ergebnisraum".

Def.: Ereignis

Eine Teilmenge $A \subseteq \Omega$ heißt **Ereignis**.

Def.: Elementarereignisse $\{\omega\}$

Eine Teilmenge von Ω , die als einziges Element ein Ergebnis ω enthält, nennt man **Elementareignis**.

Immer nur genau ein Elementarereignis tritt ein.

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Laplace-Prinzip

Prinzip von Laplace (Pierre-Simon de Laplace [1749-1827]):

*“Wenn nichts dagegen spricht, gehen wir davon aus, dass **alle Elementarereignisse gleichwahrscheinlich** sind.”*

Laplace-Wahrscheinlichkeiten

Betrachte die endliche Grundgesamtheit von Elementarereignissen

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$$

Def.: Laplace-Wahrscheinlichkeit

Für ein **Ereignis** $A \subseteq \Omega$ ist die *Laplace-Wahrscheinlichkeit*

$$P(A) := \frac{|A|}{|\Omega|} = \frac{|A|}{n},$$

wobei $|A|$ die Anzahl der Elemente in A ist.

⇒ “Anzahl *günstiger* und gleichwahrscheinlicher Fälle durch Anzahl *möglicher* und gleichwahrscheinlicher Fälle”

Folgerungen und Erweiterungen

- ▶ Jedes Elementarereignis $\omega_i, i = 1, \dots, n$, eines Laplace-W. keitsraums hat Wahrscheinlichkeit $P(\{\omega_i\}) = \frac{1}{n}$.
- ▶ Die Wahrscheinlichkeit von Ω ist $P(\Omega) = 1$.
- ▶ Die entsprechende **Abbildung** $P : \mathcal{P}(\Omega) \rightarrow [0, 1]$ nennt man auch **diskrete Gleichverteilung** auf Ω
 - ▶ $\mathcal{P}(\Omega)$ ist die **Potenzmenge** (Menge aller Teilmengen) von Ω – nicht zu verwechseln mit $P(\Omega)$!
- ▶ Die **Vereinigung** $U = A \cup B$ zweier Ereignisse A, B definiert das Ereignis “A oder B oder beide treten ein”
- ▶ Der **Schnitt** $I = A \cap B$ zweier Ereignisse A, B definiert das Ereignis “A und B treten beide ein”

Beispiel: Augensumme von zwei Würfeln

$$\Omega = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$$

$$|\Omega| = 6^2 = 36$$

Sei A_k das Ereignis "Augensumme ist k ". Dann gilt:

$$P(A_k) = \frac{6 - |k - 7|}{36} \quad \text{für } k = 2, \dots, 12$$

Interlude: Kombinatorik auf 1er Folie

Wie viele unterschiedliche Möglichkeiten, k Elemente aus n auszuwählen?

- **Auswahl mit Zurücklegen, Ergebnis mit Reihenfolgen:**

$$n^k$$

→ je n Möglichkeiten für alle k Ziehungen

- **Auswahl ohne Zurücklegen, Ergebnis mit Reihenfolge:**

$$\frac{n!}{(n-k)!} = n \cdot (n-1) \cdot \dots \cdot (n-k+1)$$

→ zuerst n Möglichkeiten, dann je $n-1$, dann je $n-2$, ...

- **Auswahl ohne Zurücklegen, Ergebnis ohne Reihenfolge:**

$$\binom{n}{k} := \frac{n!}{(n-k)!k!} = \frac{\frac{n!}{(n-k)!}}{k!}$$

→ Anzahl unterschiedlicher Auswahlen / Anzahl möglicher Reihenfolgen

- **(mit Zurücklegen, ohne Reihenfolge: $\binom{n+k-1}{k}$)**

Beispiel: Skatspiel

Beim Skatspiel werden 32 verschiedene Karten, darunter 4 Buben, an 3 Personen verteilt.

Jede:r erhält 10 Karten.

2 Karten kommen in den Skat.

Wie groß sind die Laplace-Wahrscheinlichkeiten der Ereignisse:

- ▶ $A_1 \approx \text{"Person 1 erhält alle Buben"}$
- ▶ $A_2 \approx \text{"Alle 3 erhalten genau einen Buben"}$

Laplace-Wahrscheinlichkeiten sind zu speziell:

Gegenbeispiele:

- ▶ Unfairer Würfel
- ▶ Seltene Ereignisse, z.B. hard disk failures, Mutationen, ...
~~ Elementarereignisse hier nicht gleichwahrscheinlich!
- ▶ Außerdem: Was wenn $|\Omega|$ unendlich?

Beispiel für unendliche Grundgesamtheiten

Man interessiere sich für die Anzahl der Würfe einer fairen Münze bis zum ersten Mal Zahl eintritt.

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \dots\} = \{1, 2, 3, 4, \dots\} = \mathbb{N} \implies |\Omega| = \infty!$$

Allgemein mit $\omega_i := i$:

$$P(\{\omega_i\}) = \frac{1}{2^i} \quad i = 1, 2, 3, \dots$$

$$\sum_{i=1}^{\infty} P(\{\omega_i\}) = \sum_{i=1}^{\infty} \frac{1}{2^i} = 1$$

Beweis s. geom. Reihe

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Definitionen

Def: Disjunkte Ereignisse

$A, B \subseteq \Omega$ sind *disjunkte Ereignisse* wenn gilt:

$$A \cap B = \emptyset$$

Def: Komplement/Gegenereignis

Das *Gegenereignis* oder *Komplement* \bar{A} von $A \subseteq \Omega$ ist

$$\bar{A} := \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$$

- ▶ Ereignis und Gegenereignis sind disjunkt.
- ▶ Disjunkte Ereignisse können also nie gleichzeitig bzw. gemeinsam eintreten
 \implies Elementarereignisse in Ω bilden eine Menge von *disjunkten* Ereignissen.

Kolmogorov-Axiome

Wir betrachten

- ▶ eine beliebige (abzählbare) Grundgesamtheit Ω
- ▶ und eine Funktion P auf der Potenzmenge $\mathcal{P}(\Omega)$, die jedem Ereignis $A \subseteq \Omega$ eine Wahrscheinlichkeit zuordnet.

Def.: Wahrscheinlichkeitsverteilung

P ist eine *Wahrscheinlichkeitsverteilung auf Ω* , wenn sie folgende Eigenschaften erfüllt:

- ▶ **A1:** $P(A) \geq 0$ für beliebige $A \subseteq \Omega$ (*Positivität*)
- ▶ **A2:** $P(\Omega) = 1$ (*Sicheres Ereignis*)
- ▶ **A3:** $P(A \cup B) = P(A) + P(B)$ für disjunkte Ereignisse $A, B \subseteq \Omega$ (*Additivität*)

Folgerungen

- $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$ für paarweise disjunkte Ereignisse $A_1, A_2, \dots, A_n \subset \Omega$
- $A \subseteq B \implies P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ für beliebige $A, B \subset \Omega$

Anwendung: Siebformel von Sylvester-Poincaré

James Sylvester [1814-1897], Jules Henri Poincaré [1854-1912]

Für beliebiges $n \in \mathbb{N}$ und Ereignisse $A_1, A_2, \dots, A_n \subseteq \Omega$ gilt:

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) \\ &\quad + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) \\ &\quad - \dots \pm \dots \pm \dots + (-1)^{n+1} \cdot P(A_1 \cap A_2 \cap \dots \cap A_n) \end{aligned}$$

Bsp.:

$$\begin{aligned} P(A \cup B \cup C) &= (P(A) + P(B) + P(C)) \\ &\quad - (P(A \cap B) + P(A \cap C) + P(B \cap C)) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$

Anwendung: Bonferroni Ungleichungen

⇒ Abschätzung von $P(A_1 \cup A_2 \cup \dots \cup A_n)$

Für beliebige Ereignisse A_1, A_2, \dots, A_n gilt

$$\sum_i P(A_i) \geq P\left(\bigcup_{i=1}^n A_i\right) \geq \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j)$$

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Bedingte Wahrscheinlichkeit

Def.: Bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit von A gegeben B* für Ereignisse $A, B \subseteq \Omega$ mit $P(B) > 0$ ist

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Interpretation:

- ▶ “Wie wahrscheinlich ist A wenn B bereits eingetreten ist?”
- ▶ “Wie wahrscheinlich ist A unter der Annahme, dass B der Fall ist?”

Intuition:

- ▶ “In welchem Anteil von den Fällen, in denen B eintritt, tritt auch A ein?”
- ▶ “Welchen Anteil an der Wahrscheinlichkeit für Bedingung B haben die Elementarereignisse, die auch Teil des interessierenden Ereignisses A sind?”

Beispiel: Würfelwurf

$G :=$ "Würfelergebnis gerade Zahl",

$F :=$ "Würfelergebnis mindestens 5"

$S :=$ "Würfelergebnis ist 6"

Dann z.B.:

$$\implies P(F|G) = \frac{1}{3}$$

$$P(G|F) = \frac{1}{2}$$

$$P(S|\bar{G}) = 0$$

$$P(G|S) = 1$$

Eigenschaften von bedingten Wahrscheinlichkeiten

$P(B|B) = 1$ (Sicheres Ereignis)

$P(\bar{B}|B) = 0$ (Unmögliches Ereignis)

$P(A|B) \geq 0$ für beliebige $A \subseteq \Omega$ (Positivität)

Außerdem gelten alle Aussagen über Wahrscheinlichkeiten **bei fester Bedingung** auch für *bedingte* Wahrscheinlichkeiten, z.B.:

$$P((A_1 \cup A_2)|B) = P(A_1|B) + P(A_2|B) \text{ für } A_1 \text{ und } A_2 \text{ disjunkt. (Additivität)}$$

⇒ Die bedingten Wahrscheinlichkeiten $P(A|B)$ für $A \subseteq \Omega$ definieren *bei fester Bedingung B* einfach eine **Wahrscheinlichkeitsverteilung über die neue, kleinere stochastische Grundgesamtheit $B \subset \Omega$.**

Mehrfaches Bedingen ⇒ Bedingen auf den Schnitt der Bedingungen:

$$P((A|B)|C) = P(A|(B \cap C))$$

Beispiel: Skat

Definiere

$A :=$ "Mindestens eine der acht Karokarten liegt im Skat"

$B :=$ "Spieler 1 erhält beim Austeilen keine der acht Karokarten"

Berechne $P(A)$ und $P(A|B)$ und vergleiche diese.

Multiplikationssatz

Für beliebige Ereignisse A_1, A_2, \dots, A_n mit $P(A_1 \cap A_2 \cap \dots \cap A_n) > 0$ gilt:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot \dots \cdot P(A_n | A_1 \cap \dots \cap A_{n-1})$$

wobei man die rechte Seite offensichtlich auch in jeder anderen möglichen Reihenfolge faktorisieren kann.

\vskip 3em Alternative informelle Schreibweise:

$$P(A_1, A_2) := P(A_1 \cap A_2)$$

insbesondere gilt also z.B.

$$P(A_1, A_2) = P(A_1) \cdot P(A_2 | A_1),$$

$$P(A_1, A_2, A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1, A_2), \dots$$

Satz von der totalen Wahrscheinlichkeit

Def: Disjunkte Zerlegung (Partition)

Die Mengen B_1, B_2, \dots, B_n bilden eine *disjunkte Zerlegung* von Ω , falls

- B_1, B_2, \dots, B_n paarweise disjunkt: $B_i \cap B_j = \emptyset \quad \forall i \neq j$
- und $B_1 \cup B_2 \cup \dots \cup B_n = \Omega$

Satz von der totalen Wahrscheinlichkeit

Sei B_1, B_2, \dots, B_n eine *disjunkte Zerlegung* von Ω mit $P(B_i) > 0$ für $i = 1, \dots, n$, dann gilt für jedes $A \subseteq \Omega$:

$$P(A) = \sum_{i=1}^n P(A|B_i) \cdot P(B_i)$$

Wichtiger Spezialfall

Insbesondere gilt

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

da B, \bar{B} eine disjunkte Zerlegung von Ω ist.

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Stochastische Unabhängigkeit: Motivation

Frage: Wann sind 2 Ereignisse A, B **stochastisch unabhängig**?

Motivation über bedingte Wahrscheinlichkeiten:

Zwei Ereignisse A, B sind *stochastisch unabhängig*, wenn

$$\underbrace{P(A|B)}_{\frac{P(A \cap B)}{P(B)}} = P(A)$$

$$\text{bzw. } \underbrace{P(B|A)}_{\frac{P(A \cap B)}{P(A)}} = P(B)$$

Intuition:

Zwei Ereignisse sind stochastisch unabhängig, wenn das Eintreten des einen nichts an der Wahrscheinlichkeit des Eintretens des anderen verändert.

Stochastische Unabhängigkeit

Def: Stochastische Unabhängigkeit

Zwei Ereignisse A, B sind *stochastisch unabhängig*, wenn gilt:

$$P(A \cap B) = P(A) \cdot P(B)$$

Notation: $A \perp B : \iff P(A \cap B) = P(A) \cdot P(B)$

- ▶ Voraussetzungen $P(B) > 0$ und/oder $P(A) > 0$ dafür nicht nötig.
- ▶ Unabhängigkeit überträgt sich auf die Gegenereignisse:
 - ▶ $A \perp B \iff \bar{A} \perp B$
 - ▶ $A \perp B \iff A \perp \bar{B}$
 - ▶ $A \perp B \iff \bar{A} \perp \bar{B}$

Beispiel: Zweimaliges Würfeln

Ein fairer Würfel wird zweimal hintereinander geworfen. Definiere

A : "Beim 1. Würfelwurf eine Sechs"

B : "Beim 2. Würfelwurf eine Sechs"

Bei jedem Würfelwurf ist die Grundgesamtheit $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Nach Laplace gilt $P(A) = P(B) = \frac{1}{6}$.

Bei "unabhängigem" Werfen gilt somit

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{36}$$

Beispiel: Zweimaliges Würfeln mit Tricks

Angenommen die Würfelwerfende legt es darauf an einen Pasch zu würfeln. Sie kann den zweiten Wurf so steuern, dass sie mit W.keit 0.5 das gleiche Ergebnis wie beim ersten Wurf würfelt. Die fünf anderen möglichen Ergebnisse haben dann jeweils W.keit 0.1.

Dann ist zwar $P(A) = \frac{1}{6}$ und auch $P(B) = \frac{1}{6}$, aber

$$P(A \cap B) = P(A)P(B|A) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Die Ereignisse A und B sind also abhängig, da

$$\frac{1}{12} = P(A \cap B) \neq P(A) \cdot P(B) = \frac{1}{36}$$

Unabhängigkeit von mehr als zwei Ereignissen

Allgemeiner:

Def.: Stochastische Unabhängigkeit von mehr als zwei Ereignissen

Ereignisse A_1, A_2, \dots, A_n sind stochastisch unabhängig, wenn für **alle Teilmengen** $I \subseteq \{1, 2, \dots, n\}$ mit $I = \{i_1, i_2, \dots, i_k\}$ gilt:

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k})$$

Bemerkung:

Aus paarweiser Unabhängigkeit folgt **nicht** die Unabhängigkeit von mehr als zwei Ereignissen.

Beispiel zur paarweisen Unabhängigkeit

$\Omega = \{0, 1, 2, 3\}$ Laplace-Wahrscheinlichkeitsraum

$A_i = \{0\} \cup \{i\}$ mit $i = 1, 2, 3$

(z.B.: einmaliges Ziehen aus einer Urne mit 4 nummerierten Kugeln)

Dann gilt: $P(A_i) = \frac{1}{2} \quad \forall i$

und $P(A_i \cap A_j) = P(\{0\}) = \frac{1}{4} = P(A_i) \cdot P(A_j) \quad \forall i \neq j.$

$\implies A_i$ alle paarweise unabhängig – Aber:

$$P(A_1 \cap A_2 \cap A_3) = P(\{0\}) = \frac{1}{4}$$

$$P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{1}{8}$$

$\implies A_1, A_2, A_3$ also **nicht** unabhängig.

Bedingte Unabhängigkeit

Def.: Bedingte Unabhängigkeit

Sei C ein beliebiges Ereignis mit $P(C) > 0$. Zwei Ereignisse A und B nennt man **bedingt unabhängig gegeben** C , genau dann wenn

$$P(A \cap B|C) = P(A|C) \cdot P(B|C).$$

Notation: $(A \perp B)|C : \iff P(A \cap B|C) = P(A|C) \cdot P(B|C)$

- ▶ Für feste Bedingung C gelten die selben Folgerungen wie für unbedingte stochastische Unabhängigkeit:
 $(A \perp B)|C \iff (\bar{A} \perp B)|C; (A \perp \bar{B})|C; (\bar{A} \perp \bar{B})|C$
- ▶ Aus bedingter stochastischer Unabhängigkeit folgt nicht unbedingte stochastische Unabhängigkeit!
- ▶ Aus unbedingter stochastischer Unabhängigkeit folgt nicht bedingte stochastische Unabhängigkeit!

Beispiel 1 zu bedingter (Un)abhängigkeit

Eine Box enthält 2 Münzen – eine normale und eine gezinkte, die auf beiden Seiten “Kopf” zeigt.

Eine Münze wird zufällig aus der Box gezogen und zweimal geworfen.

Definiere

$A :=$ “Kopf beim 1. Wurf”,

$B :=$ “Kopf beim 2. Wurf”,

$C :=$ “Normale Münze wurde ausgewählt”.

Dann gilt:

$A \perp\!\!\!\perp B$, aber $(A \perp B) | C$!

Also: aus “bedingt stochastisch unabhängig” folgt *nicht* “stochastisch unabhängig”.

Beispiel 2 zu bedingter (Un)abhängigkeit

Szenario: Einfacher Würfelwurf

Definiere Ereignisse $A = \{1, 2\}$; $B = \{2, 4, 6\}$; $C = \{1, 4\}$.

Dann gilt: $A \perp B$ (!), aber $(A \not\perp B) | C$!

Also: aus “stochastisch unabhängig” folgt *nicht* “bedingt stochastisch unabhängig”.

Wahrscheinlichkeit: Grundlagen & Definitionen

Wahrscheinlichkeit: Begriffsbildung & Interpretation

Laplace-Wahrscheinlichkeiten

Axiome von Kolmogorov

Bedingte Wahrscheinlichkeiten

Stochastische Unabhängigkeit

Der Satz von Bayes

Der Satz von Bayes

Thomas Bayes [1701-1761]

Dieser Satz beruht auf der Asymmetrie der Definition von bedingten Wahrscheinlichkeiten:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \implies P(A \cap B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \implies P(A \cap B) = P(B|A)P(A)$$

$$\implies P(A|B)P(B) = P(B|A)P(A)$$

Der Satz von Bayes II

Satz von Bayes

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A)} \\ &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \end{aligned}$$

Allgemeiner gilt für eine disjunkte Zerlegung B_1, \dots, B_n :

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}$$

Interpretation

$P(B_i)$ als *a-priori*-Wahrscheinlichkeiten:

- \approx Vorwissen über bzw. Plausibilität der Annahmen/Hypothesen B_i

$P(A|B_i)$ als *Likelihood* von A:

- wie plausibel ist die Beobachtung A *unter der Annahme dass B_i der Fall ist*

$P(B_i|A)$ *a-posteriori*-Wahrscheinlichkeiten:

- Auf Basis der Beobachtung von A ändert sich die Wahrscheinlichkeit der Hypothese B_i von $P(B_i)$ zu $P(B_i|A)$.

\implies Bayes liefert “Update-Regel”:

Wie verändert sich durch die Beobachtung von Daten A die Plausibilität der Vorannahme B_i ?

Bedeutung

⇒ Satz von Bayes liefert ein extrem mächtiges Verfahren zur *Umkehr der Bedingungsreihenfolge*:

- ▶ von $P(A|B_i)$:
“Wie wahrscheinlich wäre es, Daten A zu beobachten falls B_i der Fall wäre?”
- ▶ zu $P(B_i|A)$:
“Wie plausibel/wahrscheinlich ist B_i wenn ich Daten A beobachtet habe?”

Ermöglicht Rückschluss auf Plausibilität nicht direkt beobachtbarer Phänomene/Modellannahmen B_i , gegeben beobachtete Daten A und Wahrscheinlichkeitsmodell $P(A|B_i)$

Beispiel: Diagnostischer Test

$K :=$ "Person ist krank"

$T :=$ "Test auf Krankheit ist positiv"

Gegeben:

- ▶ **Sensitivität** $P(T|K)$ ("wahr-positiv"-Rate) des Tests
- ▶ **Spezifität** $P(\bar{T}|\bar{K})$ ("wahr-negativ"-Rate) des Tests
- ▶ **Prävalenz** $P(K)$ der Krankheit in der Population.

Für Therapieentscheidungen relevant:

- ▶ $P(K|T)$ ("positiv prädiktiver Wert")
- ▶ $P(\bar{K}|\bar{T})$ ("negativ prädiktiver Wert")

Beispiel: Diagnostischer Test II

Gegeben:

Sensitivität / wahr-positiv-Rate: $P(T|K) = 0.8$

Spezifität / wahr-negativ-Rate: $P(\bar{T}|\bar{K}) = 0.99$

Prävalenz: $P(K) = 0.001$

$$\begin{aligned}\implies P(T) &= P(T|K)P(K) + P(T|\bar{K})P(\bar{K}) = \\ &= P(T|K)P(K) + (1 - P(\bar{T}|\bar{K}))(1 - P(K)) \approx 0.011\end{aligned}$$

$$P(K|T) = \frac{P(T|K)P(K)}{P(T)} \approx 0.073 \text{ (!!)}$$

$$P(\bar{K}|\bar{T}) \approx 0.989$$

Odds / “Chancen”

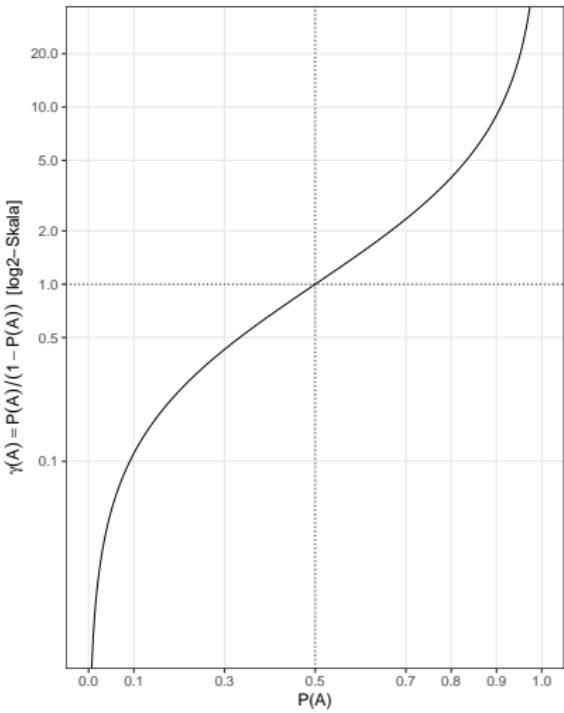
Alternative Darstellung von W.keit als (Wett-)Chance
(z.B. 1:10 (“1 zu 10”) oder 3:1, engl. “odds”):

Def.: Odds (“Chance”)

Die Odds $\gamma(A)$ für ein Ereignis A sind

$$\gamma(A) := \frac{P(A)}{1 - P(A)} \in [0, \infty]$$

$$\implies P(A) = \frac{\gamma(A)}{1 + \gamma(A)}$$



Satz von Bayes in Odds-Notation

$$\frac{P(B|A)}{P(\bar{B}|A)} = \frac{P(B)}{P(\bar{B})} \cdot \frac{P(A|B)}{P(A|\bar{B})}$$

also: $\gamma(B|A) = \gamma(B) \cdot \frac{P(A|B)}{P(A|\bar{B})}$

Posterior Odds = Prior Odds · Likelihood Ratio

Je weiter der *Likelihood Quotient* von 1 weg ist, desto aussagekräftiger ist die Beobachtung von A bezüglich der Plausibilität der Annahme B.

Bsp: Bayesianisches Update mit Odds

Mit

$$P(T|K) = 0.8; \quad P(\bar{T}|\bar{K}) = 0.99; \quad P(K) = 0.001$$

ergibt sich:

$$\text{Prior Odds: } \gamma(K) = \frac{0.001}{0.999} = \overline{0.001} = 1 : 999$$

$$\text{Likelihood Ratio: } \frac{P(T|K)}{P(T|\bar{K})} = \frac{0.8}{0.01} = 80 = 80 : 1$$

$$\implies \text{Posterior Odds: } \gamma(K|T) = \frac{1}{999} \cdot \frac{80}{1} = \overline{0.080} \approx 1 : 12!!$$

Also: Nur bei 1 von 13 positiven Test liegt tatsächlich eine Krankheit vor....

Obacht, sehr häufiger Fehlschluss "**base-rate fallacy**" –
verwechselt Likelihood(-Ratio) (hier: positiver Test 80-mal
wahrscheinlicher bei Kranken)

mit Posteriori (hier: *wahr* positive Tests sind viel seltener als falsch
positive)

und vergisst/ignoriert Einfluss der Priori/base-rate (hier: 1000-mal mehr
Gesunde als Kranke)... **remember your priors!**

Bsp 2: Bayes

75% der Mathestudierenden können Python.

15% der anderen Studierenden können Python.

Von 1100 Studierenden studieren 100 Mathe.

Sei M := “Tom studiert Mathe.”

Sei P := “Tom kann Python”

Berechnen Sie $\gamma(M|P)$ bzw $P(M|P)$.

Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

Einführung

Zurück zur Empirie –

Wie können wir in beobachteten Daten Aussagen über die
(Un-)Abhängigkeit von Merkmalen machen?

⇒ **mehrdimensionale** oder **multivariate** Daten

Multivariate Daten

- ▶ Stichprobe mit Untersuchungseinheiten $i = 1, \dots, n$
- ▶ **Beobachtungen** (x_i, y_i, z_i) der **Merkmale** (X, Y, Z)
- ▶ Daten $(x_1, y_1, z_1), \dots, (x_i, y_i, z_i), \dots, (x_n, y_n, z_n)$
(Im weiteren: Meistens nur 2 Merkmale X, Y)
- ▶ Fragestellungen:
 - ▶ $X \leftrightarrow Y$: (Wie) hängen X und Y zusammen?
Assoziation, Korrelation
("Sind X und Y unabhängig?")
 - ▶ $X \rightarrow Y$: (Wie) beeinflusst X das (Ziel-)Merkmal Y ?
Regression, Kausalität (... nicht Teil dieses Kurses)

Diskrete und gruppierte Merkmale

- ▶ Darstellung, Präsentation von diskreten Merkmalen X und Y mit den Ausprägungen

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

- ▶ Skalenniveau von X, Y hier beliebig; X, Y können auch gruppierte metrische Merkmale sein.
- ▶ benutzt wird hier allerdings nur das *Nominalskalenniveau* der Merkmale.

Kontingenztabellen

Nachwahlbefragung Bundestag 2021:

\vskip 3em

	SPD	CDU/CSU	Grüne	FDP	AfD	Linke	Rest	Σ
Männer	626	601	352	325	300	125	175	2504
Frauen	567	504	335	211	167	104	210	2098
Σ	1193	1105	687	536	467	229	385	4602

\vskip 5em Quelle: Nachwahlbefragung ARD via [statista.com](https://www.statista.com)

Beispiel: Arbeitslosigkeit

Zwei Merkmale:

- ▶ X Ausbildungsniveau mit den Kategorien
 - ▶ "keine Ausbildung"
 - ▶ "Lehre"
 - ▶ "fachspezifische Ausbildung"
 - ▶ "Hochschulabschluss"
- ▶ Y Dauer der Arbeitslosigkeit mit den Kategorien
 - ▶ "Kurzzeitarbeitslosigkeit" (≤ 6 Monate)
 - ▶ "mittelfristige Arbeitslosigkeit" (7–12 Monate)
 - ▶ "Langzeitarbeitslosigkeit" (≥ 12 Monate)

Arbeitslosigkeit

	Kurzzeit- arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit- arbeitslosigkeit	Σ
k.A.	86	19	18	123
Lehre	170	43	20	233
Fachspez.	40	11	5	56
Hochschule	28	4	3	35
Σ	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	\dots	b_m	
a_1	h_{11}	\dots	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	\dots	h_{2m}	$h_{2\cdot}$
\cdot	\cdot		\cdot	\cdot
a_k	h_{k1}	\dots	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	\dots	$h_{\cdot m}$	n

Notation

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

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

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

mit

$$h_{i\cdot} = \sum_{j=1}^m h_{ij}, \quad h_{\cdot j} = \sum_{i=1}^k h_{ij}$$

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	\dots	b_m	
a_1	f_{11}	\dots	f_{1m}	$f_{1\cdot}$
\cdot	\cdot		\cdot	\cdot
a_k	f_{k1}	\dots	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	\dots	$f_{\cdot m}$	1

Notation

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

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

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

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

Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

Bedingte Häufigkeiten

Zusammenhang zwischen X und Y aus **gemeinsamen** Häufigkeiten h_{ij} bzw. f_{ij} schwer ersichtlich.

Deshalb: Blick auf **bedingte** Häufigkeiten

⇒ Verteilung eines Merkmals für festen Wert des 2. Merkmals

Bedingte Häufigkeiten: Beispiel

Wahlverhalten nach Geschlecht, Bundestagswahl 2021:

	SPD	CDU/CSU	Grüne	FDP	AfD	Linke	Rest	
Männer	25	24	14	13	12	5	7	100
Frauen	27	24	16	10	8	5	10	100

Prozentzahlen für Parteipräferenz in den Schichten (Subpopulationen)
“Frauen”, “Männer”

= **bedingte relative Häufigkeiten für Parteipräferenzen bei gegebenem Geschlecht**

Bedingte relative Häufigkeitsverteilung

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

$$f_{Y|X}(b_1|a_i) = \frac{h_{i1}}{h_{i\cdot}}, \dots, f_{Y|X}(b_m|a_i) = \frac{h_{im}}{h_{i\cdot}}.$$

Die **bedingte Häufigkeitsverteilung von X unter der Bedingung $Y = b_j$** , kurz

$X|Y = b_j$, ist bestimmt durch

$$f_{X|Y}(a_1|b_j) = \frac{h_{1j}}{h_{\cdot j}}, \dots, f_{X|Y}(a_k|b_j) = \frac{h_{kj}}{h_{\cdot j}}.$$

Bemerkung

Wegen

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

gilt auch

$$f_{Y|X}(b_1|a_i) = \frac{f_{i1}}{f_{i\cdot}}, \dots, f_{Y|X}(b_m|a_i) = \frac{f_{im}}{f_{i\cdot}}$$

$$f_{X|Y}(a_1|b_j) = \frac{f_{1j}}{f_{\cdot j}}, \dots, f_{X|Y}(a_k|b_j) = \frac{f_{kj}}{f_{\cdot j}}.$$

Merksatz:

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

Beispiel: Wahlverhalten

	SPD	CDU/CSU	Grüne	FDP	AfD	Linke	Rest	Σ
Männer	626	601	352	325	300	125	175	2504
Frauen	567	504	335	211	167	104	210	2098

- Zeile $X = a_1$ (Männer)

Bedingte relative Häufigkeiten für Parteipräferenz der Männer

$$f_{Y|X}(Y = b_j | X = a_1):$$

“1.Zeile / Randhäufigkeit für Männer”

$$\frac{h(a_1, b_1)}{h(a_1)} = f_{Y|X}(b_1 | a_1), \dots, \frac{h(a_1, b_j)}{h(a_1)} = f_{Y|X}(b_j | a_1), \dots$$

$$\frac{626}{2504} \approx 25\%, \dots, \frac{175}{2504} \approx 7\% \text{ usw.}$$

- Zeile $X = a_2$ (Frauen) analog, z.B. $\frac{h(a_2, b_1)}{h(a_2)} = \frac{567}{2098} \approx 27\%$ usw.

Bedingte und gemeinsame Häufigkeiten

Man kann auch umgekehrt aus bedingten Häufigkeiten und Randhäufigkeiten die gemeinsamen Häufigkeiten ausrechnen. Bei der Nachwahlbefragung aus den bedingten Häufigkeiten und Randhäufigkeiten also z.B.

$$h(a_1) = 2504 \text{ Männer}, \quad h(a_2) = 2098 \text{ Frauen}; \quad n = 4602.$$

$$\begin{aligned} & h(a_1) \cdot f(b_1 | a_1) = h(a_1, b_1) \\ \implies & 2504 \cdot 25\% \approx 626 \quad \text{usw.} \end{aligned}$$

Beispiel: Arbeitslosigkeit

$f(b_j | a_i), \quad X = a_i, \quad i = 1, \dots, 4$ Ausbildungsniveau

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

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

usw.

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

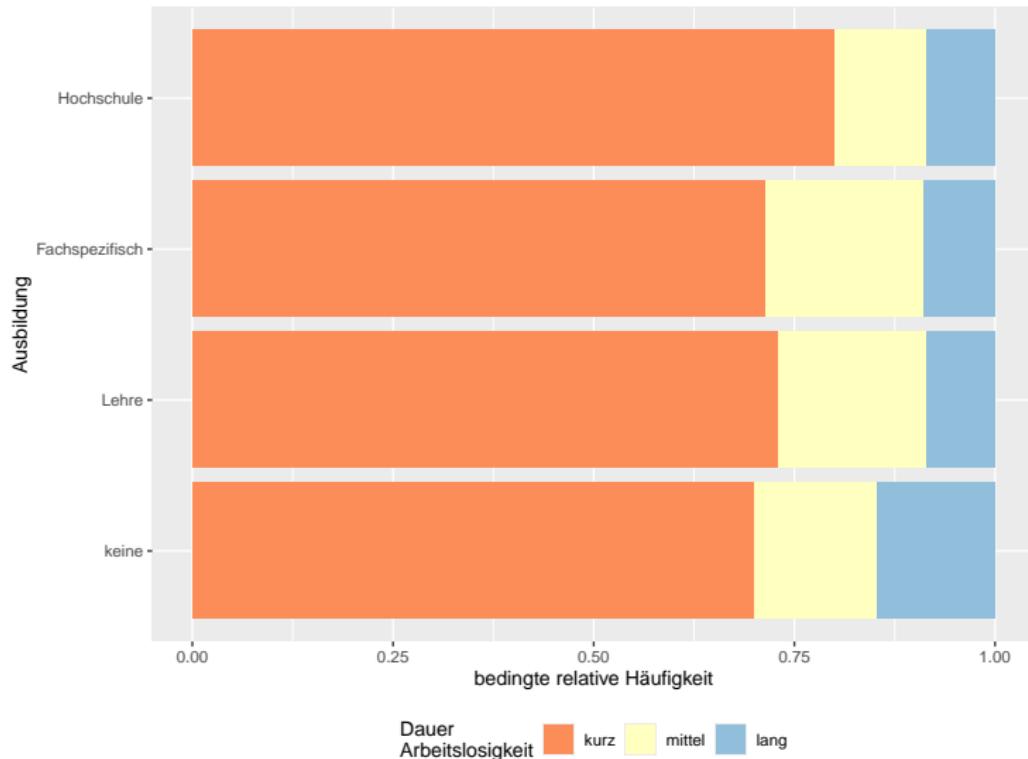
Bedingte Verteilung

	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:
 ⇒ Verteilung der Dauer der Arbeitslosigkeit für die Subpopulationen "Keine Ausbildung", "Lehre", usw.
- ▶ Verteilungen lassen sich nun miteinander vergleichen
 ⇒ Relative Häufigkeit für Kurzzeitarbeitslosigkeit ist in der Subpopulation "Hochschulabschluss" mit 0.8 am größten.

Darstellung der bedingten Verteilung

Zum Beispiel Stapeldiagramme:



Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

Zusammenhangsanalyse in Kontingenztabellen

Bisher: Tabellarische / grafische Präsentation

Jetzt: Maßzahlen für Stärke des Zusammenhangs zwischen X und Y.

Chancen und relative Chancen:

- Betrachte zunächst nur 2×2 - Kontingenztafeln

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

Odds (Chancen)

- Wir betrachten die Merkmale X und Y zunächst asymmetrisch: Die Ausprägungen von X definieren (hier 2) Subpopulationen, Y ist das interessierende dichotome Merkmal in diesen Subpopulationen
- Unter einer **Chance (odds)** versteht man nun das **Verhältnis der Häufigkeiten** von $Y = 1$ und $Y = 2$.
- **bedingte Odds:** Verhältnis der bedingten Häufigkeiten von $Y = 1$ und $Y = 2$ **in einer Subpopulation** $X = a_i$.

Chancenverhältnis (Odds Ratio)

- Die (empirischen) **bedingten Odds** für festes $X = a_i$ sind definiert als

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

- 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 Odds für $Y = 1$ gegen $Y = 2$ in der 1. Population ($X = 1$, 1. Zeile) zu den entsprechenden Odds in 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 ergeben sich für Personen mit fachspezifischer Ausbildung die Odds, kurzzeitig arbeitslos zu sein statt mittel- oder langfristig arbeitslos zu sein, als

$$\gamma(1, 2 | \text{fachsp. Ausbildung}) = \frac{40}{16} = 5 : 2 = 2.5.$$

Beispiel: Dauer der Arbeitslosigkeit

Für Arbeitslose mit Hochschulabschluss erhält man

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

Für fachspezifische Ausbildung stehen die Odds für Kurzarbeitslosigkeit somit 5 : 2, für Arbeitslose mit Hochschulabschluss 4 : 1.

Man erhält für fachspezifische Ausbildung und Hochschulabschluss das **Odds Ratio**

$$\gamma(1,2|\text{fachsp. Ausbildung, Hochschule}) = \frac{5 : 2}{4 : 1} = \frac{2.5}{4} = 0.625 = \frac{40 \cdot 7}{16 \cdot 28}$$

Odds Ratio: Interpretation

- Wegen der spezifischen Form $\gamma(1, 2|X = 1, X = 2) = (h_{11} h_{22}) / (h_{21} h_{12})$ wird das Odds Ratio auch als **Kreuzproduktverhältnis** bezeichnet. Es gilt

$\gamma = 1$ Odds in beiden Subpopulationen gleich

$\gamma > 1$ Odds in Subpopulation $X = 1$
größer als in Subpopulation $X = 2$

$\gamma < 1$ Odds in Subpopulation $X = 1$
niedriger als in Subpopulation $X = 2$.

- Das Odds Ratio gibt somit an, um welchen Faktor sich die Odds in den beiden Subpopulationen unterscheiden

Odds Ratio: Symmetrie

- Für die Kontingenztafel

h_{11}	h_{12}
h_{21}	h_{22}

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: \vskip -1em

$$\begin{aligned}\gamma(Y = 1, Y = 2 | X = 1, X = 2) &= \gamma(X = 1, X = 2 | Y = 1, Y = 2) \\&= \gamma(Y = 2, Y = 1 | X = 2, X = 1) \\&= \gamma(X = 2, X = 1 | Y = 2, Y = 1) \\&= 1/\gamma(Y = 2, Y = 1 | X = 1, X = 2) \\&= 1/\gamma(X = 2, X = 1 | Y = 1, Y = 2) \dots \text{etc.}\end{aligned}$$

Odds Ratio: Fall - Kontroll - Studie

Beispiel: Morbus Alzheimer und Genetik

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

$$OR = \gamma(\text{ApoE3, ApoE4} | \text{Kontrolle, Fall}) = \frac{2258/803}{593/620} \approx 2.94$$

- ⇒ Odds für ApoE3 bei Kontrollen um den Faktor 3 höher als bei Fällen
- ⇒ Odds für ApoE4 bei Kontrollen um den Faktor 3 niedriger als bei Fällen
- ⇒ ApoE4 Risiko-Faktor für Morbus Alzheimer

Odds Ratio: Fall - Kontroll - Studie

Zentrale Argumentation:

Odds Ratio ist *symmetrisches* Maß

d.h. Chancenverhältnis für Auftreten von ApoE4 bei Kontrolle zu Auftreten von ApoE4 bei Fällen

entspricht

Chancenverhältnis für Ausbleiben von Alzheimer bei ApoE3 zu Ausbleiben von Alzheimer bei ApoE4

⇒ Interpretation als **Risikofaktor** zulässig:

$$\gamma(\text{Kontrolle, Fall} | \text{ApoE3, ApoE4}) = \gamma(\text{ApoE3, ApoE4} | \text{Kontrolle, Fall}) \approx 2.94$$

Verallgemeinerung auf $k \times m$ Kontingenztafeln, Anmerkungen

- ▶ Verallgemeinerung des Verfahrens 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 auch das logarithmierte Odds Ratio verwendet

Anwendung: Apolipoprotein E und Morbus Alzheimer

Etablierter Zusammenhang zwischen Apolipoprotein E ϵ 4 und Morbus Alzheimer

Daten aus Metaanalyse:

ApoE genotype	$\epsilon 2\epsilon 2$	$\epsilon 2\epsilon 3$	$\epsilon 2\epsilon 4$	$\epsilon 3\epsilon 3$	$\epsilon 3\epsilon 4$	$\epsilon 4\epsilon 4$
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

ORs $\gamma(\text{Alzheimer}, \text{Control} | \varepsilon? \varepsilon?, \varepsilon3 \varepsilon3)$,
also immer im Vergleich zu $\varepsilon3 \varepsilon3$ (Referenz):

ApoE genotype	$\varepsilon2 \varepsilon2$	$\varepsilon2 \varepsilon3$	$\varepsilon2 \varepsilon4$	$\varepsilon3 \varepsilon3$	$\varepsilon3 \varepsilon4$	$\varepsilon4 \varepsilon4$
OR (klinisch)	0.99	0.7	1.93	1	2.94	11.1
OR (post mortem)	0.5	0.4	1.4	1	4.5	17.4

Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

Kontingenz- und χ^2 -Koeffizient

Jetzt: Definiere allgemein anwendbares Zusammenhangsmaß für diskrete Merkmale.

Idee:

1. Was wären die gemeinsamen Häufigkeiten \tilde{h}_{ij} bzw. \tilde{f}_{ij} , falls - bei vorgegebenen Randverteilungen - die Merkmale X und Y empirisch unabhängig wären?
2. Quantifizierte "Abstand" der beobachteten gemeinsamen Häufigkeiten h_{ij} bzw. f_{ij} von den unter Unabhängigkeit erwarteten gemeinsamen Häufigkeiten \tilde{h}_{ij} bzw. \tilde{f}_{ij} .

	b_1	...	b_m	
a_1				$h_{1\cdot}$
.				.
a_k				$h_{k\cdot}$
	$h_{\cdot 1}$...	$h_{\cdot m}$	n

Empirische Unabhängigkeit

Idee:

X und Y “empirisch unabhängig” \iff

Verteilung von Y nicht beeinflusst von X (und umgekehrt) \iff

Bedingte relative Häufigkeiten von Y sind in jeder Teilstichprobe $X = a_i$ identisch, d.h. unbeeinflusst von X:

$$f_{Y|X}(b_j|a_1) = f_{Y|X}(b_j|a_2) = \dots = f_{Y|X}(b_j|a_k), \forall j = 1, \dots, m$$

Vergleiche **stoch. Unabhängigkeit**: $A \perp B \iff P(A|B) = P(A)$

Bsp: Empirische Unabhängigkeit

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

$$f_{Y|X}(b_1|a_1) = f_{Y|X}(b_1|a_2) = f_Y(b_1) = \frac{1}{6}$$

$$f_{Y|X}(b_2|a_1) = f_{Y|X}(b_2|a_2) = f_Y(b_2) = \frac{1}{3}$$

$$f_{Y|X}(b_3|a_1) = f_{Y|X}(b_3|a_2) = f_Y(b_3) = \frac{1}{2}$$

Bemerkung: Lokale Odds Ratios sind alle 1

Bsp: Empirische Unabhängigkeit

Wie sehen also die **unter empirischer Unabhängigkeit erwarteten** (absoluten und relativen) Häufigkeiten \tilde{h}_{ij} und \tilde{f}_{ij} aus?

$$\begin{aligned} f_{Y|X}(b_1|a_i) &= f_Y(b_1), \dots, f_{Y|X}(b_m|a_i) = f_Y(b_m), \quad i = 1, \dots, k \\ \iff \frac{\tilde{h}_{ij}}{h_{i\cdot}} &= \frac{h_j}{n}; \forall, i = 1, \dots, k; j = 1, \dots, m \\ \iff \tilde{h}_{ij} &= \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}; \forall, i = 1, \dots, k; j = 1, \dots, m \\ \iff \tilde{f}_{ij} &= f_{i\cdot} \cdot f_{\cdot j}; \forall, i = 1, \dots, k; j = 1, \dots, m \end{aligned}$$

\vskip 2em (\tilde{h}_{ij} sind üblicherweise keine ganzen Zahlen...)

Unabhängigkeitstabelle

Idee:

Vergleiche für jede Zelle (i, j) die *unter Unabh. erwarteten* \tilde{h}_{ij} mit den *tatsächlich beobachteten* h_{ij}

χ^2 -Koeffizient ist bestimmt durch

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

Eigenschaften des χ^2 -Koeffizienten

- $\chi^2 \in [0, n(\min(k, m) - 1)]$
- $\chi^2 = 0 \iff X$ und Y **empirisch unabhängig**
- χ^2 groß \iff starker Zusammenhang
- χ^2 klein \iff schwacher Zusammenhang

\vskip 2em

Schwachpunkt:

Wertebereich von χ^2 hängt vom Stichprobenumfang n und von der Dimension der Tafel ab. Numerischer Wert deshalb schwierig direkt interpretierbar → Normierung

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

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

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

Eigenschaften des (korrigierten) Kontingenzkoeffizienten

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

Beispiel: Nachwahlbefragung 2021

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

	SPD	CDU/CSU	Grüne	FDP	AfD	Linke	Rest	Σ
Männer	649.12 (626)	601.24 (601)	373.80 (352)	291.64 (325)	254.10 (300)	124.60 (125)	209.48 (175)	2504
Frauen	543.88 (567)	503.76 (504)	313.20 (335)	244.36 (211)	212.90 (167)	104.40 (104)	175.52 (210)	2098
Σ	1193	1105	687	536	467	229	385	4602

Unter Unabh. erwartete Häufigkeiten \tilde{h}_{ij} und beobachtete Häufigkeiten h_{ij} (in Klammern)

Interpretation: Wären Geschlecht und Parteipräferenz unabhängig, dann wären 649.12 SPD-wählende Männer in der Stichprobe zu erwarten gewesen, tatsächlich wurden aber nur 626 beobachtet.

Interpretation:

Insgesamt hier

- ▶ $\chi^2 = 43.6$
- ▶ $K = \sqrt{\frac{\chi^2}{\chi^2+n}} = 0.097$ mit $n = 4602$
- ▶ $K^* = \frac{K}{\sqrt{(M-1)/M}} = 0.14$ mit $M = \min(7, 2)$

⇒ keine starke Abhangigkeit zwischen Geschlecht & Parteienprferenz

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

erhält man also unmittelbar

$$\chi^2 = \frac{100(19 \cdot 20 - 18 \cdot 43)^2}{37 \cdot 63 \cdot 62 \cdot 38} = 2.8$$

und $K = 0.17$, $K^* = 0.23$.

Zusammenhangsmaße für diskrete Merkmale

Kontingenztafeln für diskrete und gruppierte Merkmale

Bedingte Häufigkeiten

Diskrete Zusammenhangsanalyse: Odds

Diskrete Zusammenhangsanalyse: Unabhängigkeit & Kontingenz

Multivariate Kontingenztafeln & Mosaikplots

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
- ▶ grafische Darstellung durch Mosaik-Plot

Beispiel: Titanic

```
str(Titanic)
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
## - attr(*, "dimnames")=List of 4
##   ..$ Class    : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex      : chr [1:2] "Male" "Female"
##   ..$ Age      : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

⇒ echt multivariate Kontingenztafeln sind hochdimensionale Arrays

\vskip 1em

```
apply(Titanic, MAR = c(4, 1), FUN = sum)
```

```
##          Class
## Survived 1st 2nd 3rd Crew
##       No 122 167 528 673
##      Yes 203 118 178 212
```

⇒ (Gemeinsame) Randhäufigkeiten durch Akkumulation über restliche Merkmale/Dimensionen

Beispiel: Titanic

Randverteilung von “Klasse”:

```
apply(Titanic, FUN = sum, MAR = 1)
```

```
## 1st 2nd 3rd Crew  
## 325 285 706 885
```

Bedingte Verteilungen von “Überleben” gegeben “Klasse”:

```
apply(Titanic, FUN = sum, MAR = c(1, 4)) / apply(Titanic, FUN = sum, MAR = 1)
```

```
##      Survived  
## Class    No Yes  
##   1st 0.38 0.62  
##   2nd 0.59 0.41  
##   3rd 0.75 0.25  
## Crew 0.76 0.24
```

Beispiel: Titanic

```
apply(Titanic, FUN = sum, MAR = c(2, 4))
```

```
##           Survived  
## Sex      No Yes  
##   Male    1364 367  
##   Female   126 344
```

Bedingte Verteilung:

```
apply(Titanic, FUN = sum, MAR = c(2, 4))/apply(Titanic, FUN = sum, MAR = c(2))
```

```
##           Survived  
## Sex      No Yes  
##   Male    0.79 0.21  
##   Female 0.27 0.73
```

Überlebens-Chance Frauen: $\frac{344}{126} \approx 2.7 \approx 3 : 1$

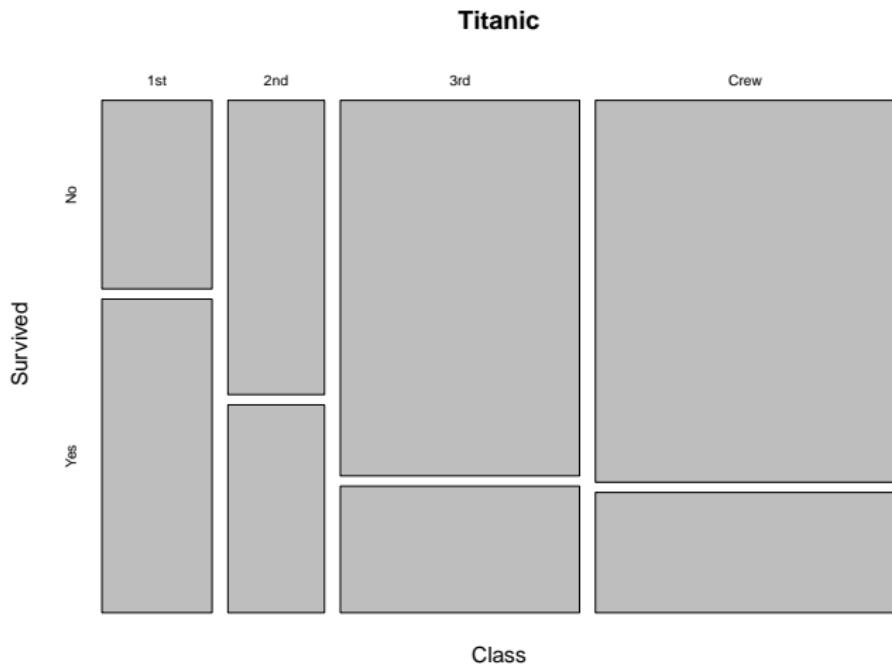
Überlebens-Chance Männer: $\frac{367}{1364} \approx 0.27 \approx 1 : 3$

Chancenverhältnis: 10

Mosaik-Plot

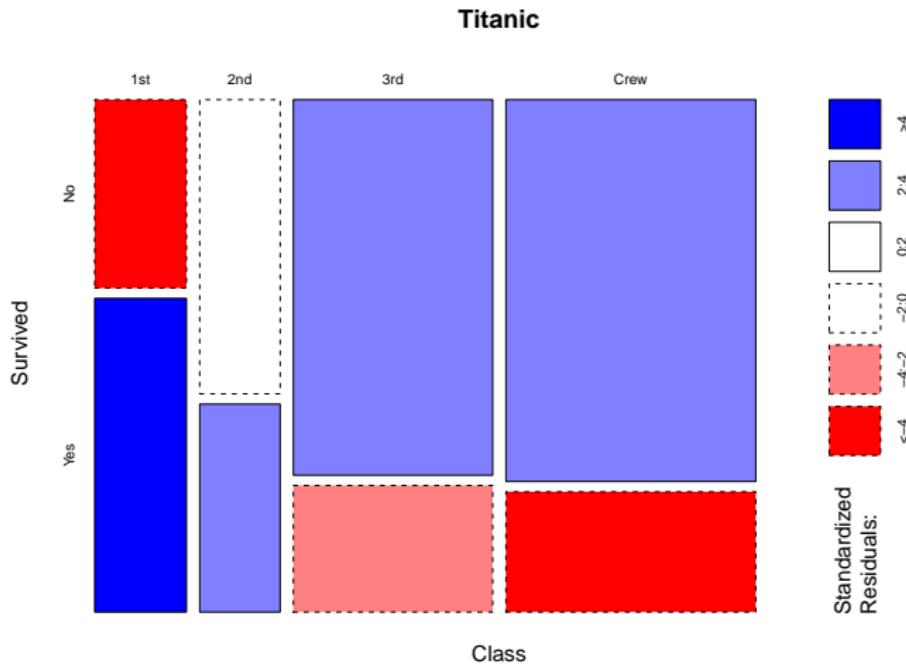
- ▶ Flächentreue Darstellung von gemeinsamen Häufigkeiten
- ▶ Aufteilung schrittweise:
 - Zuerst nach *Einflussgrößen*, dann nach *Zielgröße* aufteilen
- ▶ Gut geeignet für mehrkategoriale ordinale Daten
- ▶ Auch für höhere Dimensionen geeignet (... angeblich)
- ▶ für 2 Dimensionen: entspricht gestapeltem Balkendiagramm mit variabler Balkenbreite

Beispiel: Überlebende bei Titanic

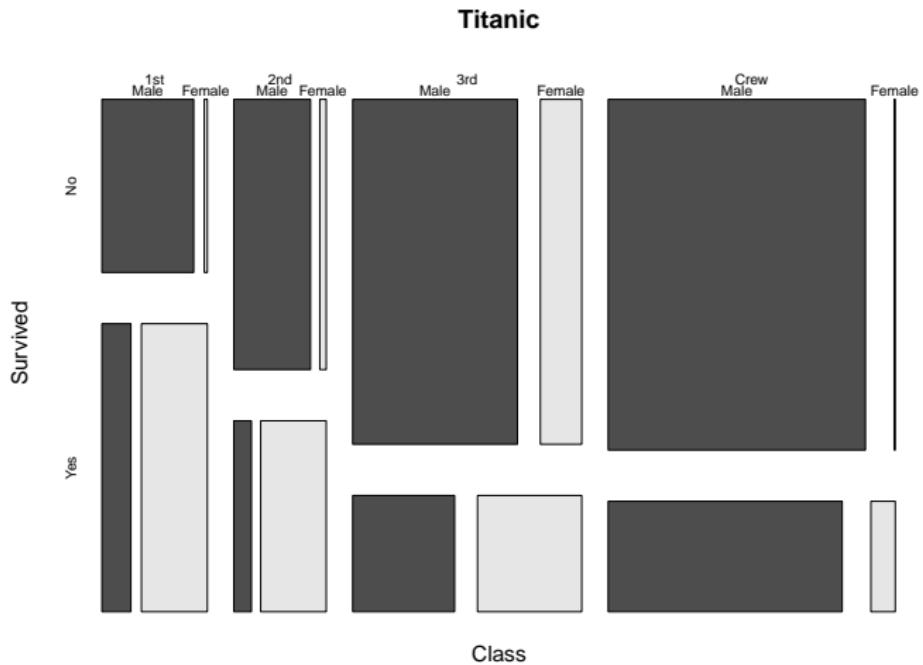


Beispiel: Titanic

Zellen eingefärbt nach $\frac{(h_{ij} - \bar{h}_{ij})}{\sqrt{\bar{h}_{ij}}}$ (*standardized pearson residuals*):



Beispiel: Titanic



Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Zufallsvariablen, Verteilungen & Häufigkeiten

Zufallsvariablen - Motivation

- ▶ Ergebnisse von Zufallsvorgängen sind nicht notwendigerweise Zahlen
- ▶ Oft ist es aber hilfreich diese durch Zahlen zu repräsentieren, um mit ihnen rechnen zu können
- ▶ Beispiel: 3-maliger Wurf einer Münze

$$\Omega = \{Z, K\} \times \{Z, K\} \times \{Z, K\}$$

$$|\Omega| = 2^3 = 8$$

z.B. $\omega = ZKZ := (Z, K, Z)$

Angenommen man interessiert sich für

$$Y := \text{"Anzahl Kopf"}$$

Dann nennt man Y eine **Zufallsvariable** (ZV) mit reellwertigen **Ausprägungen** bzw. **Realisierungen** $y \in T \subseteq \mathbb{R}$.

Man schreibt kurz $Y = y$, wenn die Ausprägung y der ZV Y eingetreten ist.

Zufallsvariable - Definition

Def.: Träger einer ZV

Die Menge der möglichen Ausprägungen einer ZV X heißt *Träger* T_X der ZV X

Im Bsp: $T_Y = \{0, 1, 2, 3\}$

Def.: Zufallsvariable

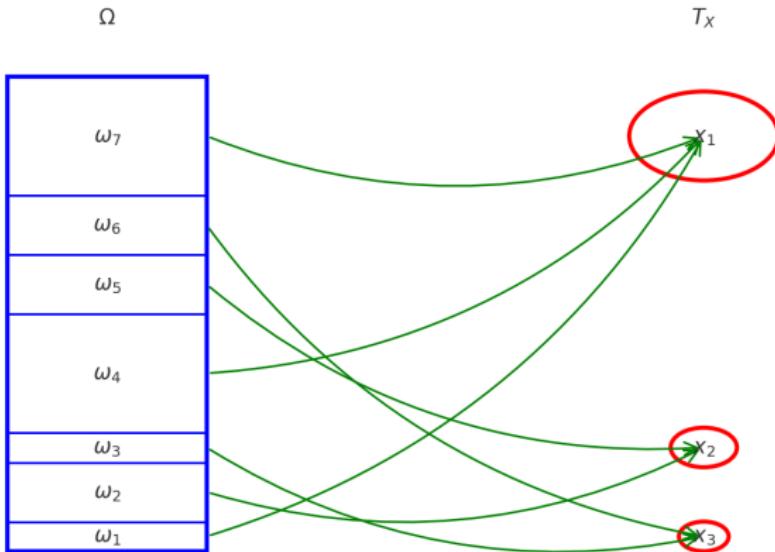
Eine Zufallsvariable X ist eine **eindeutige Abbildung** von Ω nach $T_X \subseteq \mathbb{R}$:

$$X : \Omega \rightarrow T_X$$

- ▶ Eine Zufallsvariable X ordnet also jedem Elementarereignis $\omega \in \Omega$ *genau einen* Zahlenwert $x \in T_X$ zu: $X(\omega) = x$
- ▶ Mehrere Elementarereignisse können dem selben Zahlenwert zugeordnet werden.
- ▶ Unterscheide: Zufallsvariable $X : \Omega \rightarrow T_X$
vs. Ausprägungen/Realisierungen $x = X(\omega)$ von X .

Zufallsvariable - Diagramm

Zufallsvariable $X(\omega)$: Abbildung von Ω auf T_X



Zufallsvariable - Motivation

- ▶ Man kann mit X "rechnen": z.B $P(X \leq a)$ oder $P(X^2 > b)$, oder "Welches Ergebnis erwarten wir 'im Mittel'?"
- ▶ Ursprünglicher Wahrscheinlichkeitsraum (Ω, P) wird letztendlich nicht mehr benötigt, stattdessen nur Wahrscheinlichkeiten für die Werte der ZV
 - ⇒ meist deutlich einfacher zu handhaben
 - im Bsp: Verteilung über $\{0, 1, 2, 3\}$ statt über $\{\text{KKK}, \text{KKZ}, \text{KZK}, \dots, \text{ZZK}, \text{ZZZ}\}$
- ▶ Mathematische Formalisierung eines (quantitativen) Messvorgangs – jeder möglichen Beobachtung (= Elementarereignis) wird genau ein Zahlenwert zugeordnet.
 - in vielen Anwendungen Ω kaum zugänglich oder mathematisch formalisierbar und nur Ergebnis des Messvorgangs überhaupt *beobachtbar*
 - ⇒ ZV zentraler Begriff für statistische Anwendungen und Ansatzpunkt für mathematische Formalisierung empirischer Phänomene

Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Diskrete Zufallsvariable

Def.: Diskrete Zufallsvariable

Eine ZV X heißt **diskret**, falls sie nur endlich oder abzählbar unendlich viele Werte x_1, x_2, \dots annehmen kann, also: falls die Menge $T = \{x_1, x_2, \dots\}$ der möglichen Ausprägungen x_i von X mit $P(X = x_i) > 0$ abzählbar ist.

Def.: Wahrscheinlichkeitsfunktion einer diskreten ZV

Die Wahrscheinlichkeitsfunktion von X ist durch

$$f_X(x_i) := P(X = x_i) = P(\{\omega \in \Omega : X(\omega) = x_i\})$$

für $x_i \in \mathbb{R}$ gegeben.

- Alternativ: "Wahrscheinlichkeitsdichte"

Folgerungen

Als Funktion von $B \subset \mathbb{R}$ ist also

$$P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

eine Wahrscheinlichkeitsverteilung auf \mathbb{R} .

Man nennt diese die **Verteilung** der Zufallsvariable X .

Sie ist durch die Abbildung X und die Wahrscheinlichkeitsverteilung P über $\omega \in \Omega$ induziert.

Für $x \notin T_X$ ist $f_X(x) = P(\emptyset) = 0$.

Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Die Verteilungsfunktion

Def.: Verteilungsfunktion (diskret)

Die *Verteilungsfunktion* einer diskreten ZV ist definiert als

$$F_X(x) := P(X \leq x) = \sum_{i: x_i \leq x} f_X(x_i).$$

Kennt man also die **Wahrscheinlichkeitsfunktion** $f(x)$ für alle $x \in T$, so kennt man auch die **Verteilungsfunktion** $F(x)$ und umgekehrt.

Eigenschaften der Verteilungsfunktion

- ▶ $F(x)$ ist monoton wachsend (“Treppenfunktion”)
- ▶ $F(x)$ ist stückweise konstant mit Sprungstellen an Werten x_i mit $f(x_i) > 0$, d.h. an allen Realisierungen $x_i \in T$
- ▶ Die Höhe des Sprungs an der Stelle $x_i \in T$ ist $f(x_i)$
- ▶ $\lim_{x \rightarrow \infty} F(x) = 1$
- ▶ $\lim_{x \rightarrow -\infty} F(x) = 0$

Beispiel: 3-maliger Münzwurf

$X := \text{"Anzahl Kopf"}, T = \{0, 1, 2, 3\}$

$$\begin{array}{lll} f(0) &= P(\{\text{ZZZ}\}) &= 1/8 \\ f(1) &= P(\{\text{KZZ}, \text{ZKZ}, \text{ZZK}\}) &= 3/8 \\ f(2) &= P(\{\text{KKZ}, \text{KZK}, \text{ZKK}\}) &= 3/8 \\ f(3) &= P(\{\text{KKK}\}) &= 1/8 \end{array}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/8 & 0 \leq x < 1 \\ 4/8 & 1 \leq x < 2 \\ 7/8 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Beachte: $f(x) = 0$ für alle $x \notin T$

Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Stetige Zufallsvariable - Definition 1

Def.: Stetige Zufallsvariable

Eine Zufallsvariable X ist **stetig**, falls ihr Träger eine überabzählbare Teilmenge der reellen Zahlen \mathbb{R} ist.

Beispiel:

Glücksrad mit stetigem Wertebereich $[0^\circ, 360^\circ)$

Von Interesse also die Zufallsvariable, die den *exakten* Winkel angibt, an dem das Glücksrad stehen bleibt.

Stetige Zufallsvariablen - Definition 2

Def.: Stetige Zufallsvariable

Eine Zufallsvariable X heißt *stetig*, wenn es eine Funktion $f(x) \geq 0$ gibt, so dass sich die Verteilungsfunktion $F_X(x) := P(X \leq x)$ von X wie folgt darstellen lässt:

$$F_X(x) = \int_{-\infty}^x f_X(u)du$$

Def.: Dichtefunktion einer (stetigen) ZV

Diese nicht-negative Funktion $f(x)$ zur Verteilungsfunktion $F(x)$ heißt *Wahrscheinlichkeitsdichte* (auch: *Dichte* oder *Dichtefunktion*) von X .

Beachte: bei *diskreten* Zufallsvariablen gilt $F(x) = \sum_{i:x_i \leq x} f(x_i)$.

Folgerungen

- ▶ $\int_{-\infty}^{+\infty} f(x) dx = 1$
- ▶ Mit Dichte- bzw. Verteilungsfunktion lassen sich W.keiten für beliebige Teilmengen des Trägers ausrechnen:
 $P(X \in [a, b]) = \int_a^b f(x) dx = F(b) - F(a)$
- ▶ Unintuitive Konsequenz:
für stetige ZV X gilt $P(X = x) = 0 \forall x \in \mathbb{R}$!

Video dazu von 3blue1brown

Folgerungen

- ▶ $\lim_{x \rightarrow -\infty} F(x) = 0$
- ▶ $\lim_{x \rightarrow \infty} F(x) = 1$
- ▶ $F'(x) = f(x)$ überall dort wo $f(x)$ stetig ist.
- ▶ $P(a \leq X \leq b) = F(b) - F(a)$
- ▶ $P(X > a) = 1 - F(a)$

Ausblick

Stringentere Definitionen in späteren Vorlesungen zu W.keitstheorie

Eine allgemein gültige Definition ohne Fallunterscheidungen zwischen

- ▶ diskreten ZVn,
- ▶ stetigen ZVn,
- ▶ und ZVn, deren Träger teils stetig und teils diskret ist,

benötigt Maßtheorie mit einem verallgemeinertem Integrationsbegriff.

Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Empirische Häufigkeitsverteilungen

Zurück zur Empirie

Zufallsvariablen sind Bestandteil der mathematischen Formalisierung eines zufälligen Prozesses.

Jetzt: Betrachte *empirische* Entsprechungen der eben eingeführten Begriffe.

Häufigkeitsverteilung: Notation & Terminologie

Im Weiteren:

- ▶ X, Y, \dots Bezeichnung für **Merkmale**
- ▶ n ist Anzahl der **Untersuchungseinheiten**
- ▶ $x_i, i \in \{1, \dots, n\}$ ist **beobachteter Wert** bzw. **Merkmalsausprägung** von Merkmal X für i -te Untersuchungseinheit
- ▶ x_1, \dots, x_n **Rohdaten, Urliste**
- ▶ $a_1, a_2, \dots, a_k, k \leq n$ (evtl. der Größe nach geordnete) *verschiedene Werte* der Urliste x_1, \dots, x_n : die **Menge der beobachteten Merkmalsausprägungen** von X .

Eindimensionale Häufigkeitsverteilung

- ▶ Sortierung der Daten nach einem Merkmal
- ▶ Auszählen der Häufigkeiten der einzelnen Merkmalsausprägungen
- ▶ **Relative Häufigkeiten** = Häufigkeit einer Merkmalsausprägung / Anzahl der Untersuchungseinheiten (Anteil)
- ▶ **Kumulative relative Häufigkeiten** bei ordinal oder metrisch skalierten Merkmalen sinnvoll:
 $F(x) := \text{"Anteil UE mit Merkmalsausprägung } \leq x\text{"}$ heißt **empirische Verteilungsfunktion**

Häufigkeitsverteilungen

Bemerkungen:

- ▶ Für Nominalskalen hat die Anordnung “ $<$ ” keine inhaltliche Bedeutung.
- ▶ Bei kategorialen Merkmalen $k = \text{Anzahl der Kategorien}$
- ▶ Bei stetigen Merkmalen 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, \dots, x_n mit $x_i = a_j$

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

h_1, \dots, h_k **absolute Häufigkeitsverteilung**

f_1, \dots, f_k **relative Häufigkeitsverteilung**

Bemerkungen:

- ▶ Wenn statt der Urliste bereits die Ausprägungen a_1, \dots, a_k und ihre Häufigkeiten f_1, \dots, f_k bzw. h_1, \dots, h_k vorliegen, sprechen wir von **Häufigkeitsdaten**.
- ▶ **Klassenbildung, gruppierte Daten:** Vor allem bei metrischen, stetigen (oder quasi-stetigen) Merkmalen oft Aggregation der Urliste durch Bildung geeigneter Klassen

Beispiel Nettomieten I

Wir greifen aus dem gesamten Datensatz des Münchner Mietspiegels 2015 die Wohnungen ohne zentrale Warmwasserversorgung (`zh0 == 1`) und mit einer Wohnfläche kleiner als $60m^2$ (`wfl < 60`) heraus.

Die folgende Urliste zeigt, bereits der Größe nach geordnet, die Nettomieten dieser $n = 88$ Wohnungen:

```
mietspiegel_url <- "http://chris.userweb.mwn.de/statistikbuch/mietspiegel2015.txt"
mietspiegel <- read.table(file = mietspiegel_url, header = TRUE)
klein_und_kalt <- subset(mietspiegel, zh0 == 1 & wfl < 60)
sort(klein_und_kalt[, "nm"]) |>
  print(digits = 4)

## [1] 174.8 175.0 211.0 211.9 212.9 245.0 257.2 266.8 276.6 278.1 282.4 287.4
## [13] 294.9 299.5 302.9 308.4 315.7 318.0 322.0 323.0 330.0 336.0 336.0 338.5
## [25] 339.7 343.0 346.5 362.0 363.5 370.0 375.1 379.6 380.0 390.4 391.4 395.1
## [37] 408.9 410.0 415.6 421.8 437.7 440.0 440.0 443.6 446.6 451.6 452.2 461.9
## [49] 470.0 472.4 480.8 490.0 495.0 500.0 510.0 513.0 515.0 515.0 517.1 518.0
## [61] 520.2 521.6 523.0 526.5 540.0 543.0 550.0 551.0 556.0 570.0 570.0 573.5
## [73] 576.0 590.0 590.0 590.0 600.7 610.0 610.0 630.0 638.0 660.0 695.0 700.0
## [85] 700.0 726.9 730.0 790.0
```

Alle Werte verschieden:

$$\Rightarrow k = n \text{ und } \{x_1, \dots, x_n\} = \{a_1, \dots, a_k\}; f_j = \frac{1}{88} \quad \forall j = 1, \dots, 88.$$

Beispiel Nettomieten II

Gruppert man die Urliste in 7 Klassen mit gleicher Klassenbreite von 100€, so erhält man folgende Häufigkeitstabelle:

```
gruppierung <- seq(from = 150, to = 850, by = 100)
klein_und_kalt[, "nm_gruppiert"] <- cut(klein_und_kalt[, "nm"], breaks = gruppierung)
table(klein_und_kalt[, "nm_gruppiert"])
```

```
## 
## (150,250] (250,350] (350,450] (450,550] (550,650] (650,750] (750,850]
##      6       21       18       22       14       6       1
```

Klasse	absolute H.keit	relative H.keit	kumulative rel. H.keit
(150,250]	6	0.07	0.07
(250,350]	21	0.24	0.31
(350,450]	18	0.20	0.51
(450,550]	22	0.25	0.76
(550,650]	14	0.16	0.92
(650,750]	6	0.07	0.99
(750,850]	1	0.01	1.00

Zufallsvariablen, Verteilungen & Häufigkeiten

Diskrete Zufallsvariablen

Verteilungsfunktion

Stetige Zufallsvariablen

Empirische Häufigkeitsverteilungen

Empirische Verteilungsfunktion

Empirische Verteilungsfunktion

Häufigkeitsfunktion:

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

Verteilungsfunktion:

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

bzw.

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

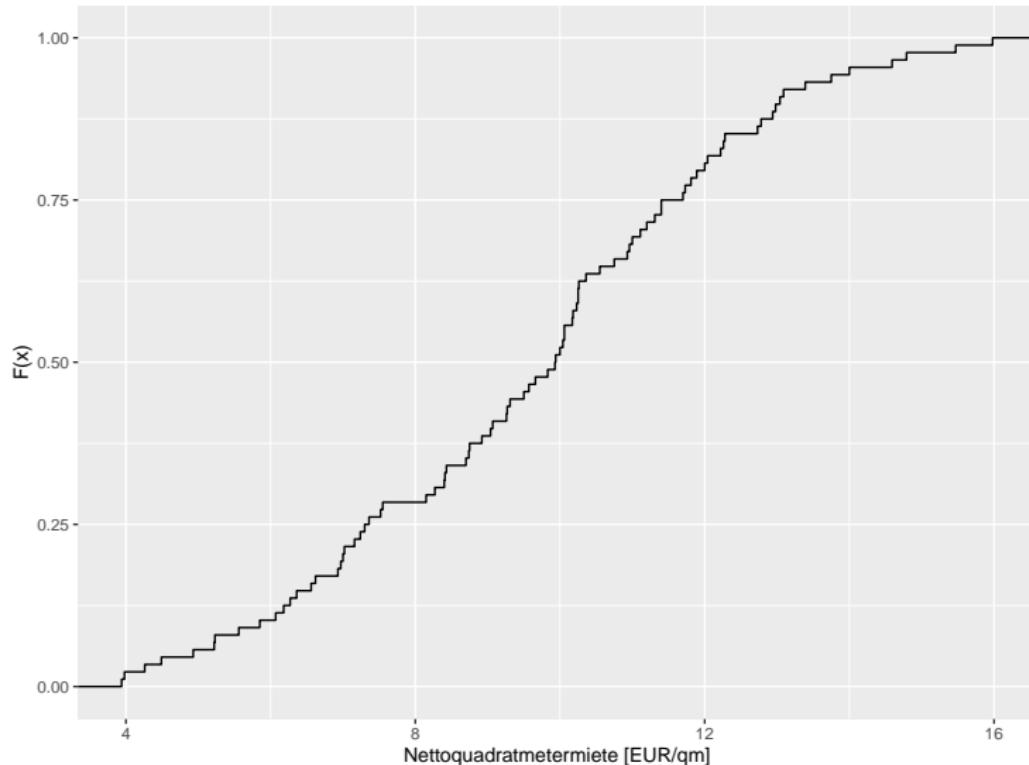
für $a_j \leq x$ und $a_{j+1} > x$.

ECDF (empirical cumulative distribution function)

Eigenschaften der ECDF $F_n(x)$

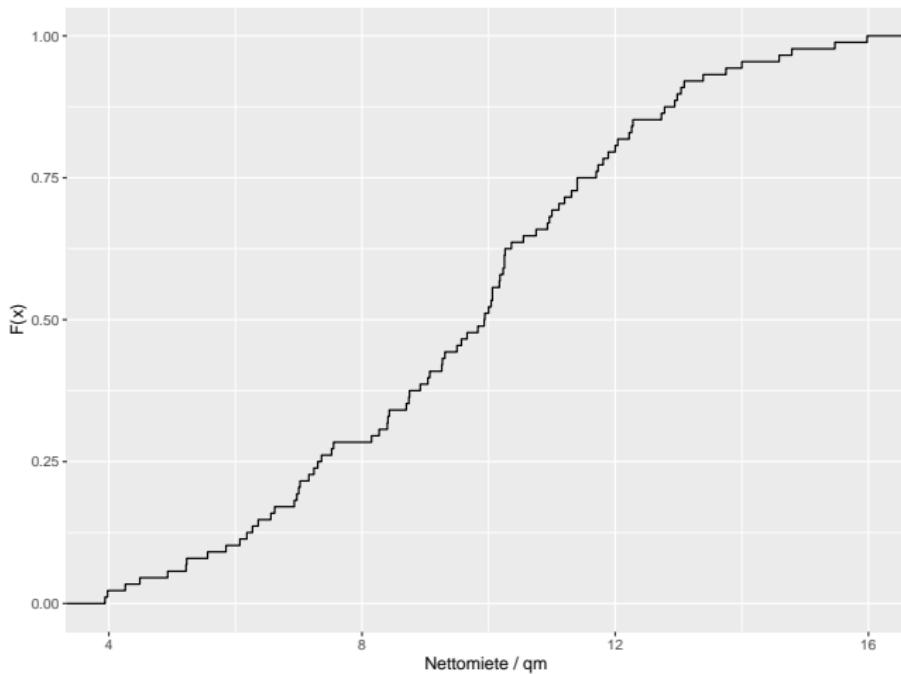
- ▶ monoton wachsende Treppenfunktionen mit Sprüngen an den Ausprägungen a_1, \dots, a_k
- ▶ Sprunghöhen: f_1, \dots, f_k
- ▶ rechtsseitig stetig
- ▶ $F_n(x) = 0$ für $x < a_1$, $F_n(x) = 1$ für $x \geq a_k$

Beispiel: F_n (Quadratmetermiete) für Klein & Kalt



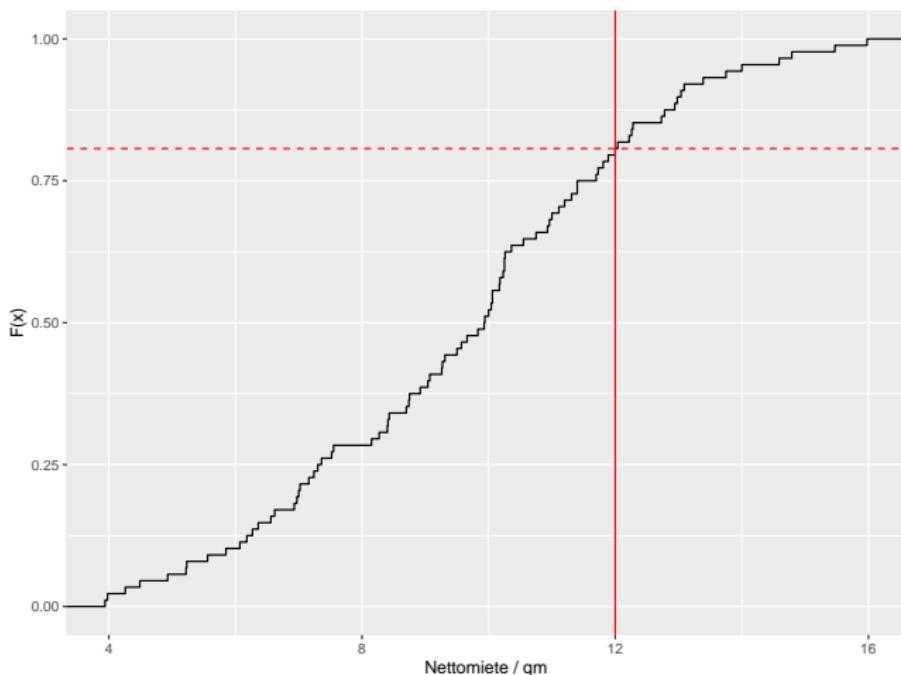
Beispiel: F_n (Quadratmetermiete) für Klein & Kalt

Wie groß ist der Anteil von kleinen Wohnungen ohne ZH mit $\text{qm-Miete} \leq 12$ €? Wie viel kosten die günstigsten 25% der Wohnungen höchstens?



Beispiel: F_n (Quadratmetermiete) für Klein & Kalt

Wie groß ist der Anteil von kleinen Wohnungen ohne ZH mit qm-Miete ≤ 12 €?

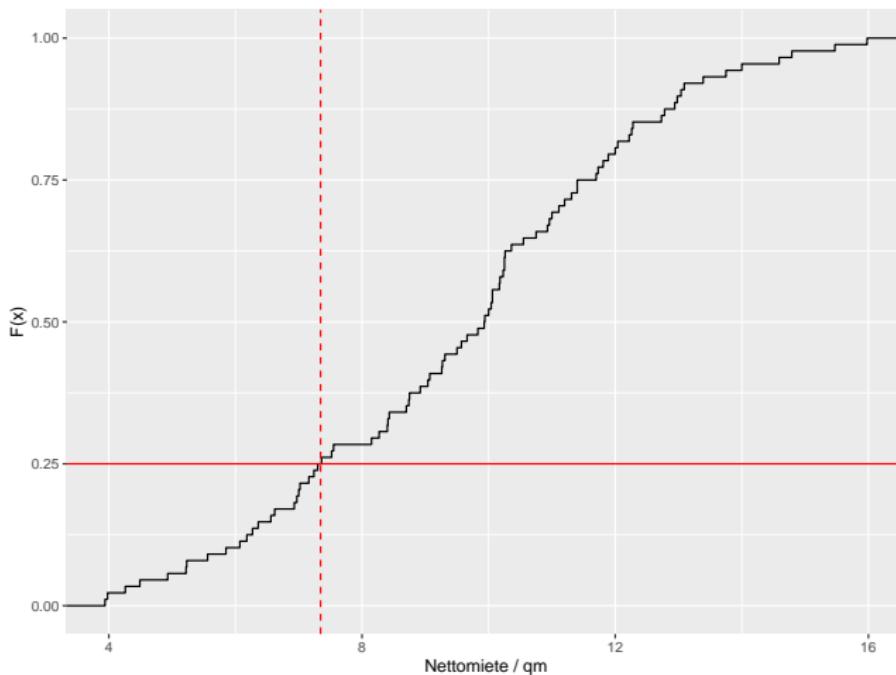


$\Rightarrow \approx 80\%$

Beispiel: F_n (Quadratmetermiete) für Klein & Kalt

Wie viel kosten die günstigsten 25% der Wohnungen höchstens?

Was ist die maximale qm-Miete im billigsten unteren Viertel der Wohnungen?



$\Rightarrow \approx 7.3\text{€}$

Theorie & Empirie

Theorie		Empirie
Zufallsvariable X	\equiv	Merkmal X
Träger T_X	\supseteq	Beobachtete Merkmalsausprägungen $\{a_1, \dots, a_k\}$ von X
W.keitsfunktion $f_X(x)$	\leftrightarrow	relative H.keiten $f_X(a_j)$
Verteilungsfunktion $F_X(x)$	\leftrightarrow	kumul. relative H.keiten / empirische Verteilungsfunktion $F_n(x)$

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Statistische Grafik

Statistische Grafiken

repräsentieren die *Ausprägungen* gewisser *Merkmale* in einem *Datensatz* durch die *ästhetischen Eigenschaften geometrischer Objekte*.

Aus dieser Definition können wir eine formale “Grammatik” für grafische Darstellungen ableiten.

(s.a. Wilkinson (2005) *The Grammar of Graphics*,

Wickham et al. (2020) *ggplot2: Elegant Graphics for Data Analysis*)

Grammar of Graphics

(Fast) jede sinnvolle Grafik lässt sich als Kombination der folgenden Basis-Elemente beschreiben:

- ▶ *Daten*: enthalten die *Beobachtungen* der dargestellten *Merkmale* (Oft auch: daraus abgeleitete *Statistiken* wie Anteile, Mittelwerte, etc)
- ▶ *Geometrische Elemente*: z.B. Punkte, Linien, Rechtecke, etc...
- ▶ *Ästhetische Zuordnungen*: die *Ausprägungen* der in der Grafik dargestellten *Merkmale* werden durch sichtbare Eigenschaften der geometrischen Elemente repräsentiert, z.B.
 - ▶ Position (vertikal/horizontal)
 - ▶ Farbe
 - ▶ Größe
 - ▶ Form

Grammar of Graphics

Also: Daten werden über ästhetische Eigenschaften von geometrischen Elementen dargestellt.

zusätzlich:

- ▶ *Datentransformationen/Statistiken*: Oft werden nicht Rohdaten selbst, sondern daraus abgeleitet Größen (Mittelwerte, Anteile,) abgebildet. Komplexere geometrische Elemente basieren oft intern auf Datentransformationen (→ Balkendiagramme, geglättete Trendlinien, Boxplots,).
- ▶ *Skalen* ordnen für ein gegebenes *Merkmal/Statistik* und ihrer ästhetischen Zuordnung jeder Ausprägungen bestimmte Werte zu (z.B. Achsenabschnitte für Position, Farbpaletten für Farbe, etc...) und legen damit auch Legenden (z.B. für Farbskalen) und Achsenbeschriftungen fest.
- ▶ *Koordinatensysteme*: kartesische / logarithmische / Polarkoordinatensysteme; Kartenprojektionen

Grammar of Graphics

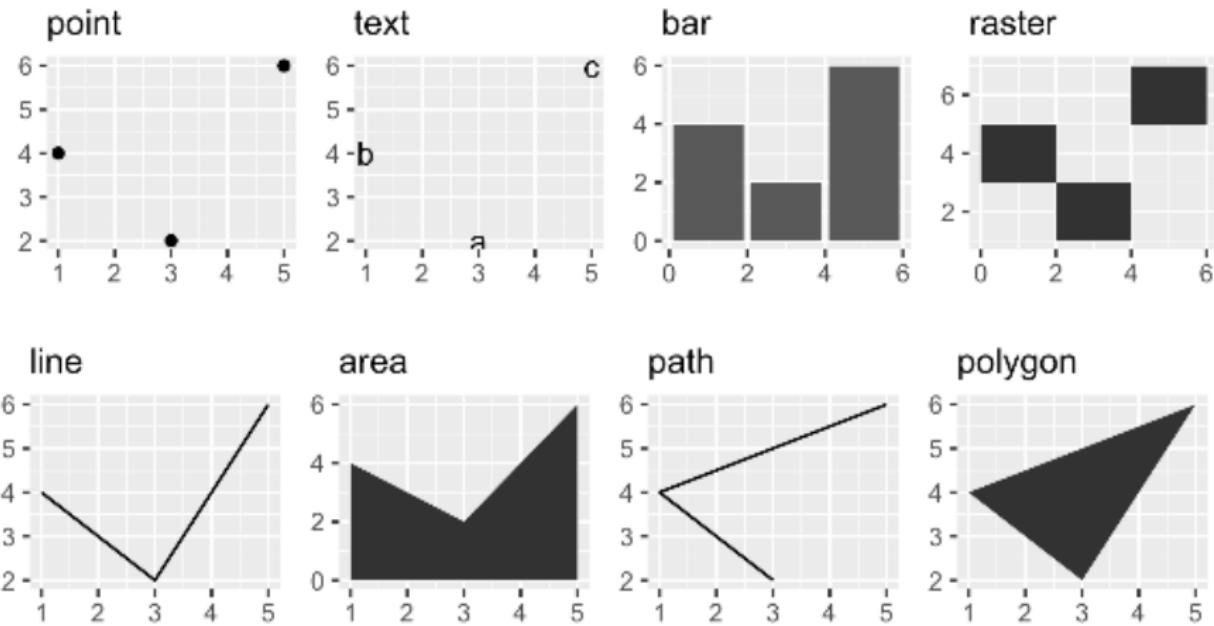
zusätzlich:

- ▶ *Facettierung* definiert anhand welcher Merkmale die Daten in Subgruppen aufgeteilt werden und wie die Darstellungen der verschiedenen Subgruppen arrangiert werden (s.a.: *small multiples plot, lattice plots*).
- ▶ *Theme*: umfasst sämtliche Designaspekte wie Fonteigenschaften, Gitterlinien, Hintergrundfarben, Layout von Textelementen und Legenden, ...

Implementation in R: `ggplot2`

Python: `plotnine`

Beispiele: Geometrien

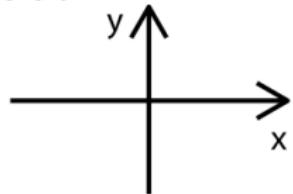


.. und viele viele mehr ...

Abb: Wickham et al. (2020)

Beispiele: Ästhetiken

position



shape



size



color



line width



line type

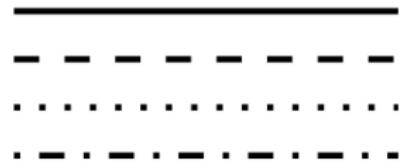


Abb: Wilke (2020)

Grammar of Graphics: Beispiele mit Code

Aden-Buie, 2018

Reynolds, 2019

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Wahrnehmung von Grafiken

Wahrnehmungspsychologische Experimente zeigen deutliche Unterschiede in der korrekten Interpretation der folgenden Grafiktypen:

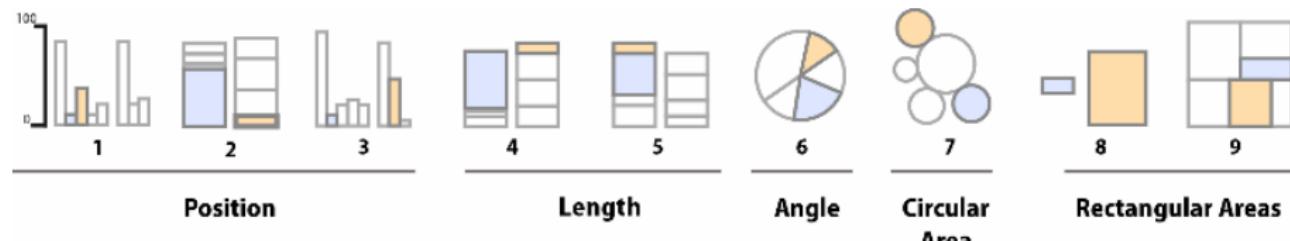


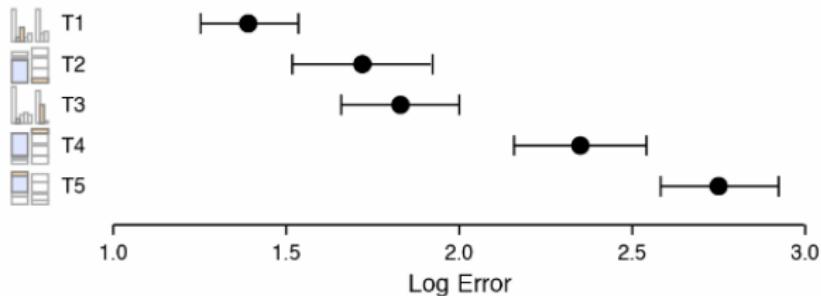
Abb.: Healy, K. (2018). *Data visualization: a practical introduction*. Princeton University Press. (übernommen aus Heer & Bostock)

\vskip 3em Cleveland, W.S., McGill, R. (1984): Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *JASA* 79(387), 531–554.

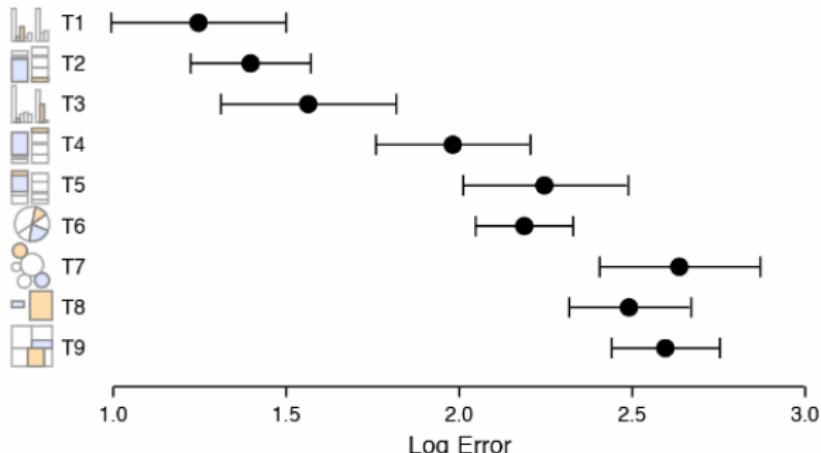
Heer, J., Bostock, M. (2010): Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design. *SIGCHI Proceedings*, 203–212.

Wahrnehmung von Grafiken

Cleveland & McGill's Results



Crowdsourced Results



Wahrnehmung von Grafiken

Hierarchie der korrekten Interpretation:

1. Position (gemeinsame/parallele Skala > verschobene Skalen)
2. Abstände & Längen
3. Steigung & Winkel
4. Flächen
5. Volumen
6. Farbton-Farbsättigung-Farbhelligkeit

Da **Position & Abstände** am besten wahrgenommen werden, müssen diese verwendet werden um die wichtigsten Merkmale zu codieren.

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*.

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

Goldene Regeln für Grafikgestaltung

- Verständnis von Grafiken erfordert:
 - *Detection*: Erkennen welche geometrischen Elemente und ästhetische Eigenschaften welche Werte repräsentieren
 - *Synthese*: Gruppierung & In-Relation-Setzen der entdeckten Inhalte
 - *Evaluation*: Bewertung der relativen Größen und Bedeutung dieser Inhalte.
Speziell: Unterscheidung, Rangbildung, Abschätzung von Verhältnissen.
- alle drei Phasen unterstützt durch *bewusste Wahl*
 - geeigneter Ästhetiken & Geometrien (s.o.: Abstände > Farbe, etc.)
 - geeigneter Achsen, Gitterlinien und Seitenverhältnisse
 - inhaltlich adäquater Sortierung/Reihenfolge der Ausprägungen qualitativer Merkmale
 - entsprechender Annotationen & Hervorhebungen

⇒ **Kommunikationsabsicht** klarmachen:

Welche Informationen will ich primär vermitteln? Welche zusätzlich? Auf welche Vergleiche soll Aufmerksamkeit gelenkt werden?

⇒ **Lesbarkeit** maximieren – sowohl Genauigkeit als auch Geschwindigkeit/Schwierigkeit sind wichtig.

Harrell, F.E. (2017), *Principles of Graph Construction*;

Grolemund, G., Wickham, H. (2017) *R for Data Science*, Ch. 28; Rauser, J. (2016) *How Humans See Data*

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

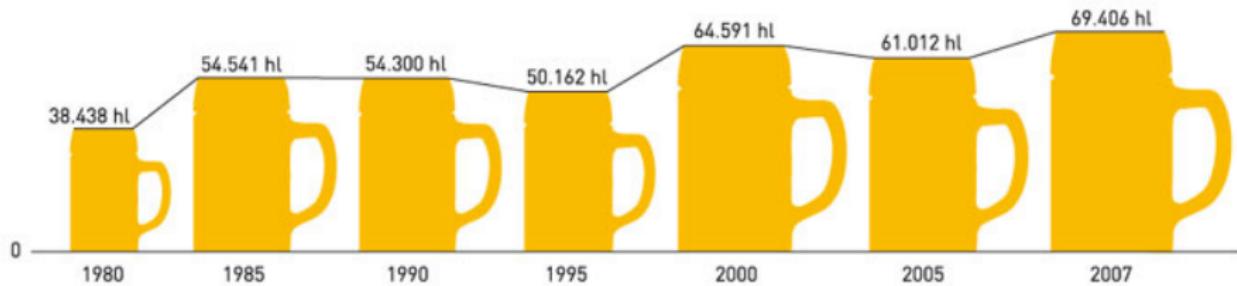
Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

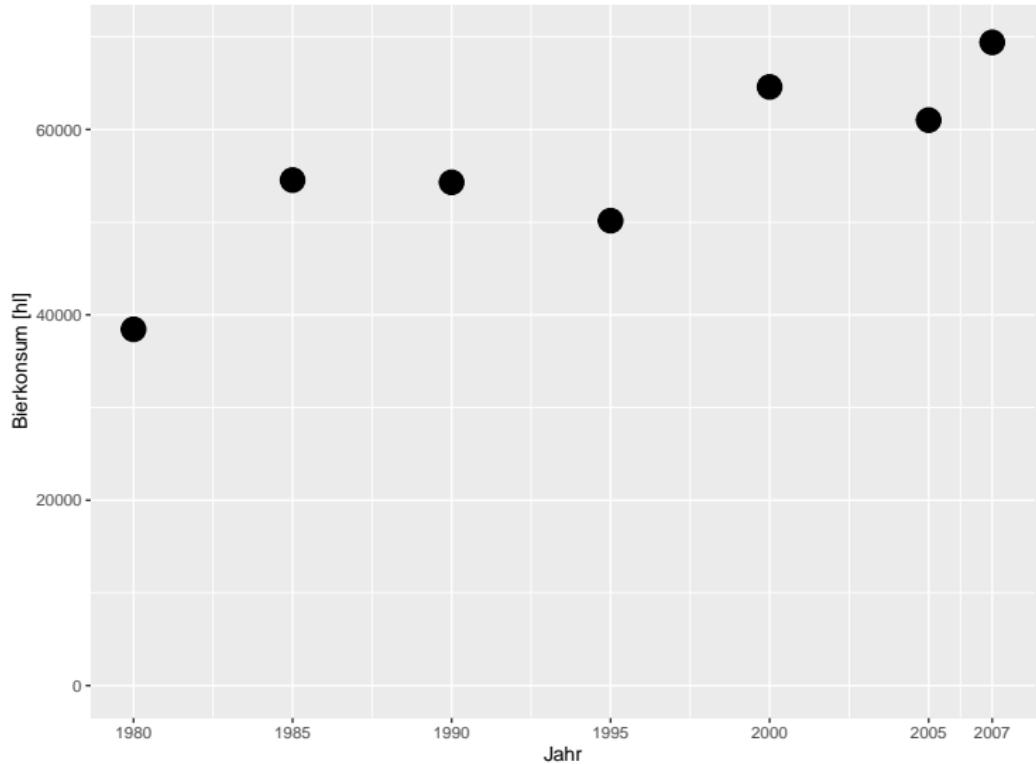
Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

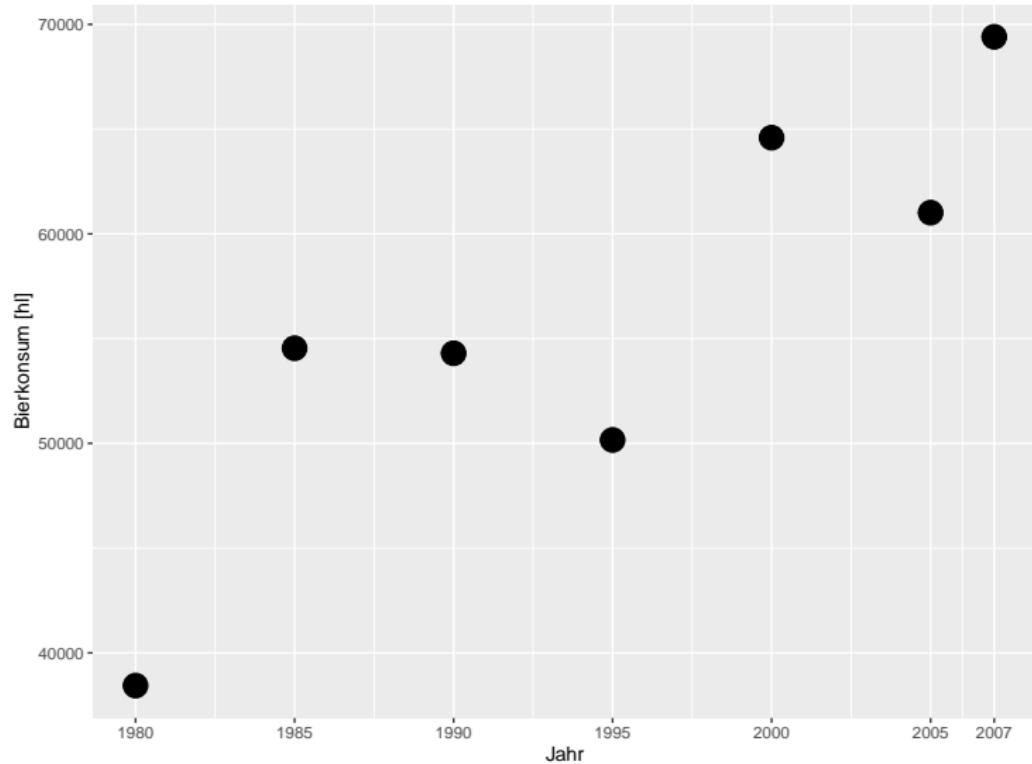
“Infoviz” vs Statistische Grafiken I



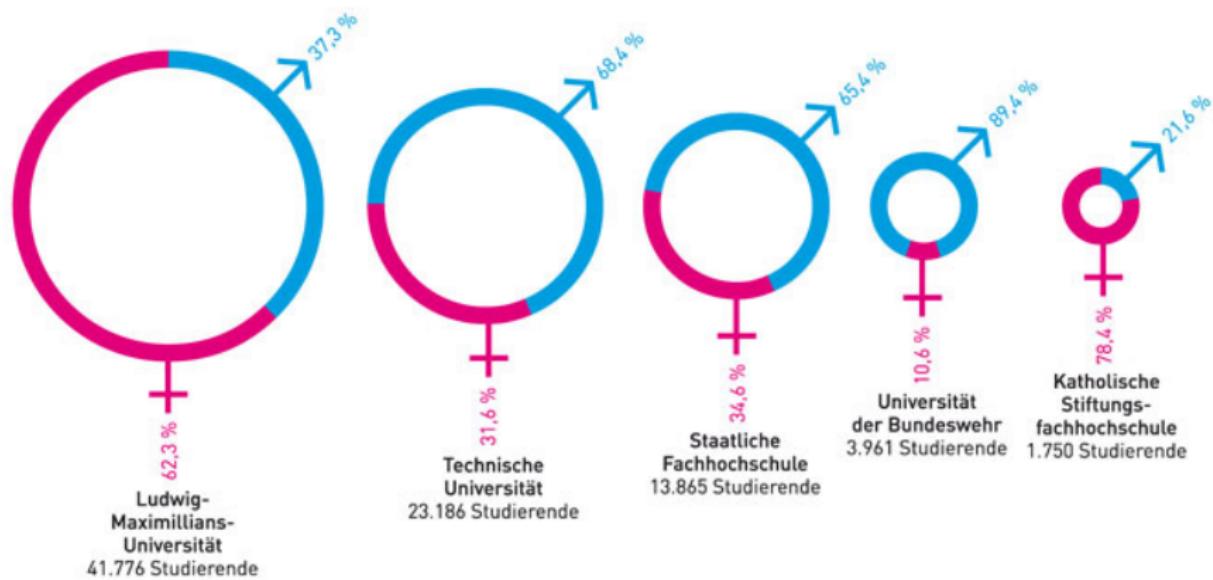
“Infoviz” vs Statistische Grafiken I



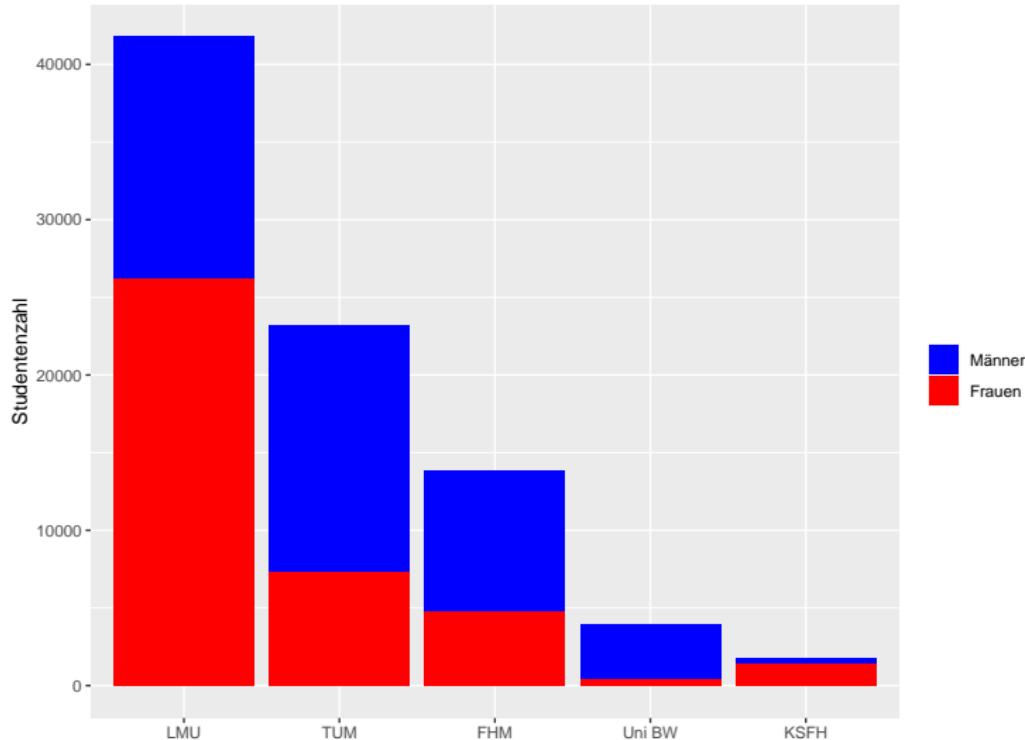
“Infoviz” vs Statistische Grafiken I



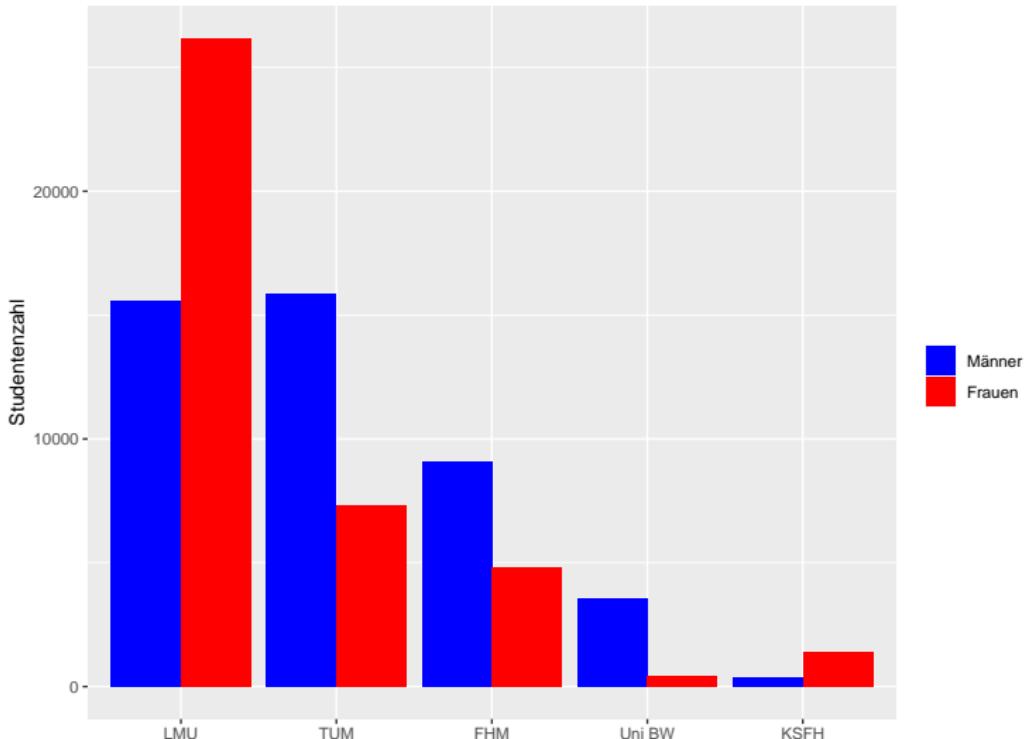
“Infoviz” vs Statistische Grafiken II



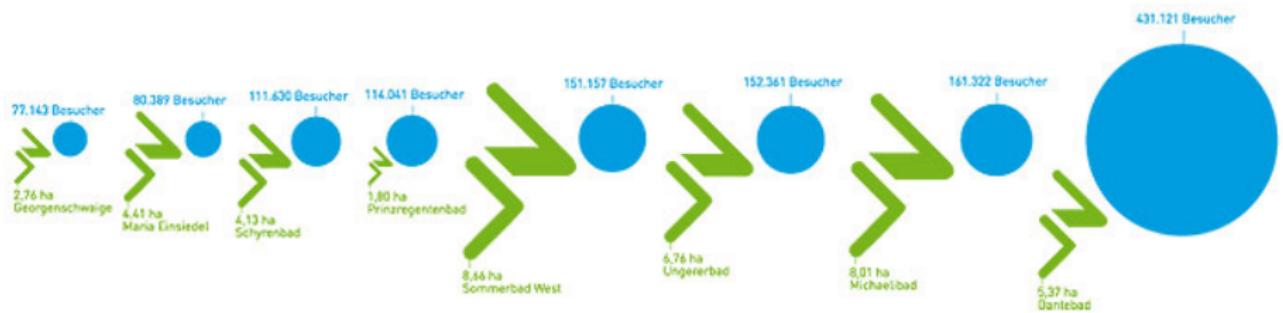
“Infoviz” vs Statistische Grafiken II



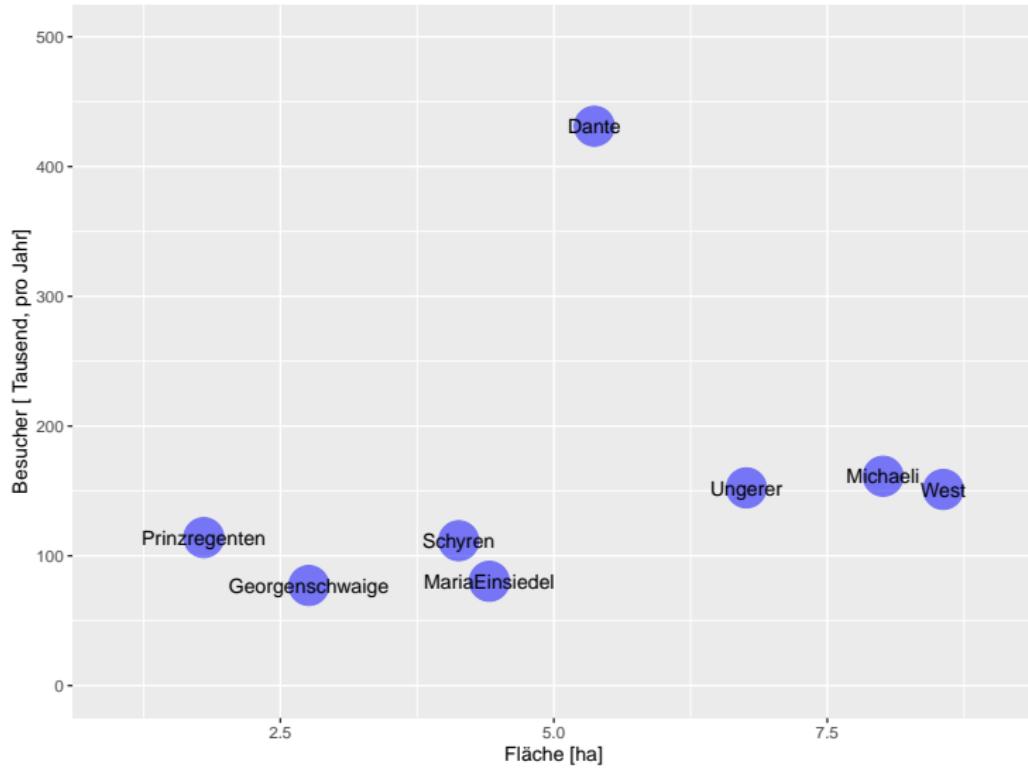
“Infoviz” vs Statistische Grafiken II



“Infoviz” vs Statistische Grafiken IV



“Infoviz” vs Statistische Grafiken IV



Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Wahrnehmung Ästhetischer Zuordnungen

Kodieren von Information durch:

- ▶ **Position, Längen & Abstände:** am einfachsten zu erkennen, v.a. für gemeinsame/parallele Skalen
- ▶ **Winkel & Steigung:** visueller Eindruck abhängig von Richtung & vom Seitenverhältnis der Darstellung
- ▶ **Flächen:**
 - ▶ lang/dünn erscheint größer als kompakt/konvex
 - ▶ abhängig von Farbe: hellere Flächen wirken größer
- ▶ **Farbe:** oft schwierig präzise Vergleiche abzulesen
 - ▶ dennoch allgegenwärtig da *kombinierbar* mit anderen grafischen Elementen.

Farbwahrnehmung

- ▶ Farbwahrnehmung ist komplex
- ▶ Farben für Grafiken so wählen dass **Wahrnehmungseinheitlichkeit** berücksichtigt wird:
 - ▶ Grün-/Gelbtöne werden bei gleicher Sättigung intensiver wahrgenommen als andere Farben
 - ▶ Sehschwächen mitbedenken

Farbwahrnehmung

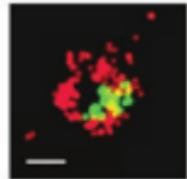
Rot-Grün-Sehschwäche

(Protanopia: Prävalenz 6%, Deutanopia: Prävalenz 2%)

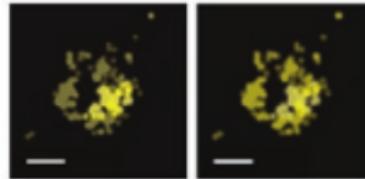
a

Original image
with red and
green color
coding

Natural color
images



Simulated colors as seen by:
protanope deutanope



b

Image with red
replaced by
magenta

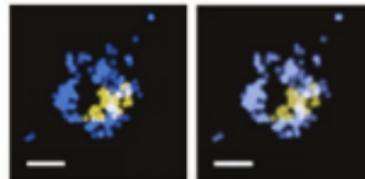
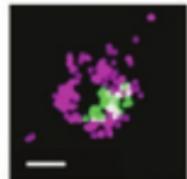
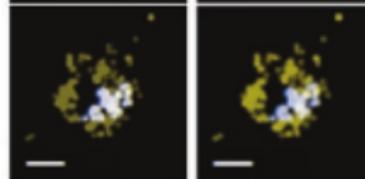
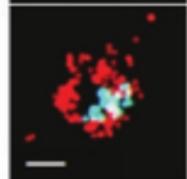
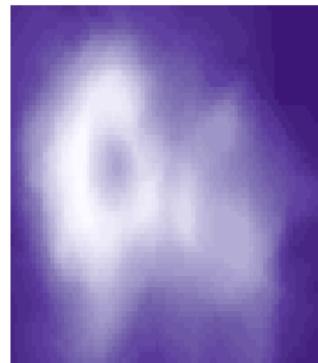
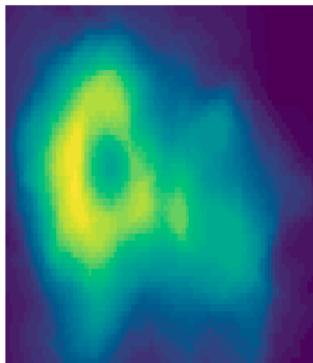
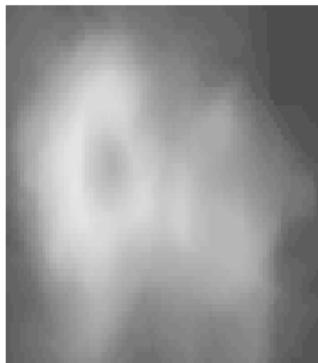
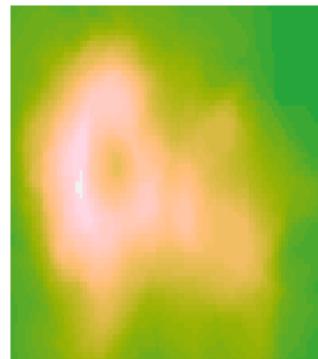
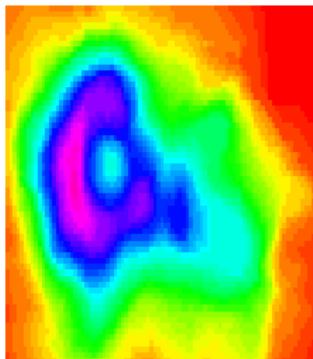
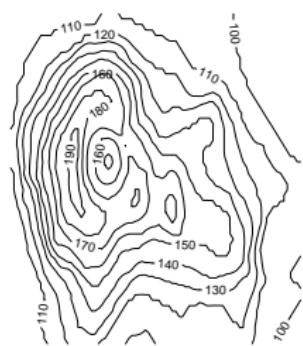


Image with
green replaced
by turquoise



aus Wong, B. *Nature Methods* 8(6), p. 441 (2011).

Wahrnehmungseinheitlichkeit: Veranschaulichung



Farträume

- ▶ repräsentieren Farben als (Vektoren von) Zahlen
- ▶ "technische" Farträume: RGB (Rot-Grün-Blau, für Bildschirme) oder CMYK (Cyan-Magenta-Yellow-Black, für Druck)
- ▶ Additive Farträume wie RGB / CMYK entsprechen nicht Funktionsweise der menschlichen Wahrnehmung, deswegen:
- ▶ Farträume:
 - ▶ HCL: hue-chroma-luminance
 - ▶ HSV: hue-saturation-value
 - ▶ Lab: Lightness-a (grün-rot Achse) - b (blau-gelb Achse)

HCL - Farbraum

- ▶ **Farbton:** dominante Wellenlänge (*hue*)
 - ▶ kreisförmige Dimension (grün - gelb - rot - lila - blau - grün)



- ▶ **Farbsättigung** (*chroma*)
- ▶ **Helligkeit** (*luminance*)

HCL - Farbraum

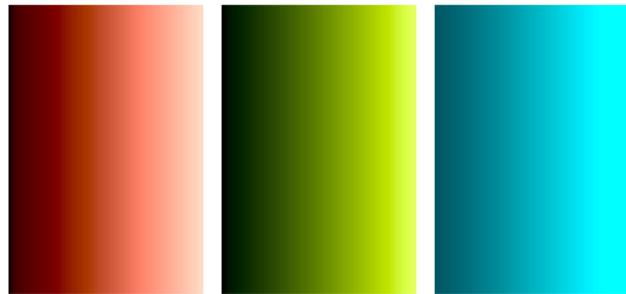
- ▶ **Farbton:** dominante Wellenlänge (*hue*)
 - ▶ kreisförmige Dimension (rot - gelb - grün - blau - lila - rot)
- ▶ **Farbsättigung** (*saturation, chroma*)



- ▶ **Helligkeit** (*brightness, luminance*)

HCL - Farbraum

- ▶ **Farbton:** dominante Wellenlänge (*hue*)
 - ▶ kreisförmige Dimension (rot - gelb - grün - blau - lila - rot)
- ▶ **Farbsättigung** (*saturation, chroma*)
- ▶ **Helligkeit** (*brightness, luminance*)



HCL - Farbraum

- ▶ **Farbton:** dominante Wellenlänge (*hue*)
 - ▶ kreisförmige Dimension (rot - gelb - grün - blau - lila - rot)
- ▶ **Farbsättigung** (*saturation, chroma*)
- ▶ **Helligkeit** (*brightness, luminance*)



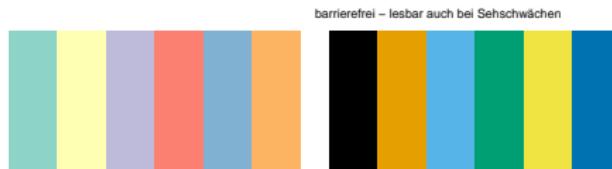
Farbskalentypen

- ▶ **Qualitativ:** (eher) nur für nominales Skalenniveau.
- ▶ **Sequentiell:** mindestens ordinale Skalenniveau
- ▶ **Divergierend:** mindestens ordinale Skalenniveau, mit “neutralem” mittleren Wert.

Farbskalen für nominales Skalenniveau

Qualitative Farbskalen für nominale Variablen:

- Fokus auf **Unterscheidbarkeit** bei gleichbleibender “Prägnanz”
- Variierender Farbton bei konstanter Sättigung & Helligkeit (korrigiert für menschliche Wahrnehmung)



Beispiel



Erster Vokal im Namen

Vowel	Color
a	Pink
e	Yellow-Gold
i	Green
o	Cyan
u	Purple

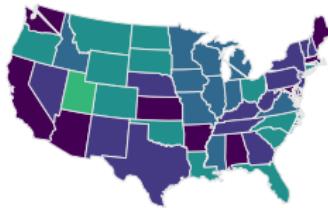
Sinnvoll, aber viel rot-grün Kontrast



Erster Vokal im Namen

Vowel	Color
a	Brown
e	Light Green
i	Dark Green
o	Cyan
u	Blue

Bessere Balance, weniger rot-grün



Erster Vokal im Namen

Vowel	Color
a	Dark Purple
e	Medium Purple
i	Dark Blue
o	Medium Blue
u	Light Blue

Ok, aber Farben implizieren Ordnung



Erster Vokal im Namen

Vowel	Color
a	Red
e	Green
i	Blue
o	Cyan
u	Purple

Schlechte Farbskala! Zu viel (rot-grün) Kontrast & Sättigung.

Sequentielle Farbskalen

- ▶ konstanter Farbton und Sättigung, variiere nur Helligkeit



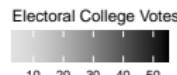
- ▶ mehrere Farbtöne, mit variierender Sättigung und abnehmender Helligkeit



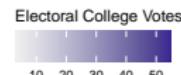
- ▶ Höchste Prägnanz für Werte die mit niedriger Helligkeit kodiert werden.

Beispiel

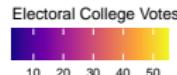
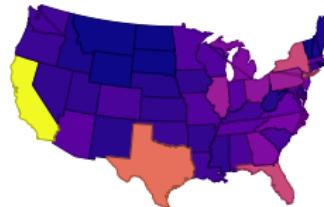
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



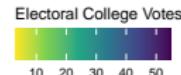
Sinnvoll, aber evtl. kleine Unterschiede verwischt



Sinnvoll, aber evtl. kleine Unterschiede verwischt



Sinnvoll, aber besser dunklere Farben für 'wichtiger' Werte



Sinnvoll, überbetont evtl. kleine Unterschiede durch Farbübergänge

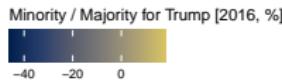
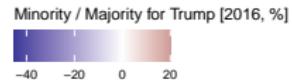
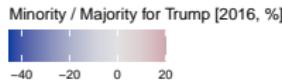
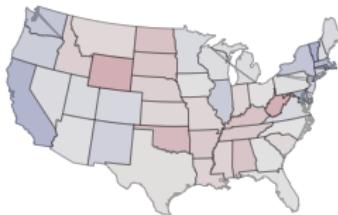
Divergierende Farbskalen

- ▶ 2 sequentielle Skalen mit je konstantem Farbton werden kombiniert

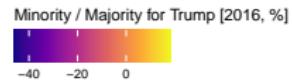


- ▶ "neutrale" Mitte: höchste Prägnanz für obere und untere Enden der Skala

Beispiel: Divergierende Farbskalen



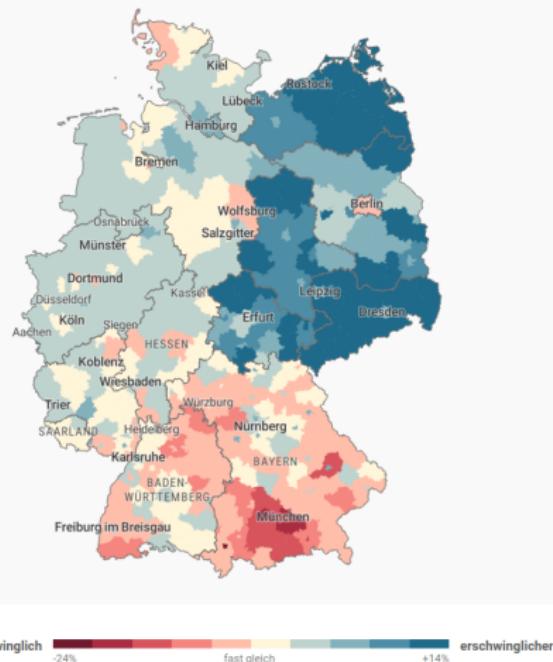
Mitte nicht viel weniger prägnant als Enden.



Nicht divergierend: Mitte nicht viel weniger prägnant als Enden, zu bunt

Beispiel: Divergierende Farbskalen

Differenz der Änderungsraten von Median-Bruttolöhnen und Neuvertragsmieten 2014 - 2018:



Quelle: IW, Berliner Morgenpost (17.01.2020)

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Visualisierung der Häufigkeiten diskreter Merkmale

(Bedingte) Häufigkeiten für diskrete/gruppierte Merkmale darstellbar u.a.
als

- ▶ Stab-, Balken- und Säulendiagramm
- ▶ Histogramm
- ▶ (Kreis-/Tortendiagramm)

Stab-, Säulen-, Balkendiagramm

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

Anwendungen:

- ▶ Ordinale Merkmale
- ▶ Metrische Merkmale mit *wenigen Ausprägungen*
- ▶ Nominale Merkmale (Problem: Anordnung der Ausprägungen beliebig!)

Stab-, Säulen-, Balkendiagramm

Darstellung der absoluten oder relativen Häufigkeiten als Höhen (Längen).

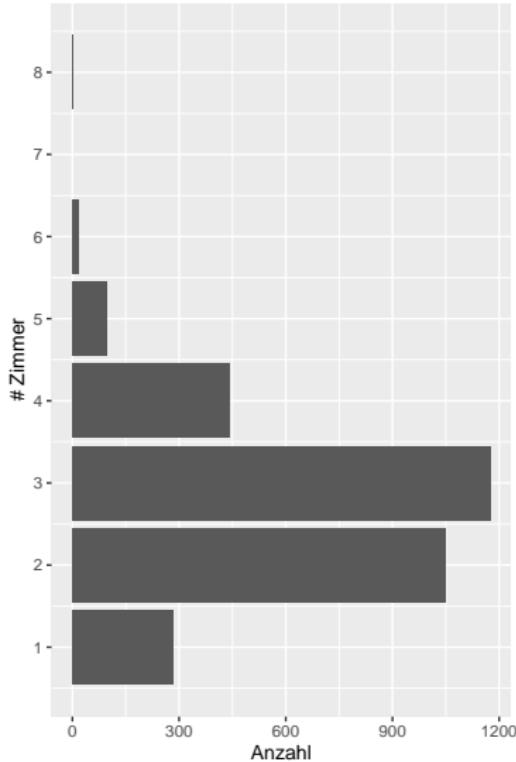
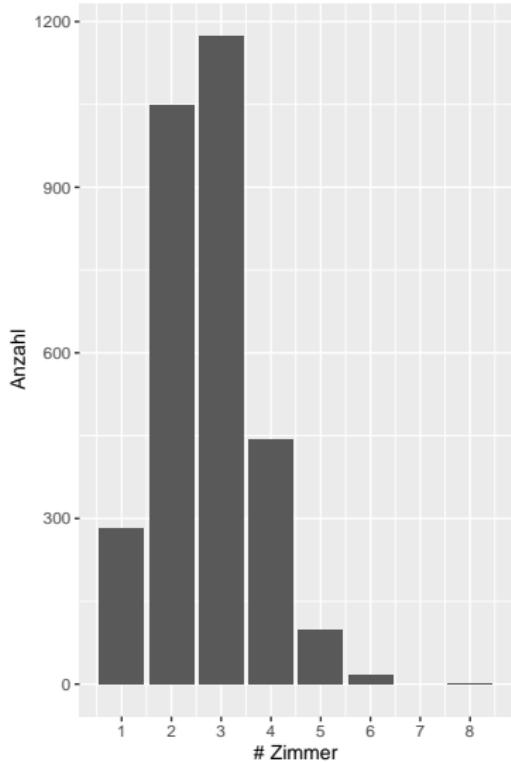
Verwendete Geometrien:

Vertikale oder horizontale Linien oder Rechtecke

Ästhetische Zuordnungen:

- ▶ x-Position: Ausprägungen des Merkmals
- ▶ y-Position des oberen Endes: absolute oder relative Häufigkeiten
- ▶ (x und y umgekehrt für Balkendiagramm)

Beispiel Mietspiegel: Säulendiagramm / Balkendiagramm



Stapeldiagramm

Darstellen absoluter oder relativer Häufigkeiten als Länge. Die Abschnitte werden übereinander gestapelt und unterschiedlich eingefärbt.

Anwendungen:

- ▶ Nominale Merkmale (aber: Reihenfolge?!)
- ▶ Ordinale / gruppierte Merkmale
- ▶ Metrische Daten mit wenigen Ausprägungen

Besonders geeignet für den Vergleich verschiedener Gruppen durch nebeneinander liegende Stapel.

⇒ also: *bedingte* Häufigkeiten gegeben Gruppe

Zu beachten ist dann die Unterscheidung:

relative Häufigkeit (Anteil) ↔ absolute Häufigkeit (Anzahl)

Stapeldiagramm

Darstellung der absoluten oder relativen Häufigkeiten als Segmente unterschiedlicher Länge (vertikal) oder Breite (horizontal).

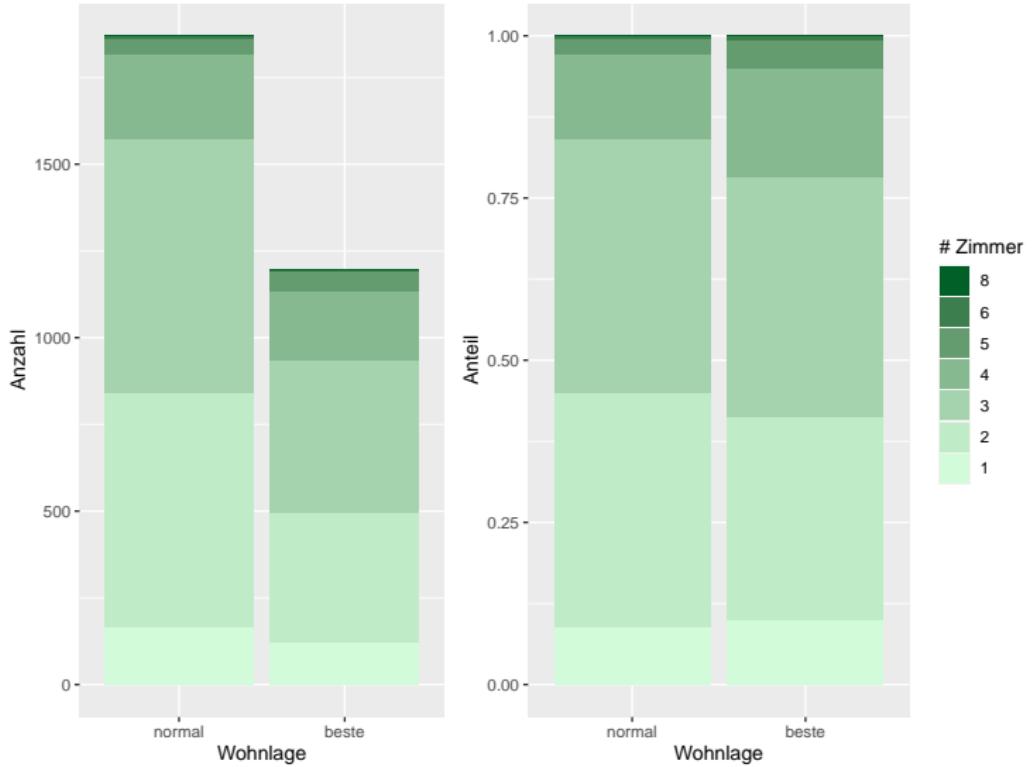
verwendete Geometrien:

Vertikale oder horizontale Rechtecksegmente

Ästhetische Zuordnungen:

- ▶ Ausdehnung: absolute oder relative Häufigkeiten der Merkmalsausprägungen
- ▶ (Füll-)farbe: Ausprägungen des Merkmals
- ▶ optional: x- oder y-Position des Segments: Ausprägungen des bedingenden Merkmals

Beispiel Mietspiegel: Stapeldiagramme



Kreisdiagramm, Tortendiagramm

Darstellung von relativen/absoluten Häufigkeiten als Anteile der Fläche eines Kreises.

Suboptimale Visualierung –

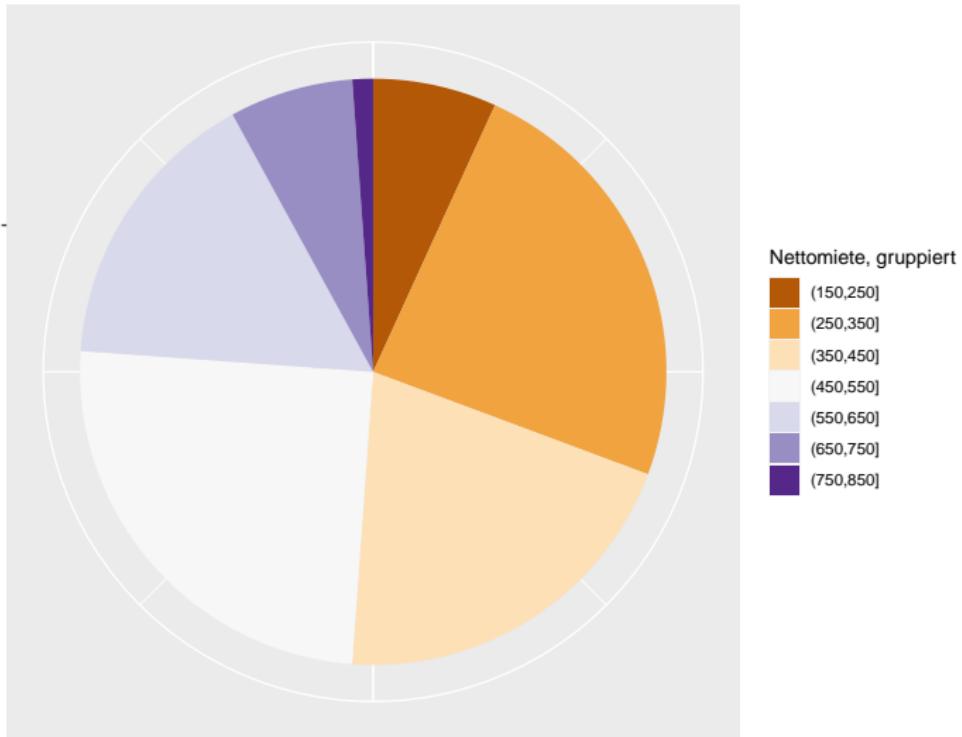
Darstellung über Längen » Darstellung über Winkel!

Grundsätzlich anwendbar für

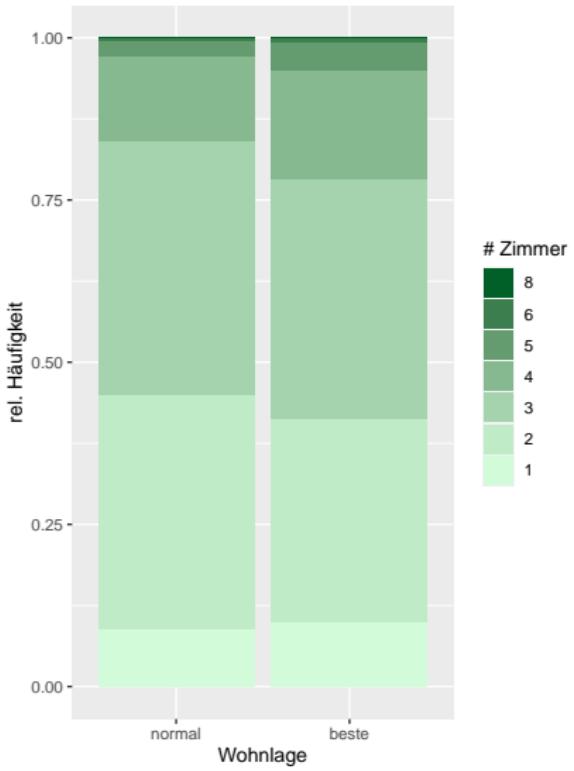
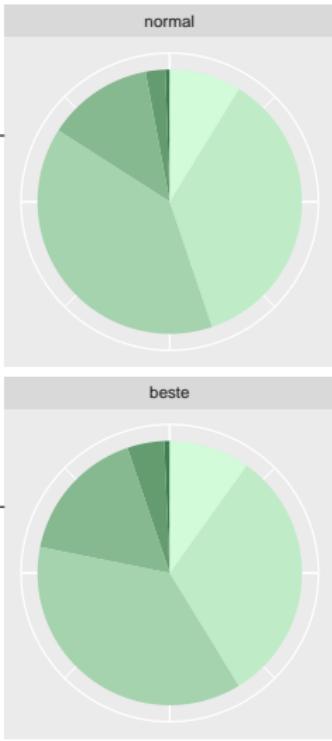
- ▶ Nominale Merkmale
- ▶ Ordinale Merkmale (Problem: Ordnung nicht korrekt wiedergegeben)
- ▶ Gruppierte Daten

pie chart

Tortendiagramm: Klein & Kalt



Beispiel Mietspiegel: Vergleich mit Kreisdiagramm



Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Dotplots

“Eindimensionales Streudiagramm”

Darstellung der einzelnen Beobachtungen als Punkte (Breiten) entlang der Achse.

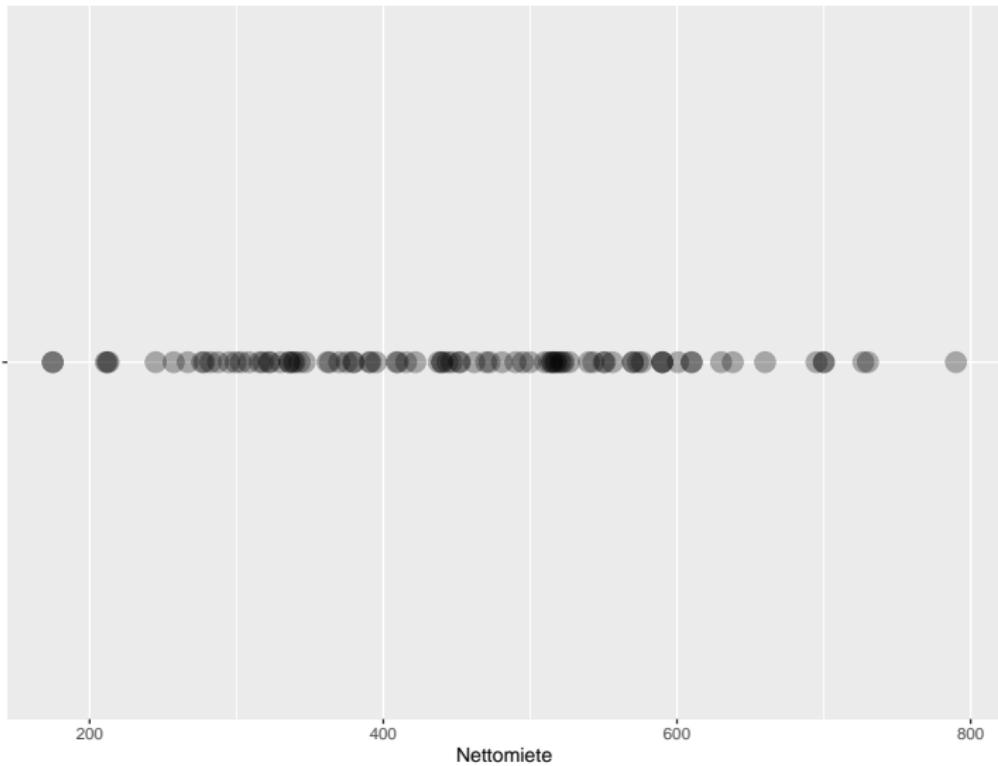
Verwendete Geometrien:

Punkte (oder andere Symbole)

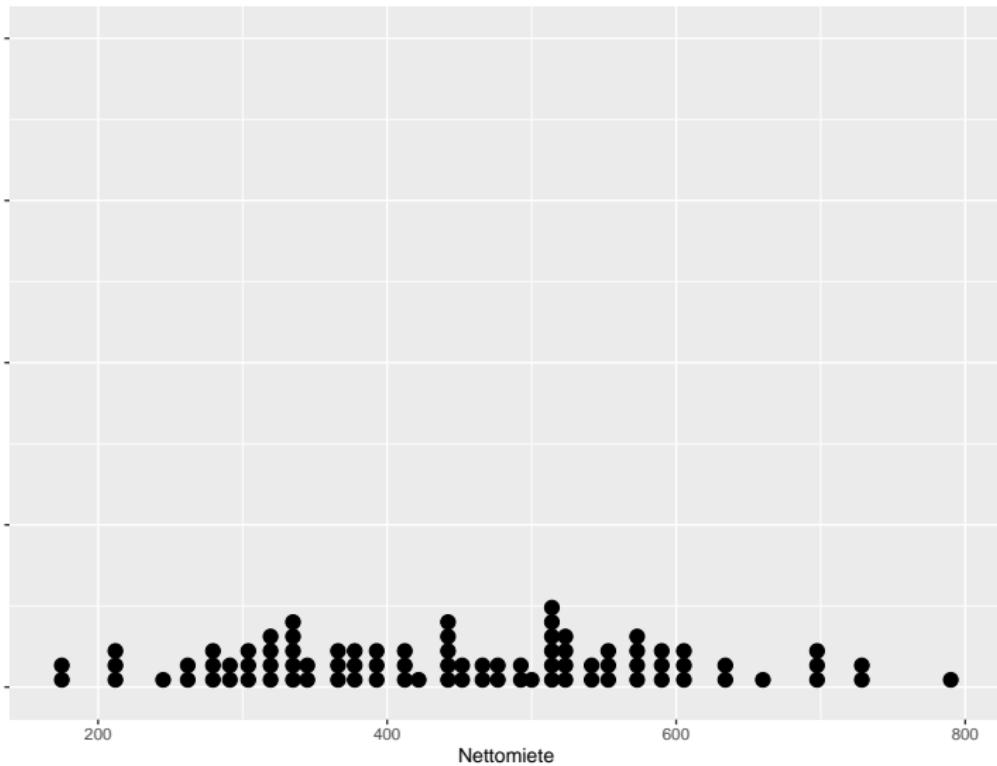
Ästhetische Zuordnungen:

- ▶ x-Position: beobachtete Merkmalsausprägungen der UE
- ▶ optional: y-Position: Ausprägungen des bedingenden Merkmals
- ▶ (oder auch umgekehrt)

Beispiel: Dotplot

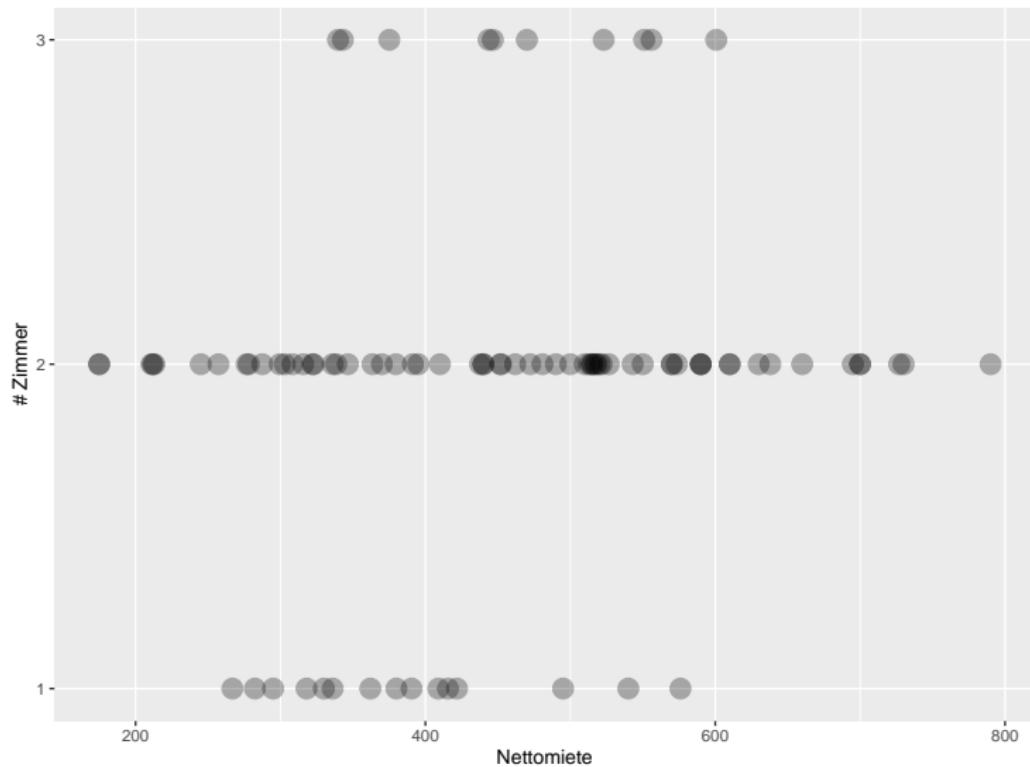


Variante: Gruppiertter Dotplot

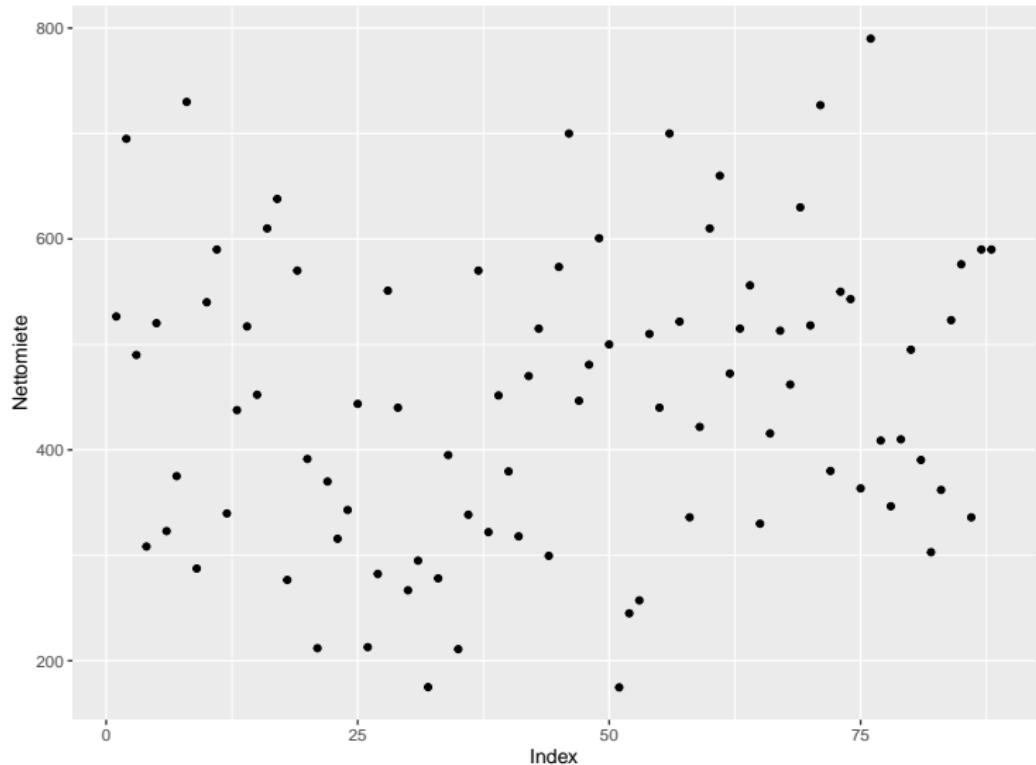


(bin width = 10)

Beispiel: Dotplot für bedingte Verteilungen



Negativbeispiel: Streudiagramm



→ Ungeeignet, da horizontale Achse ohne Informationsgehalt.

Histogramm

Grafiktyp für Häufigkeitsverteilungen auf mindestens Intervallskala.

Darstellung der relativen Häufigkeiten durch *Flächen* (Prinzip der **Flächentreue**)

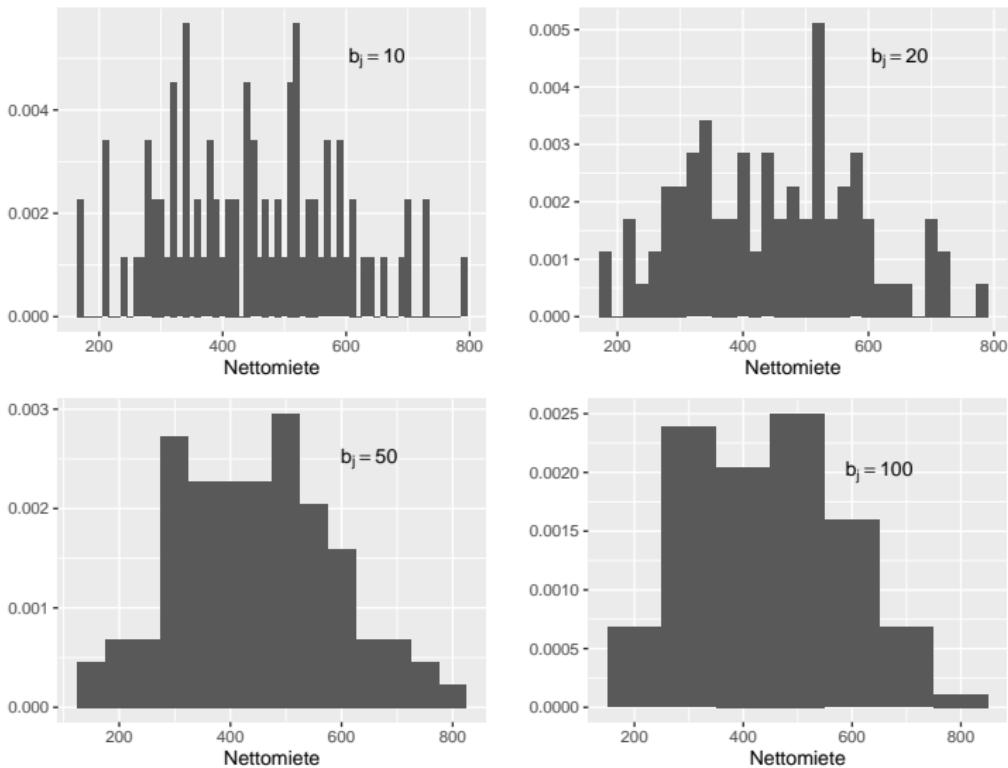
Vorgehen:

1. Aufteilung in Klassen (falls die Daten noch nicht gruppiert sind)
2. Bestimmung der Klassenhäufigkeiten n_j und der relativen Häufigkeiten

$$f_j = \frac{n_j}{n}$$

3. Bestimmung der Histogramm-Höhen y_j , so dass gilt:
 $b_j \cdot y_j = f_j$, wobei b_j die Breite der Klasse j ist.

Beispiel: Nettomiete Klein & Kalt



Histogramm

- ▶ Flächentreue Darstellung ähnlich zu Idee von *Dichtefunktionen*: Histogramm ist stückweise konstante Approximation der Dichte.
- ▶ nur bei metrischen Daten sinnvoll
- ▶ Beachte: Interpretation der Balkenhöhe bei unterschiedlichen Klassenbreiten nicht angemessen – Flächentreue!
 ⇒ unterschiedlichen Klassenbreiten vermeiden zu Gunsten einfacher Interpretierbarkeit
- ▶ Nachteil: Visueller Eindruck von Klassenbreite abhängig
 ⇒ verschiedene Varianten ansehen, falls möglich: Klassenbreiten inhaltlich begründen
- ▶ Vorsicht bei Rändern (s.a. Kapitel “Kerndichteschätzer”)

Histogramm

Flächentreue Darstellung der Verteilung eines metrischen Merkmals, unterteilt in Klassen.

Verwendete Geometrien:

Achsenparallele Rechtecke

Ästhetische Zuordnungen:

- ▶ x-Position der Ecken: Histogrammklassengrenzen
- ▶ Fläche (*implizit also: Höhe*): relative Häufigkeit der Klasse

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

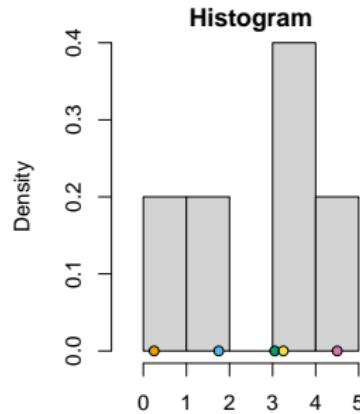
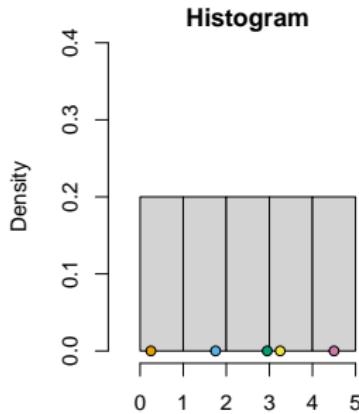
Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Dichtefunktion I

Histogramm:

- ▶ Relative Häufigkeit in einer Histogramm-Klasse = Fläche der Histogramm-Säule = Fläche “unter der Kurve”
- ▶ Aber: Histogramm ist stückweise konstante Funktion
- ▶ Sehr problematisch: Abhängigkeit von Wahl der Klassengrenzen



- ▶ \Rightarrow Ersetze Histogramm durch glatte Funktion f

Dichtefunktion II

Wiederholung: Für eine **Dichte**(-funktion) (*density*) gilt:

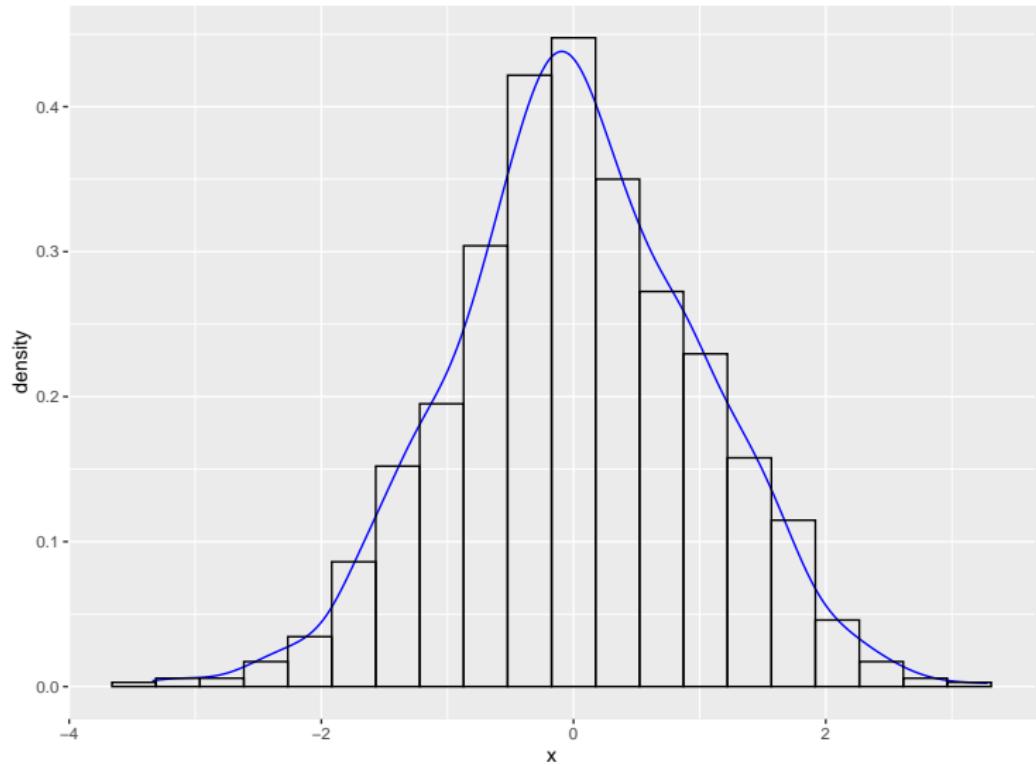
- ▶ $f(x) \geq 0 \quad \forall x$ und
- ▶ $\int_{-\infty}^{\infty} f(x)dx = 1.$

Die Fläche unter der Dichte über beliebige Intervalle soll in etwa den relativen Häufigkeiten in diesen Intervallen entsprechen, d.h.

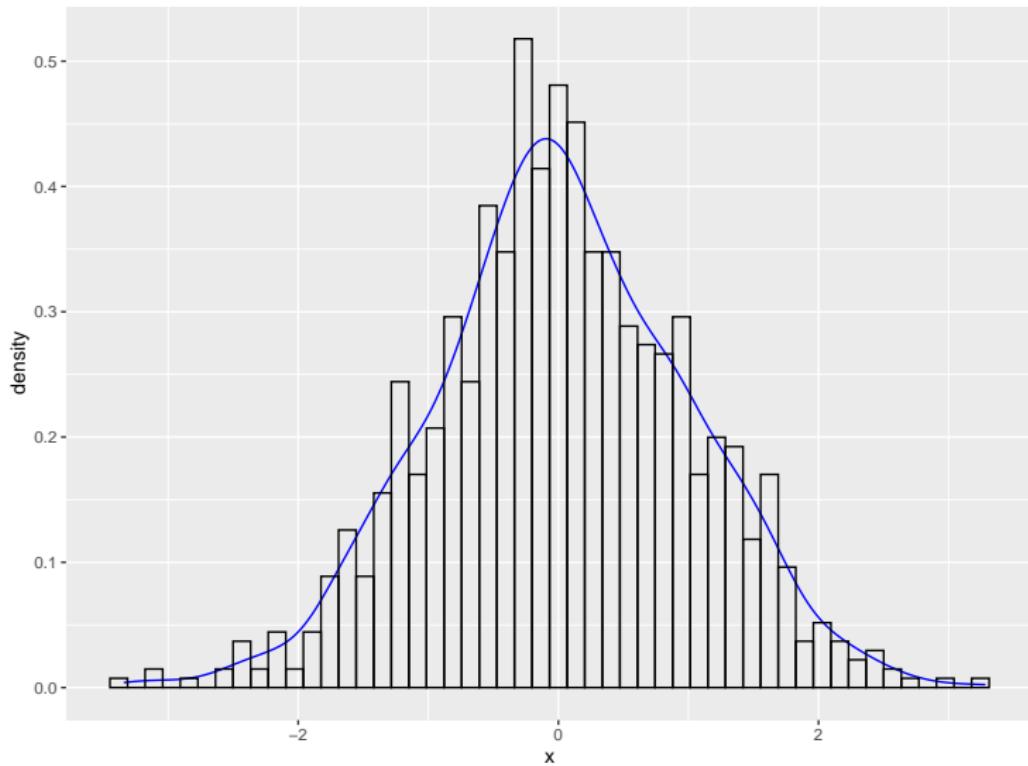
$$\int_a^b f(x)dx \approx \frac{1}{n} |\{x_i : a < x_i \leq b\}|$$

$|\{\dots\}|$: Mächtigkeit der Menge, also Anzahl der Elemente der Menge

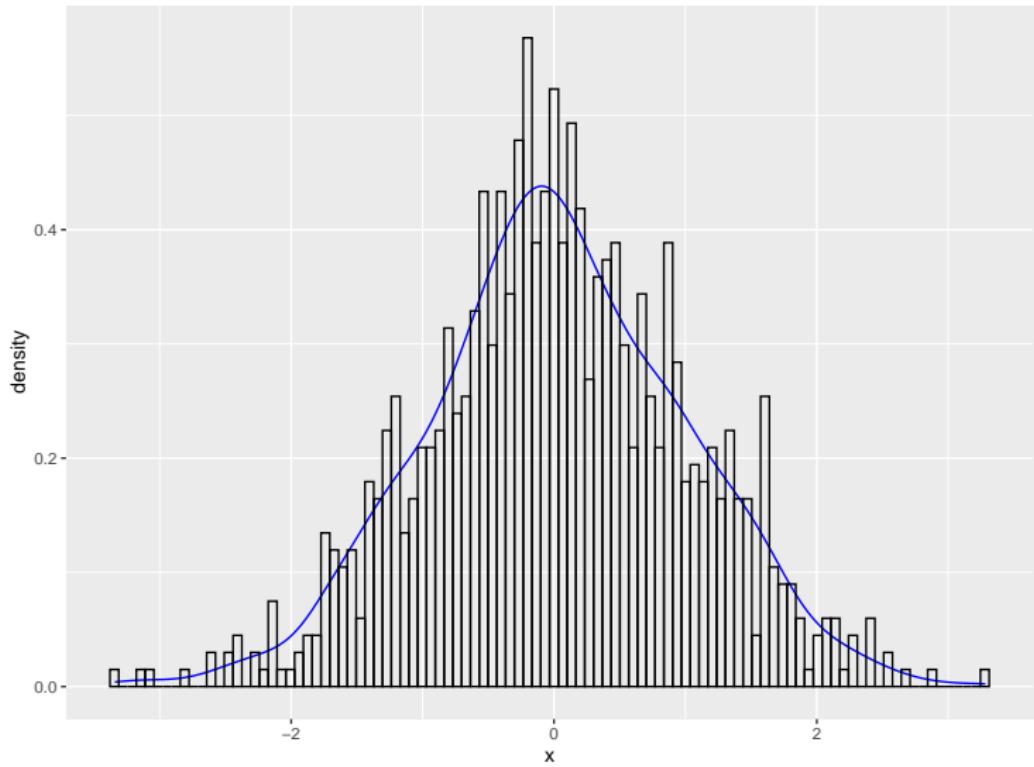
Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Berechnung von Dichte-Kurven

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

⇒ Gleitendes Histogramm

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

mit “Rechteck”-**Kernfunktion**

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

Kerndichteschätzer

$k(u)$ sei **Kernfunktion**, d.h. $k(u) \geq 0 \forall u$ und $\int_{-\infty}^{\infty} k(u)du = 1$

Dann ist der **Kerndichteschätzer** (auch: KDE - *kernel density estimator*)

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

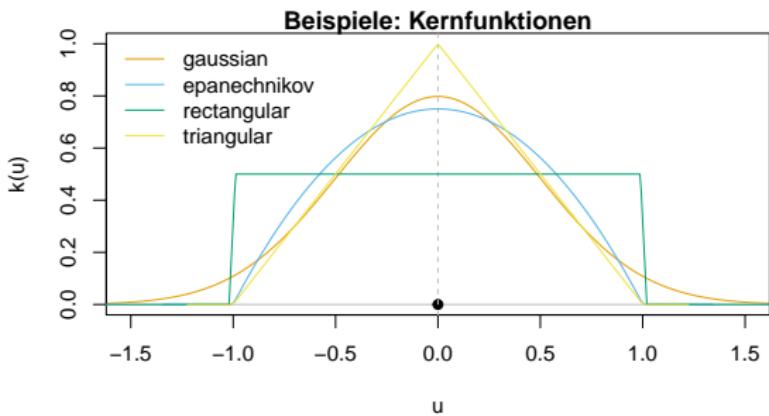
Beispiele für Kernfunktionen:

Gauß-Kern $k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$

Epanechnikov-Kern $k(u) = \max\left(0, \frac{3}{4}(1 - u^2)\right)$

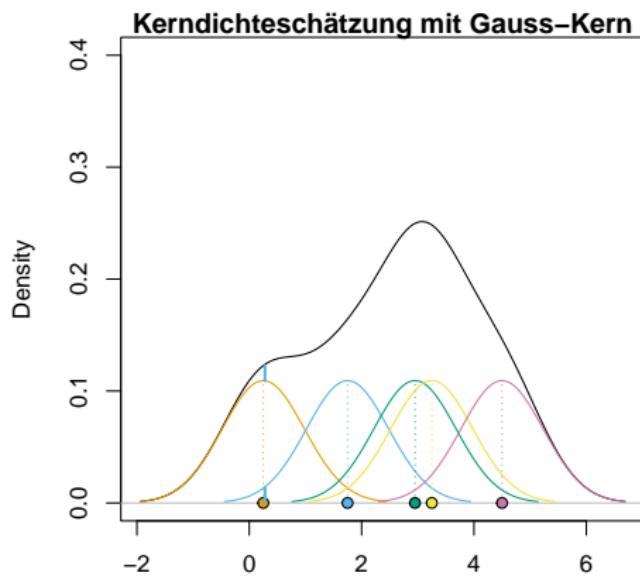
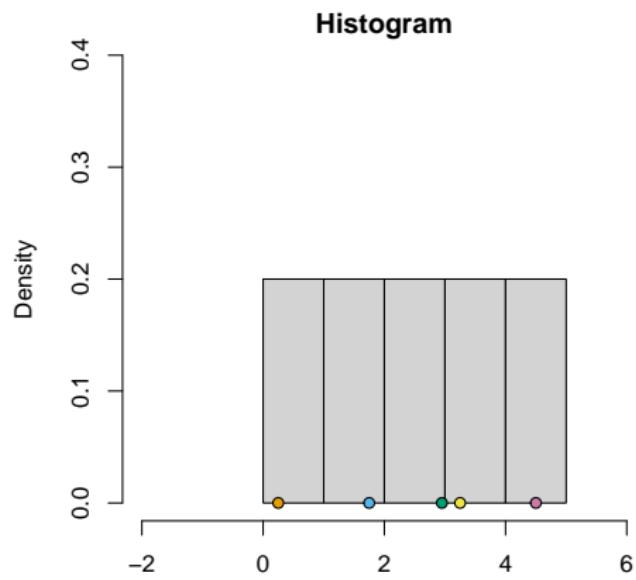
Dreieck-Kern $k(u) = \max(0, 1 - |u|)$

Kerndichteschätzer



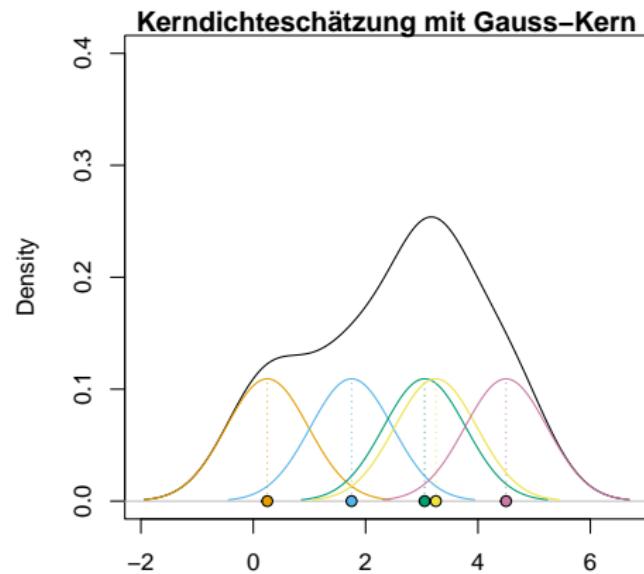
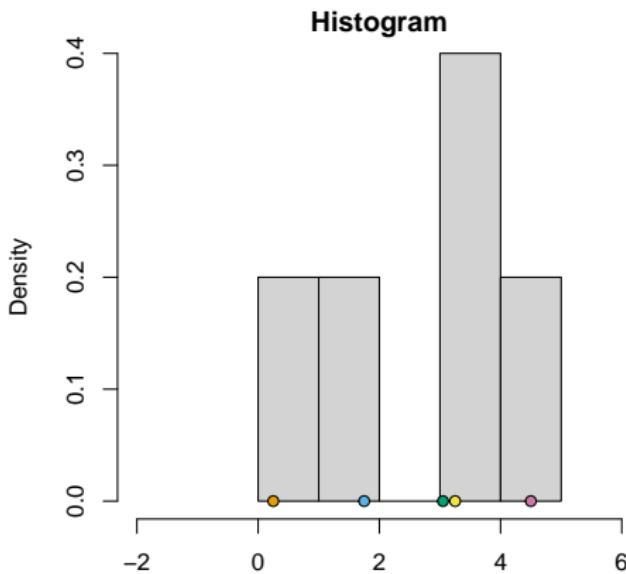
Kerndichteschätzer

- Histogramme berücksichtigen nicht ob Beobachtungen zentral oder am Rand der Klasse liegen, zählen nur die *Anzahl der Beobachtungen innerhalb der Klassengrenzen*
- Kerndichteschätzungen berücksichtigen die *Entfernung der benachbarten Punkte*, mit abnehmender Gewichtung über die Distanz:



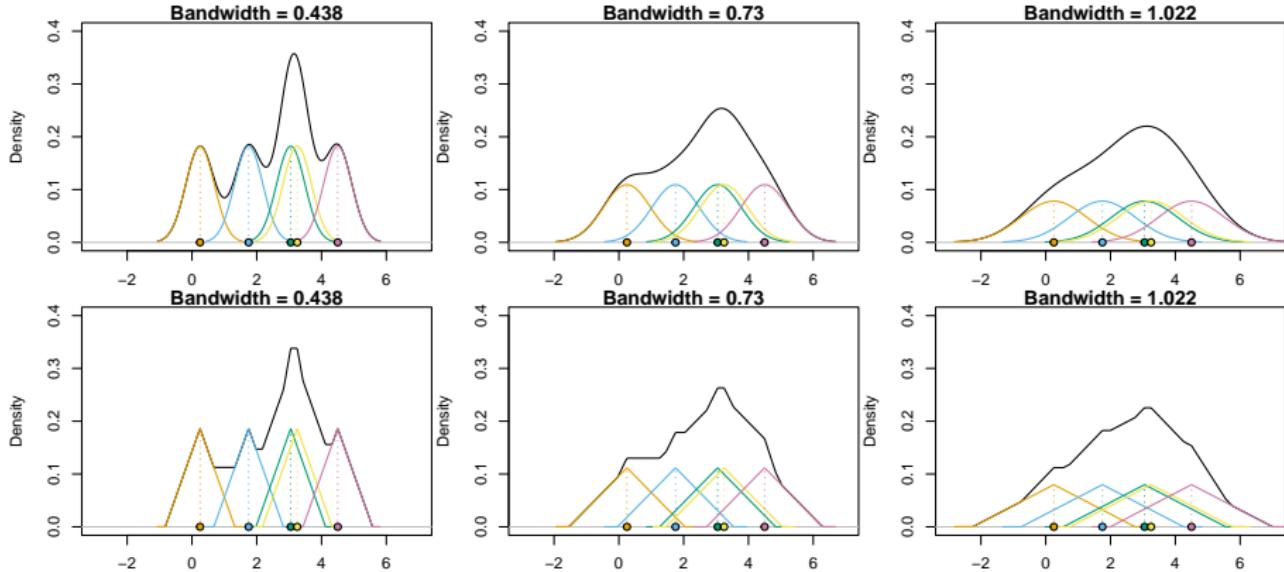
Kerndichteschätzer

- Histogramme berücksichtigen nicht ob Beobachtungen zentral oder am Rand der Klasse liegen, zählen nur die *Anzahl der Beobachtungen innerhalb der Klassengrenzen*
- Kerndichteschätzungen berücksichtigen die *Entfernung der benachbarten Punkte*, mit abnehmender Gewichtung über die Distanz:



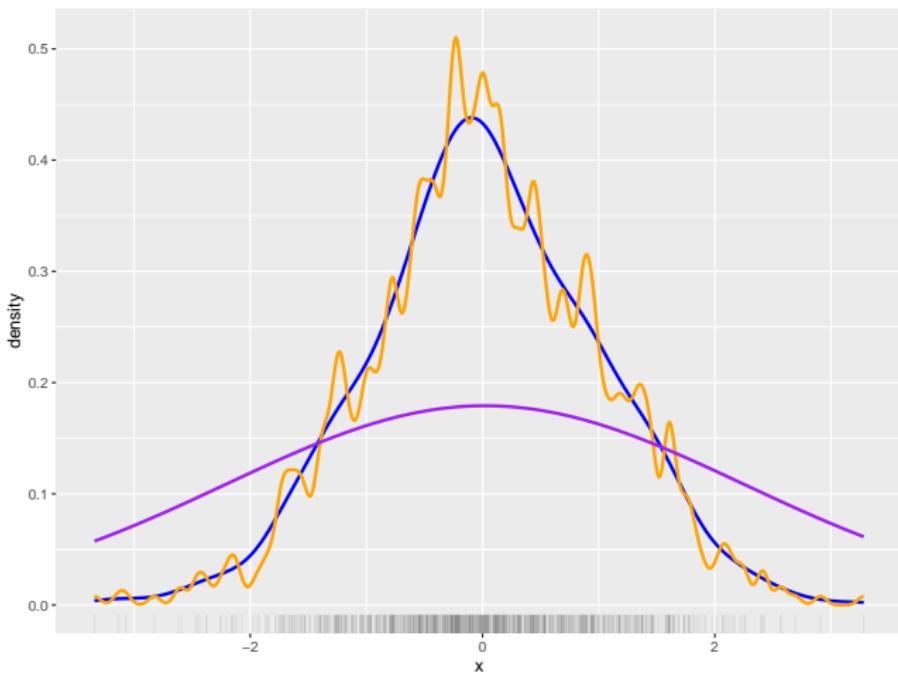
⇒ KDE

Kerndichteschätzer: Bandbreite



Obere Reihe: Gauss-Kern, untere Reihe: Dreiecks-Kern

Kern-Dichteschätzer



Kerndichteschätzer mit “optimaler” Bandbreite h (blau), zu kleiner (orange) und zu großer (lila) Bandbreite h .

Bemerkungen zur Dichteschätzung

- ▶ Abhängigkeit von der Bandbreite $h \rightarrow$ Verfahren zur Bestimmung von h aus den Daten
- ▶ Abhängigkeit von der Wahl des Kerns eher unbedeutend
- ▶ Kerndichteschätzungen sind insbesondere bei größeren Datenmengen und (quasi-)stetigen Merkmalen Histogrammen vorzuziehen.
- ▶ Kerndichteschätzungen sind immer Histogrammen mit unterschiedlichen Klassenbreiten vorzuziehen.

[Animation 1](#), [Animation 2](#)

Statistische Grafiken

Grammar of Graphics

Grafiken: Wahrnehmung

Grafiken: Infoviz vs Statistische Grafiken

Farbskalen

Diskrete Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Metrische Merkmale: Visualisierung von Häufigkeiten und Verteilungen

Kerndichteschätzung

Metrische Merkmale: Visualisierung gemeinsamer Verteilungen

Bivariate metrische Daten

Daten liegen zu zwei metrischen Merkmalen vor:

Datenpaare $(x_i, y_i), i = 1, \dots, n$

Fragen:

Gibt es einen Zusammenhang zwischen diesen Merkmalen?

Wie lässt sich dieser Zusammenhang beschreiben?

Was ist die *gemeinsame* Verteilung dieser Merkmale?

Einfachste grafische Darstellung: **Streudiagramm**.

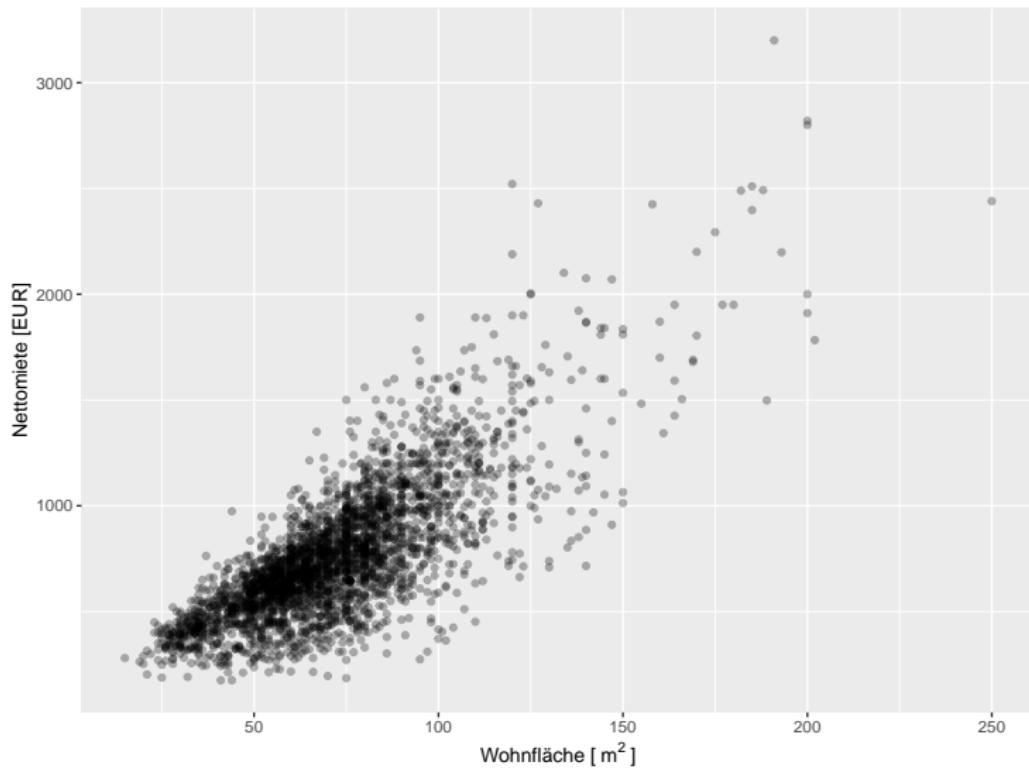
Die Datenpaare entsprechen Punkten in der Ebene (“Punktwolke”)

Beispiele:

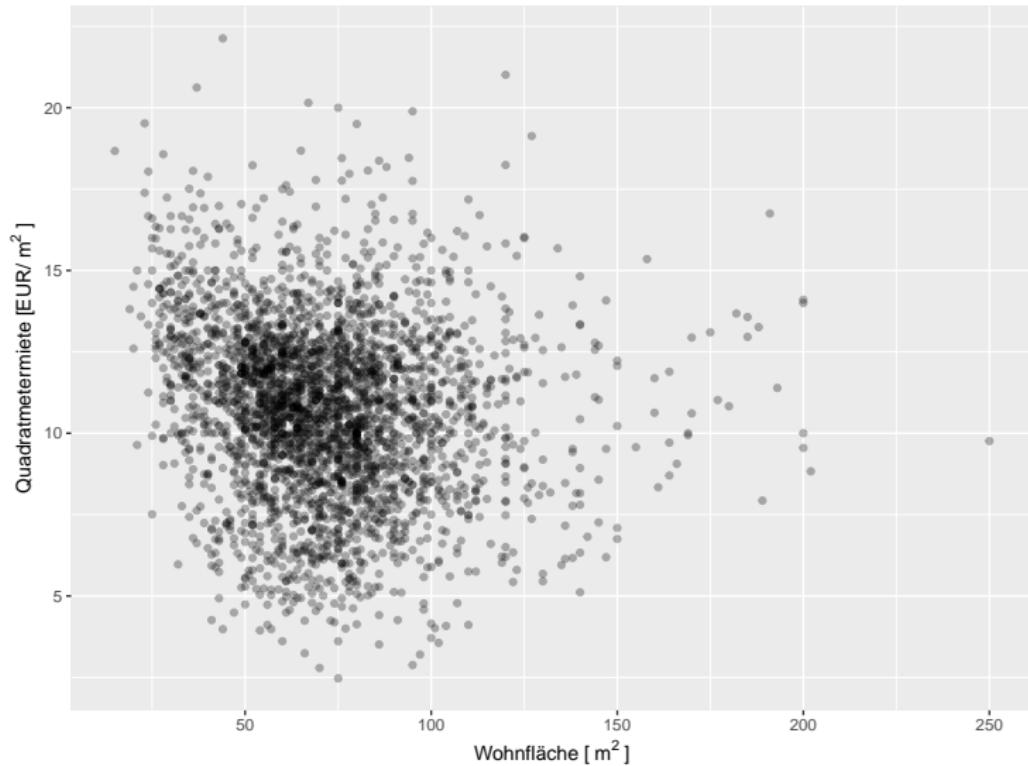
X: Wohnfläche [m^2]

Y: Nettomiete [€] oder Quadratmetermiete [€/ m^2]

Streudiagramm

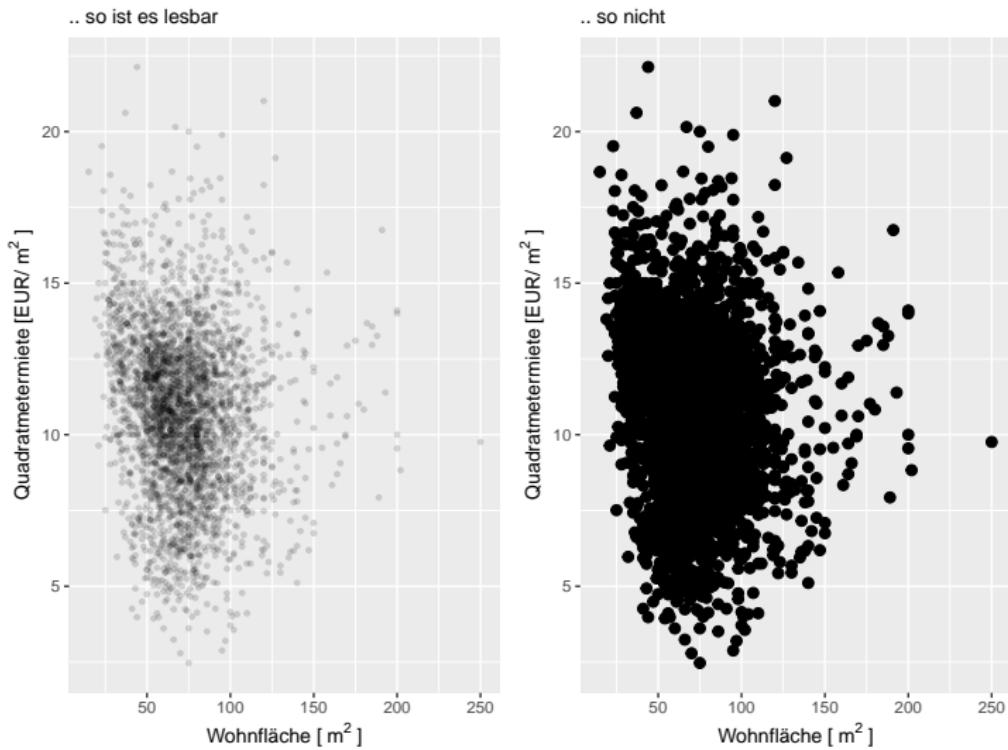


Streudiagramm



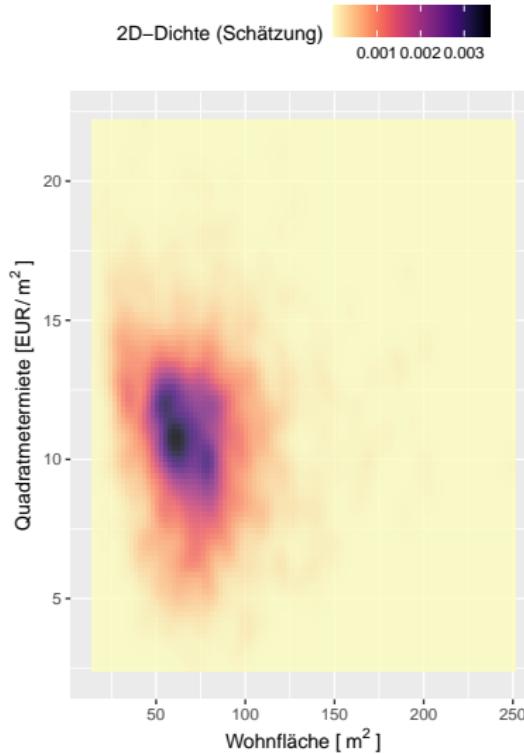
Darstellung für größere Datenmengen:

Besser mit halbdurchsichtigen & kleineren Symbolen - Overplotting vermeiden:

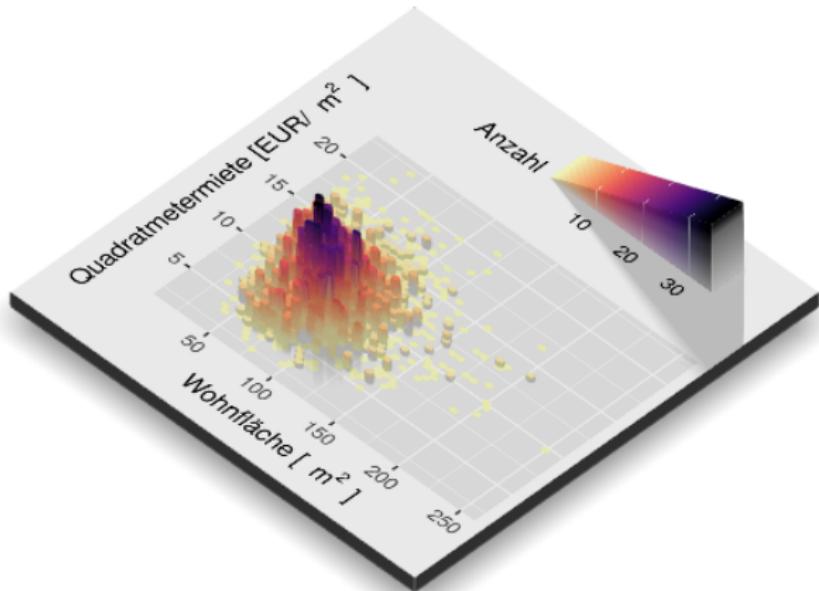


Darstellung für größere Datenmengen

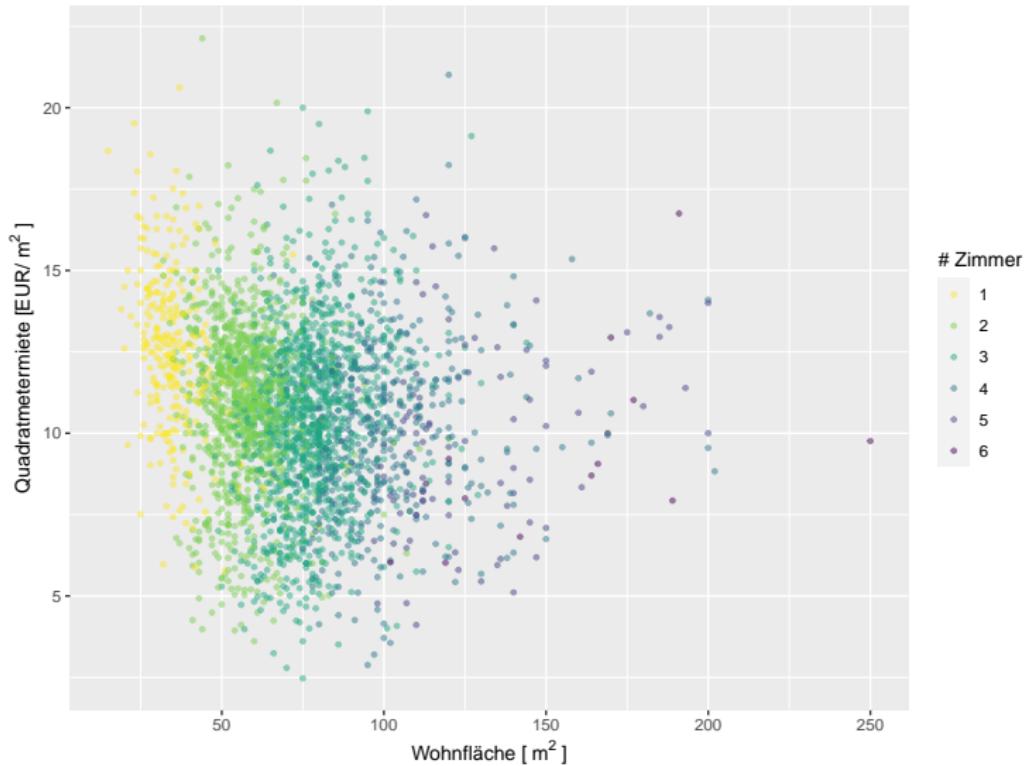
Alternativen: Anzahlen/Dichte direkt über Farbe codieren (hexbin-Plots, 2D-Kerndichte):



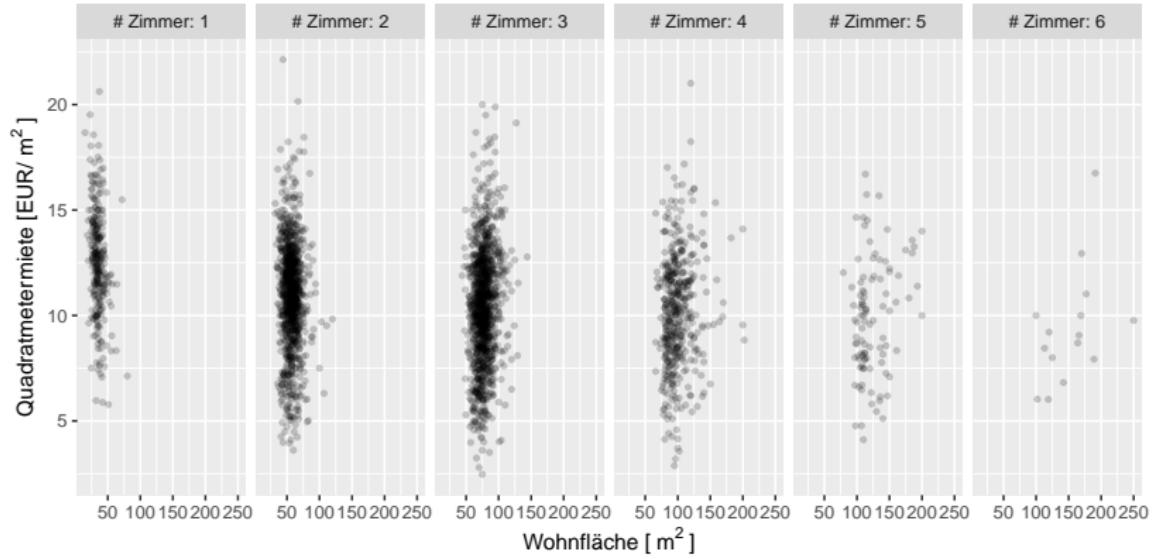
Darstellung für größere Datenmengen



Streudiagramme mit diskreter Drittvariable



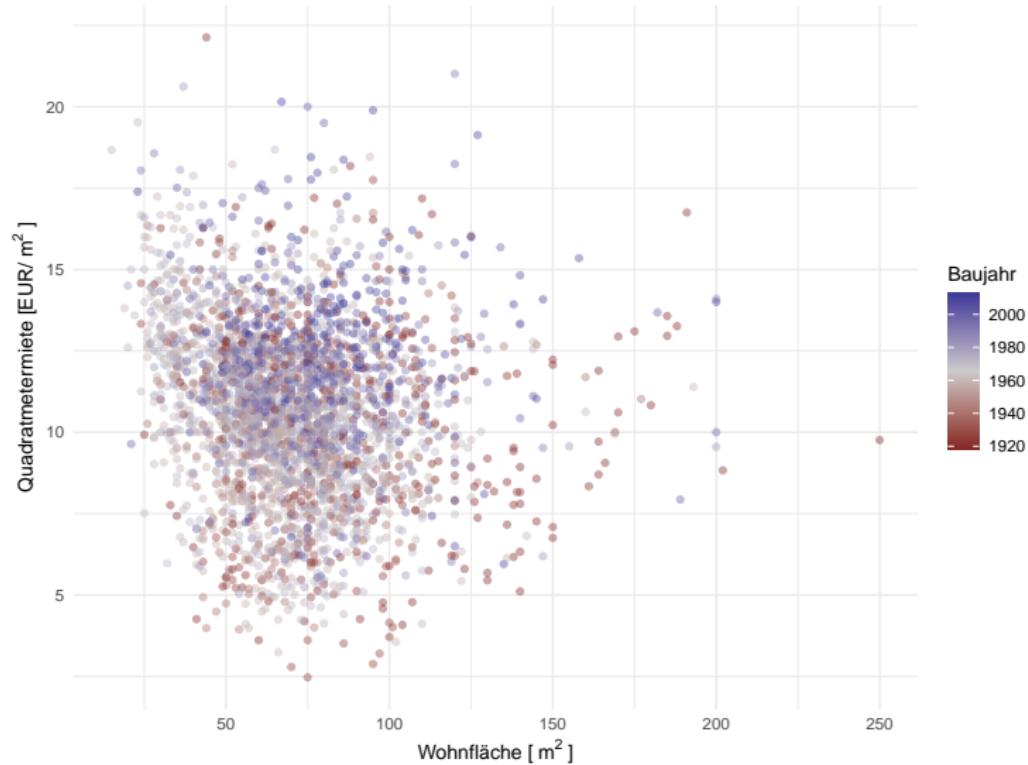
Streudiagramme mit diskreter Drittvariable



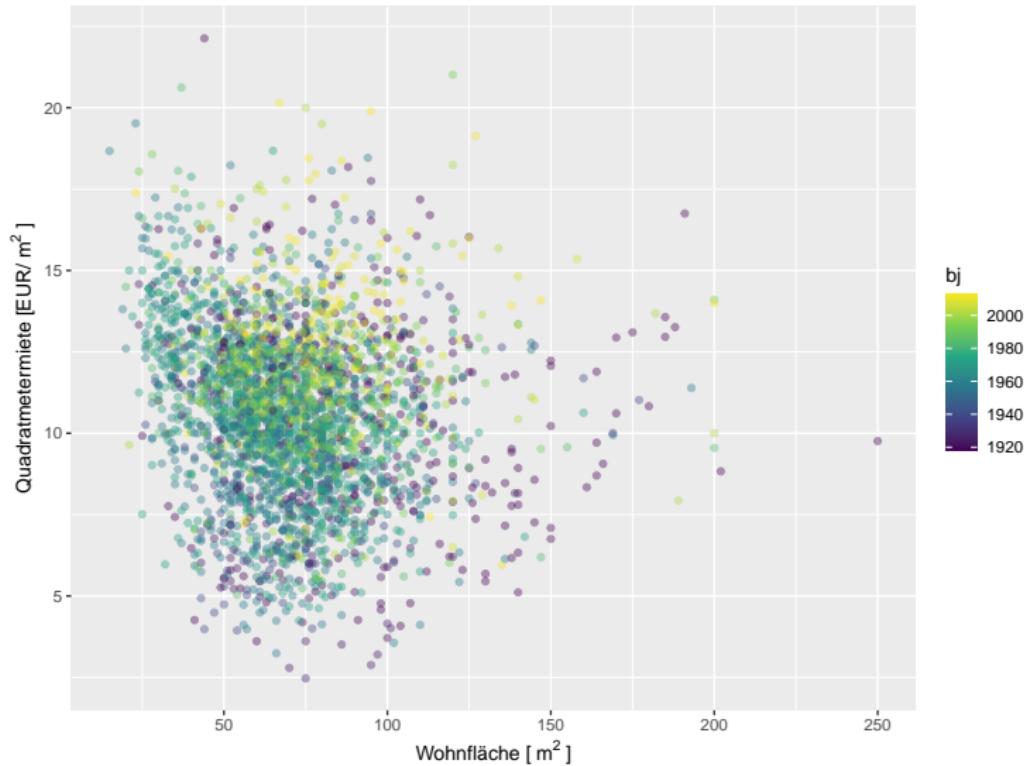
Streudiagramme mit diskreter Drittvariable



Streudiagramme mit stetiger Drittvariable



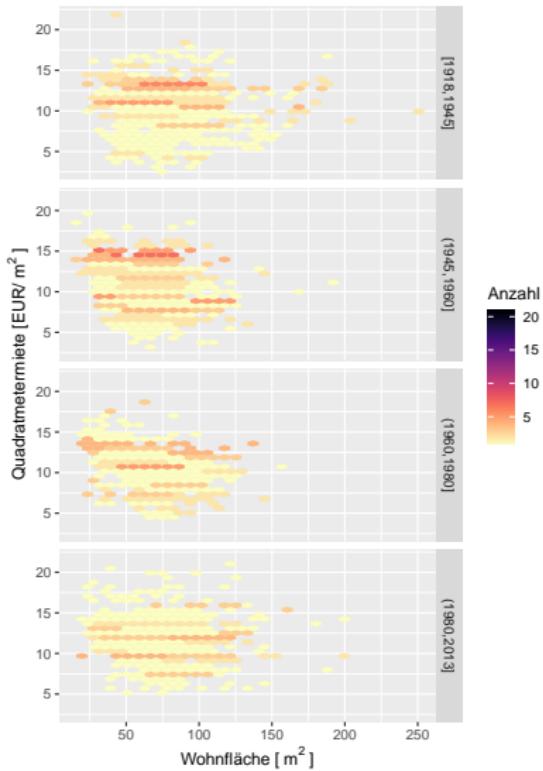
Streudiagramme mit stetiger Drittvariable



Streudiagramme mit gruppiertter Drittvariable



Streudiagramme mit gruppieter Drittvariable



Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Statistische Kennwerte

Lagemaßzahlen

- ▶ Wo liegt die “Mitte” der beobachteten Werte?
- ▶ Welche Merkmalsausprägung ist “typisch” für dieses Merkmal?
- ▶ Wo liegen die meisten beobachteten Daten?

\vskip 3em

Streuungsmaßzahlen

- ▶ Wie groß ist die Schwankung der beobachteten Werte?
- ▶ Über welchen Bereich erstrecken sich die Merkmalsausprägungen?
- ▶ Wie nah zusammen/weit entfernt voneinander liegen die beobachteten Werte?

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Lagemaß: Modus

Definition: **Häufigster Wert**

Eigenschaften:

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

Lagemaß: Median

Definition: Der **Median** (\tilde{x}_{med}) ist der Wert für den gilt

- mindestens 50% der Daten sind kleiner oder gleich \tilde{x}_{med} ,
- mindestens 50% der Daten sind größer oder gleich \tilde{x}_{med} .

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

- $x_{(1)}, \dots, x_{(n)}$ sind **geordnete Werte**
- Alternative Definition: $\tilde{x}_{\text{med}} \in [x_{(k)}, x_{(k+1)}]$ falls $k = \frac{n}{2}$ ganze Zahl.

Eigenschaften des Medians

- ▶ anschaulich
- ▶ geeignet für mindestens ordinale Daten
- ▶ robust gegenüber extremen Werten
- ▶ direkt übertragbar bei monotonen Transformationen

Formal:

- ▶ Wert, der die Summe der *absoluten* Differenzen zu den beobachteten Werten minimiert:

$$\tilde{x}_{\text{med}} = \arg \min_x \sum_{i=1}^n |x_i - x|$$

Lagemaß: Quantil

Definition: Das p -**Quantil** ist der Wert \tilde{x}_p für den gilt

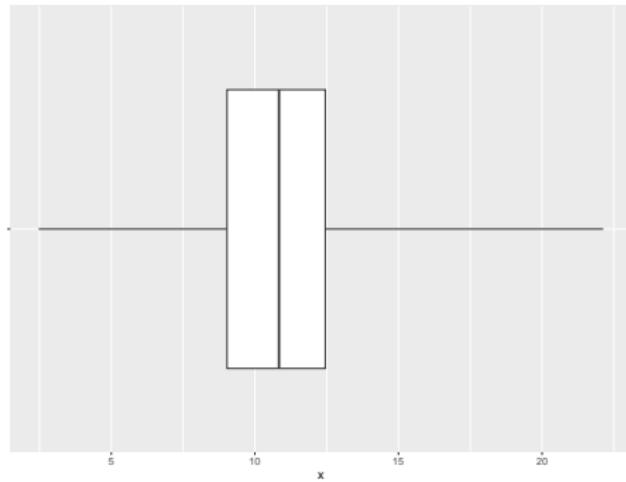
- mindestens Anteil p der Daten sind kleiner oder gleich \tilde{x}_p ,
- mindestens Anteil $1 - p$ der Daten sind größer oder gleich \tilde{x}_p .

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

- Viele alternative Definitionen von Quantilen (in R 9 Typen!), die sich aber meist nur für extreme Quantile relevant unterscheiden.
- p -Quantil $\equiv (100 \cdot p)$ -Perzentil
- Der Median ist das 0.5-Quantil bzw. 50-Perzentil

Anwendung in Visualisierung: Boxplot

- Überblicks-Darstellung der Verteilung eines Merkmals
- Visualisieren der 5-Punkte-Zusammenfassung:
Minimum, 25-, 50-, 75-Perzentile, Maximum



Boxplot

Einfacher **Boxplot**:

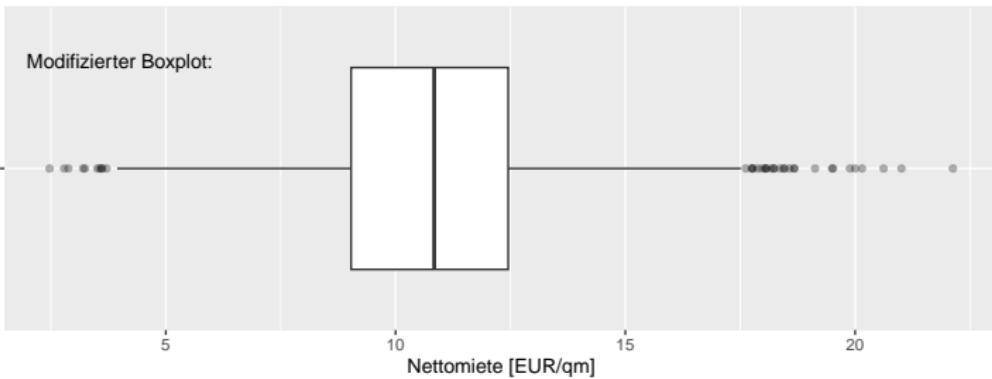
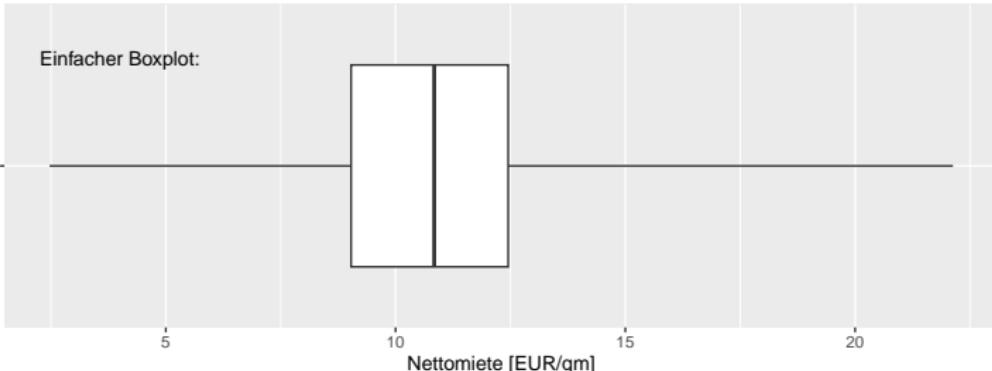
- ▶ $\tilde{x}_{0.25}$ = Anfang der Schachtel (Box) (= unteres **Quartil**)
- ▶ $\tilde{x}_{0.75}$ = Ende der Schachtel (= oberes **Quartil**)
- ▶ d_Q = Länge der Schachtel (= **Inter-Quartile-Range** (IQR) $\tilde{x}_{0.75} - \tilde{x}_{0.25}$)
- ▶ Der **Median** wird durch den Strich in der Box markiert
- ▶ Zwei Linien (*whiskers*) außerhalb der Box gehen bis zu x_{min} und x_{max} .

Boxplot

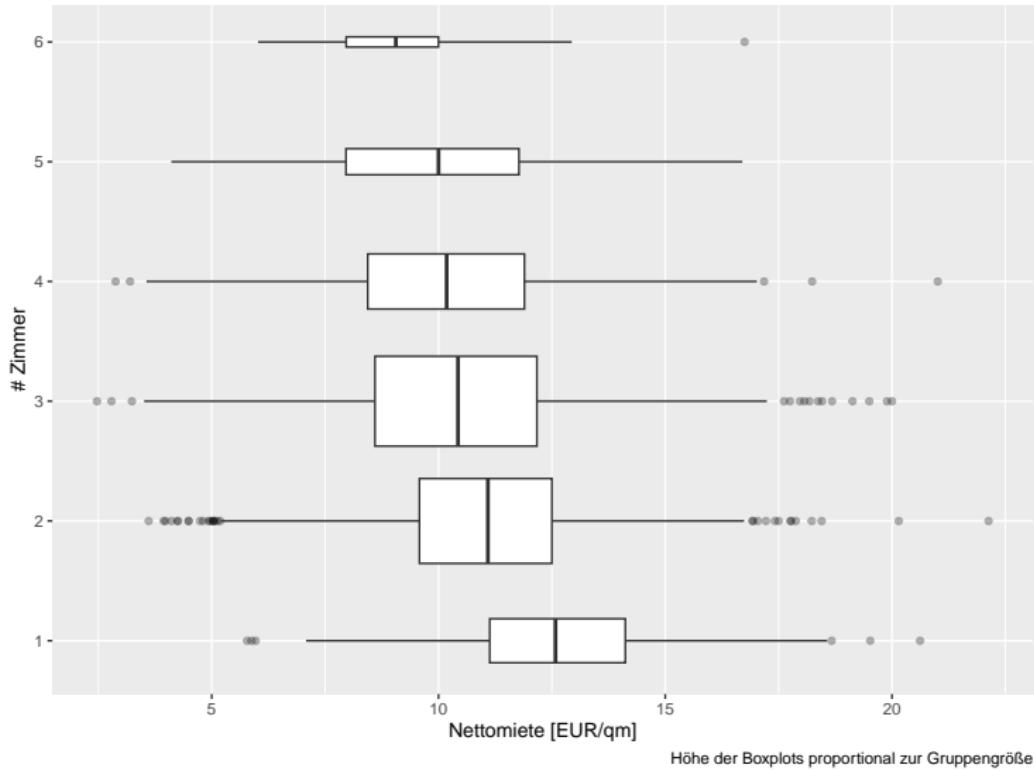
Modifizierter Boxplot:

- ▶ whiskers 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.
Üblicherweise: $z_u = \tilde{x}_{0.25} - 1.5d_Q$, $z_o = \tilde{x}_{0.75} + 1.5d_Q$
- ▶ Ansonsten gehen die Linien nur bis zum kleinsten bzw. größten Wert innerhalb der Zäune, die außerhalb liegenden Werte werden individuell als Punkte/Symbole eingezeichnet.

Boxplot: Beispiel Quadratmetermiete Mietspiegel



Gruppiertter Boxplot:



Gruppengrößen darstellbar über Höhe (horizontale Box) bzw. Breite (vertikale B.).

Boxplot: Vor- und Nachteile

+:

- ▶ kompakt
- ▶ geeignet für Vergleiche
- ▶ zentraler Bereich der Daten einfach ablesbar
- ▶ **Ausreißer** sichtbar
- ▶ **Schiefe** sichtbar

-:

- ▶ gegen Intuition (Viel Farbe – wenig Daten) da Ausreißer sehr prominent
- ▶ **Multimodale** Verteilungen nicht sichtbar
- ▶ (Höhe der Box trägt keine Information)

Einfacher Boxplot: Grammar of Graphics

verwendete Geometrie:

Rechteck mit vertikalem Band und horizontalen “whiskers”.

zugeordnete Ästhetiken:

- ▶ horizontale Ausdehnung des Rechtecks: von $\tilde{x}_{0.25}$ bis $\tilde{x}_{0.75}$
- ▶ horizontale Position des vertikalen Bands: \tilde{x}_{med}
- ▶ horizontale Positionen der Enden der “whiskers”: x_{min} bzw. x_{max}
- ▶ optional: vertikale Ausdehnung der Box: Anzahl der eingeflossenen Beobachtungen

⇒ zugeordnete Ästhetiken in statistischen Grafiken repäsentieren oft zusammenfassende Kennwerte der zugrundeliegenden Daten!

(hier nur für *einfachen, horizontale* Boxplots, für vertikale Boxplots analog mit vertikaler Position.)

Boxplot: Grammar of Graphics

- ▶ Für *modifizierten* Boxplot zusätzlich:
 - ▶ Geometrie “Punkte/Symbole” für Beobachtungen außerhalb der Zäune
 - ▶ entsprechend modifizierte Position der “whiskers”
- ▶ Alternative “Boxplots” auf Basis anderer Lage- und Streuungsmaße oder anderer Definition der “Zäune” auch zulässig und oft sinnvoll.

Der Mittelwert (arithmetisches Mittel)

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

- ▶ bekanntestes Lagemaß
- ▶ stark beeinflusst von extremen Werten
- ▶ geeignet für intervallskalierte Daten
- ▶ “Durchschnitt” oder “Schwerpunkt” der Daten

Formal:

- ▶ Wert, der die Summe der *quadrierten* Differenzen zu den beobachteten Werten minimiert:

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

Mittelwert bei gruppierten Daten

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

h_j : Häufigkeit von a_j

Gewichteter Mittelwert

Allgemeiner gilt:

Der *gewichtete Mittelwert* mit Gewichten $w_i \geq 0, i = 1, \dots, n$ ist

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

Das geometrische Mittel

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

- ▶ rücktransformiertes arithmetisches Mittel auf der log-Skala:

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

- ▶ nur geeignet für positive, mindestens intervallskalierte Merkmale
- ▶ Anwendung: Durchschnitt von Änderungsraten, z.B. durchschnittliche Verzinsung

Allgemein: Das geometrische Mittel ist der richtige Mittelwert für Merkmale, deren Ausprägungen als multiplikative Faktoren zu interpretieren sind.

Das geometrische Mittel: Beispiel

Anfangskapital 100 €, Entwicklung über 3 Quartale

- ▶ Q1: 20% Gewinn (Faktor 1.2)
- ▶ Q2: 50% Gewinn (Faktor 1.5)
- ▶ Q3: 40% Verlust (Faktor 0.6)

```
r <- c(1.2, 1.5, 0.6)
cumprod(c(100, r)) # Kapital nach jedem Quartal
## [1] 100 120 180 108

mean(r) # arithmetisches Mittel der Renditen
## [1] 1.1 # --> "durchschnittlich" 10% Gewinn pro Quartal ?

exp(mean(log(r))) # geom. Mittel der Renditen
## [1] 1.026 # --> "durchschnittlich" 2.6% Gewinn pro Quartal !

100 * mean(r)^3 # "effektive" Verzinsung pro Quartal in%??
## [1] 133.1
# 10% pro Quartal entspräche 100 -> 133 über 3 Quartale, also falsch

100 * exp(mean(log(r)))^3 # effektive Verzinsung pro Periode in %!!
## [1] 108
# 2.6% pro Quartal ergibt 100 -> 108 über 3 Quartale, also korrekt
```

Das harmonische Mittel

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

Das harmonische Mittel entspricht dem inversen Mittelwert der inversen Werte:

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

- ▶ Anwendung für Mittelwerte von Quotienten/Verhältnissen unterschiedlicher Skalen, z.B.
 - ▶ Geschwindigkeiten (in $\frac{km}{h}, \frac{m}{s}$ etc)
 - ▶ Kurs-Gewinn-Verhältnisse (PE ratios, in $\frac{\text{Aktienkurs in €}}{\text{Gewinn in €}}$)

Bsp: harmonisches Mittel

Sie fahren die ersten 100 km mit 30km/h und die zweiten 100km mit 100 km/h.

Was ist Ihre Durchschnittsgeschwindigkeit?

Gesamtstrecke: 200 km

Gesamtzeit: $3.33h + 1h = 4.33h$

$$\Rightarrow \frac{200\text{km}}{4.33h} = 46.2\text{km/h.}$$

Arithm. Mittel:

$$\bar{x} = \frac{30+100}{2} = 65, \text{ wäre also } 65 \text{ km/h und offensichtlich falsch.}$$

Harmonisches Mittel:

$$\bar{x}_H = \frac{1}{\frac{1}{30} + \frac{1}{100}} = \frac{600}{13} = 46.2, \text{ also } 46.2 \text{ km/h!}$$

Allgemeine Transformation des Arithm. Mittels

Lineare Transformation:

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

d.h.

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

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

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 "unten" und "oben" jeweils der Anteil α der extremsten Werte abgeschnitten:

α -getrimmtes Mittel

Alternative:

Winsorisiertes Mittel (gestutztes Mittel):

Der Anteil α der extremsten Werte wird **durch die entsprechenden oberen/unteren Quantile ersetzt.**

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Einleitung

Bereits kennengelernt:

- ▶ *Empirische Lage- und Streuungsmaße*
(Mittelwerte, Median, Modus, Stichprobenvarianz, MAD, ...) die Verteilung der *Beobachtungen eines Merkmals in einer Stichprobe* beschreiben.

Jetzt:

- ▶ *Theoretische Entsprechungen für Verteilungen von Zufallsvariablen.*

(Spätere Veranstaltungen: Schätztheorie – welche empirischen Maße sind wie gut geeignet um bestimmte theoretische Größen aus Daten zu schätzen?)

Modus einer Zufallsvariablen

Definition: Modus einer Zufallsvariable

Der Modus einer Zufallsvariable X ist ein Wert x_{Mod} , für den gilt:

$$f(x_{Mod}) \geq f(x) \quad \forall x \in T_X$$

⇒ Modi (auch: Modalwerte) sind die globalen Maximumsstellen der Dichte-/Wahrscheinlichkeitsfunktion von X .

- ▶ Genau wie der empirische Modus eines beobachteten Merkmals ist der Modus einer ZV weder notwendigerweise eindeutig noch muss er existieren.

Median und Quantile einer Zufallsvariablen

Mediane und andere Quantile einer Zufallsvariable X sind inhaltlich analog zu ihren empirischen Analoga definiert – es gilt:

Definition: Quantil einer Zufallsvariable

Das p -Quantil \tilde{x}_p einer ZV X , mit $p \in [0, 1]$, ist

$$\tilde{x}_p := \arg \min_{x \in T_X} \{x : F_X(x) \geq p\}.$$

Also:

- ▶ der kleinste Wert für den die Verteilungsfunktion mindestens den Wert p annimmt.
- ▶ der kleinste Wert der mindestens mit Wahrscheinlichkeit p unterschritten wird
- ▶ für stetige ZV mit strikt positiver Dichte bzw. streng monotoner Verteilungsfunktion gilt:

$$\tilde{x}_p = F_X^{-1}(p)$$

► $F_X^{-1}(p)$ heißt auch Quantilfunktion von X .

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Erwartungswert einer ZV

Definition: Diskreter Erwartungswert

Der Erwartungswert (EW) $E(X)$ einer diskreten ZV X mit Träger T_X ist definiert als

$$\begin{aligned} E(X) &:= \sum_{x \in T_X} x \cdot P_X(X = x) = \sum_{\omega \in \Omega} P(\{\omega\}) X(\omega) \\ &= \sum_{x \in T_X} x \cdot f_X(x) \end{aligned}$$

Definition: Stetiger Erwartungswert

Der Erwartungswert $E(X)$ einer stetigen ZV X mit Dichtefunktion $f_X(x)$ ist definiert als

$$E(X) := \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

Erwartungswert einer ZV

Intuition:

Der Erwartungswert ist *gewichtetes arithmetisches Mittel* der möglichen Werte einer ZV, wobei die Gewichte den Wahrscheinlichkeiten entsprechen, einen bestimmten Wert zu beobachten.

- ▶ Der Erwartungswert existiert bzw. ist endlich, wenn gilt:
 - ▶ für diskrete ZV: $\sum_{x \in T_X} |x| \cdot f(x) < \infty$ (absolut konvergente Summe),
 - ▶ für stetige ZV: $\int_{-\infty}^{\infty} |xf(x)| dx = \int_{-\infty}^{\infty} |x|f(x) dx < \infty$ (absolut integrierbar).
- ▶ Beachte: $\sum_{x \in T_X} x \cdot P(X = x) = \sum_{x \in \mathbb{R}} x \cdot P(X = x)$, da $P(X = x) = 0 \forall x \notin T_X$

Eigenschaften des Erwartungswertes

- Sei $X = a$ mit Wahrscheinlichkeit 1 ("deterministische ZV"). Dann gilt:

$$E(X) = a$$

- **Linearität des Erwartungswertes:**

Sei $a, b \in \mathbb{R}$ und X, Y beliebige ZV. Dann gilt:

$$E(a \cdot X + b \cdot Y) = a \cdot E(X) + b \cdot E(Y)$$

⇒ Für beliebige $a, b \in \mathbb{R}$ gilt daher

$$E(aX + b) = a \cdot E(X) + b$$

- Ist $f(x)$ symmetrisch um einen Punkt c , d.h. $f(c - x) = f(c + x) \forall x \in T_X$, dann ist $E(X) = c$.

Eigenschaften des Erwartungswertes II

Allgemeiner auch für beliebige $a_1, \dots, a_n \in \mathbb{R}$ und beliebige ZV X_1, \dots, X_n :

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \cdot E(X_i)$$

Transformationsregel für Erwartungswerte

Sei X eine ZV und $g : \mathbb{R} \rightarrow \mathbb{R}$ eine reelle Funktion. Dann gilt für $Y = g(X)$:

$$E(Y) = E[g(X)] = \begin{cases} \sum_{x \in T} g(x) f(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & X \text{ stetig} \end{cases}$$

- Im Allgemeinen gilt **nicht**: $E(g(X)) = g(E(X))!$

Beispiel

Sei X eine ZV mit folgender Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} 1/4 & \text{für } x = -2 \\ 1/8 & \text{für } x = -1 \\ 1/4 & \text{für } x = 1 \\ 3/8 & \text{für } x = 3 \end{cases}$$

Berechne den Erwartungswert von $E(X^2)$

Erwartungswert von ZVn mit Träger \mathbb{N}^+

Für ZV X mit Träger $T_X = \mathbb{N}^+$ gilt: $E(X) = \sum_{k=1}^{\infty} P(X \geq k)$

Beweis:

$$\begin{aligned}\sum_{k=1}^{\infty} P(X \geq k) &= \sum_{k=1}^{\infty} \sum_{t=k}^{\infty} P(X = t) \\&= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + \dots \\&\quad + P(X = 2) + P(X = 3) + P(X = 4) + \dots \\&\quad + P(X = 3) + P(X = 4) + \dots \\&\quad + P(X = 4) + \dots \\&\quad + \dots \dots \dots \\&= \sum_{t=1}^{\infty} t \cdot P(X = t) \\&= E(X)\end{aligned}$$

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Maße für die Streuung

- ▶ Spannweite
- ▶ Interquartilsabstand
- ▶ Standardabweichung und Varianz
- ▶ Variationskoeffizient

Die Spannweite (Range)

Definition:

$$sp = x_{\max} - x_{\min}$$

- ▶ Größe des Intervalls in dem die Daten liegen
- ▶ Anwendung primär für Kontrolle der Datenqualität:
Plausibilität, Eingabe-/Codierungsfehler, Existenz von Fehlmessungen und Ausreißern
- ▶ extrem sensibel gegen Ausreißer

Der Quartilsabstand

Definition:

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

- ▶ Größe des Bereichs in dem die “mittlere Hälfte” der Daten liegt
 - ▶ Länge der Box des Boxplots
- ▶ Alternativer Name: *interquartilerange*: IQR
- ▶ Robust gegen Ausreißer
- ▶ Bei ordinal skalierten Daten Angabe von $x_{0.75}$ und $x_{0.25}$: definiert Bereich der zentralen 50% der Daten.

Standardabweichung und Stichprobenvarianz

Definition:

$$\text{Stichprobenvarianz } s_x^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Standardabweichung } s_x := \sqrt{s_x^2}$$

- ▶ s_x lose interpretierbar als “mittlere Abweichung vom Mittelwert”
- ▶ Mindestens Intervallskala
- ▶ Empfindlich gegen Ausreißer
- ▶ Verwende $\tilde{s}_x^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ für Vollerhebungen, Division durch $n - 1$ nur bei Stichproben sinnvoll

Transformationsregel

$$y_i = a + bx_i \implies \tilde{s}_y^2 = b^2 \tilde{s}_x^2$$
$$\tilde{s}_y = |b| \tilde{s}_x$$

(Analog für s_x, s_y)

Varianz und Standardabweichung sind also für lineare Transformationen einfach umrechenbar.

Verschiebungssatz:

Für jedes $c \in \mathbb{R}$ gilt:

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

Setze $c = 0 \implies \tilde{s}_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$

$$\tilde{s}_x^2 = \overline{x^2} - \bar{x}^2$$

Beachte:

Der Verschiebungssatz ist für die Berechnung präziser numerischer Ergebnisse am Computer **nicht geeignet** (Auslöschung von Gleitkommazahlen).

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

$$x_1, \dots, x_{n_1}, x_{n_1+1}, \dots, x_{n_1+n_2}, \dots, x_n$$

$$\text{mit } n = \sum_{j=1}^r n_j$$

Schichtmittelwerte:

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

Schichtvarianzen:

$$\tilde{s}_{x1}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2, \quad \tilde{s}_{x2}^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2, \text{ usw.}$$

Streuungszerlegung II

Dann gilt, mit $\bar{x} = \frac{1}{n} \sum_{j=1}^r n_j \bar{x}_j$:

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

Gesamtstreuung = Streuung innerhalb + Streuung zwischen
 der Schichten den Schichten

Streuungszerlegung: Nettomiete & Zimmerzahl

Streuungszerlegung der Netto-Quadratmetermiete bezüglich Zimmerzahl:

Zimmer	n_j	\bar{x}_j	\tilde{s}_{xj}^2
8	2	6.2	1.2
6	16	10.0	13.2
5	99	9.9	7.3
4	442	10.2	7.0
3	1175	10.4	7.1
2	1049	10.9	6.1
1	282	12.6	6.2

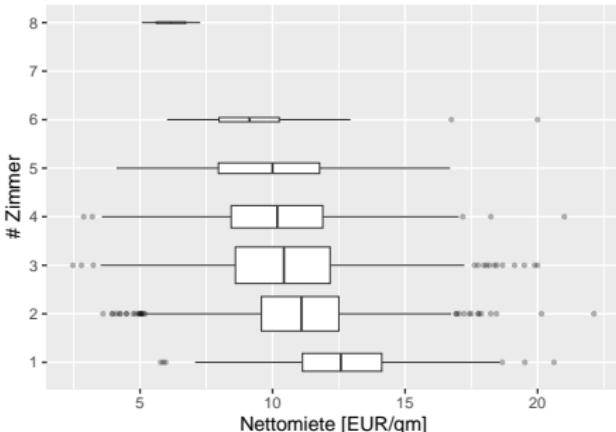
Gesamtvarianz: $\tilde{s}_x^2 \approx 7.15$

Innerhalb: $\frac{1}{n} \sum_{j=1}^r n_j \tilde{s}_{xj}^2 = 6.7$

Zwischen:

$\frac{1}{n} \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2 = 0.45$

\vskip 1.5em \Rightarrow nur $\frac{0.45}{7.15} = 6.25\%$ der Gesamtvarianz der Quadratmetermiete entfallen auf Unterschiede zwischen Wohnungen mit unterschiedlicher Zimmerzahl.



Höhe der Boxplots proportional zur Gruppengröße

Variationskoeffizient

Das Verhältnis von Standardabweichung und Mittelwert ist gegeben durch

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

Der Variationskoeffizient hat keine Einheit und ist skalenunabhängig.

Er ist eine Maßzahl für die relative Schwankung um den Mittelwert und nur für Merkmale mit positiven Ausprägungen sinnvoll.

Mittlere absolute Abweichung (MAD)

Die **mittlere absolute Abweichung** (*mean absolute deviation*) ist definiert als

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

Es gilt: $\text{MAD}_x \leq \tilde{s}_x$

Der **Median der absoluten Abweichungen** ist

$$\text{MedAD}_x := \text{Median}(|x_i - x_{med}|)$$

- ▶ MAD_x einfacher interpretierbar als s_x
- ▶ beide weniger ausreißer-empfindlich, vor allem MedAD_x
- ▶ beide weniger "schöne" theoretische Eigenschaften als s_x und werden deswegen seltener angewendet

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Varianz

Def.: Varianz einer ZV

Die Varianz $\text{Var}(X)$ einer ZV X ist definiert als:

$$\begin{aligned}\text{Var}(X) &:= E[(X - E(X))^2] \\ &= \begin{cases} \sum_{x \in T_X} (x - E(X))^2 f_X(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx & X \text{ stetig} \end{cases}\end{aligned}$$

⇒ “erwartete quadratische Abweichung vom Erwartungswert”

Beachte:

- ▶ Varianz nicht immer endlich! ($\text{Var}(X) = \infty \iff \text{Varianz existiert nicht.}$)
- ▶ Existiert der Erwartungswert nicht, so existiert auch die Varianz nicht.

Eigenschaften der Varianz

- Zur einfacheren Berechnung kann man häufig den **Verschiebungssatz** verwenden:

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

- $\text{Var}(aX + b) = a^2 \text{Var}(X) \quad \forall a, b \in \mathbb{R}$
- für unabhängige Zufallsvariablen X, Y gilt: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

Ungleichung von Tschebyscheff

Tschebyscheff-Ungleichung

Sei X eine beliebige ZV. Dann gilt:

$$P(|X - E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

Beispiel:

Für $\text{Var}(X) = 1$ und beliebiges $E(X)$ gilt also

$$P(|X - E(X)| \geq 1) \leq 1$$

$$P(|X - E(X)| \geq 2) \leq \frac{1}{4}$$

$$P(|X - E(X)| \geq 3) \leq \frac{1}{9}$$

Standardabweichung

Def.: Standardabweichung einer ZV

Die *Standardabweichung* $\sigma(X)$ einer ZV X ist die positive Wurzel ihrer Varianz:

$$\sigma(X) = +\sqrt{\text{Var}(X)}$$

Es gilt:

$$\sigma(aX + b) = |a| \cdot \sigma(X)$$

Die *erwartete absolute Abweichung* $E(|X - E(X)|)$ wäre zwar das intuitivere Streuungsmaß, ist aber mathematisch deutlich schwieriger zu handhaben.

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

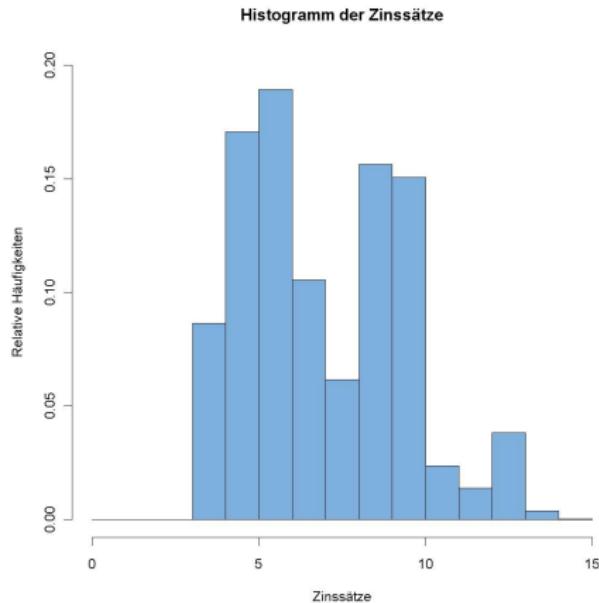
Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

Uni- und multimodale Verteilungen

unimodal = eingipflig, **multimodal** = mehrgipflig



Das Histogramm der Zinssätze zeigt eine bimodale (trimodale...?) Verteilung.

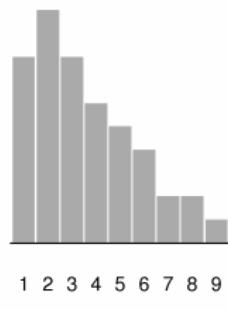
Symmetrie und Schiefe I

symmetrisch \Leftrightarrow Rechte und linke Hälften der Verteilung sind annähernd zueinander spiegelbildlich

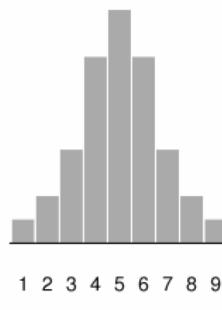
**linkssteil
(rechtsschief)** \Leftrightarrow Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab

**rechtssteil
(linksschief)** \Leftrightarrow Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab

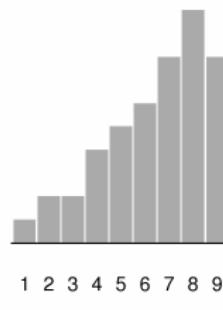
Symmetrie und Schiefe II



(a)



(b)



(c)

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

Lageregeln

- ▶ Symmetrische und unimodale Verteilung:

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

- ▶ Linksssteile Verteilung: $\bar{x} > x_{med} > x_{mod}$

- ▶ Rechtssteile Verteilung: $\bar{x} < x_{med} < x_{mod}$

- ▶ Bei gruppierten Daten: Auch für Histogramme gültig

Beachte:

- ▶ Lageregeln sind Daumenregeln, gelten nicht notwendigerweise in jedem Einzelfall.
- ▶ Form der Verteilung bleibt bei linearen Transformationen gleich, aber ändert sich bei nichtlinearen Transformationen.

Maßzahlen für die Schiefe I

Quantilskoeffizient:

$$g_p = \frac{(\tilde{x}_{1-p} - \tilde{x}_{med}) - (\tilde{x}_{med} - \tilde{x}_p)}{\tilde{x}_{1-p} - \tilde{x}_p}; \quad \text{mit } 0 < p < 0.5$$

$g_{0.25}$ bezeichnet man als **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:

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

bzw.

$$g_m = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

Werte des Momentenkoeffizienten:

$g_m = 0$ für symmetrische Verteilungen

$g_m > 0$ für linkssteile Verteilungen

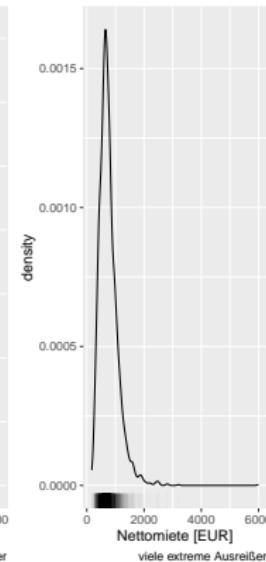
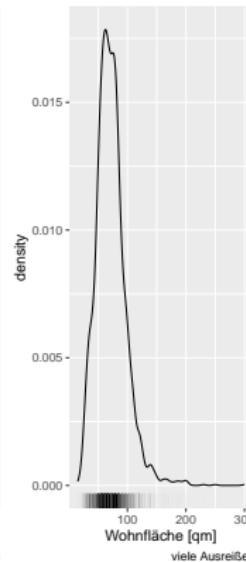
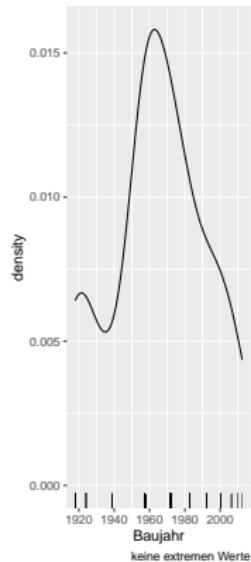
$g_m < 0$ für rechtssteile Verteilungen

Wölbung und Extremwerte

Sehr wichtige Verteilungseigenschaft für viele Anwendungsbereiche:

Wie häufig oder selten treten extreme Werte auf, die weit entfernt von der "Mitte" der Verteilung liegen?

Wie häufig sind Beobachtungen aus den Rändern der Verteilung (im Vergleich zur "Mitte")?



Wölbung und Extremwerte

⇒ Wölbung/**Kurtosis** misst Häufigkeit (und Distanz) extremer Werte:

$$\begin{aligned} k &= \frac{m_4}{s_x^4} \quad \text{mit } m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^4 \end{aligned}$$

Meist betrachtet: **Exzess-Kurtosis**

$$k^* = k - 3$$

\vskip -1em

- ▶ $k^* \approx 0$: normalgipflig (vgl. *Normalverteilung*), *mesokurtisch*.
- ▶ $k^* > 0$: steilgipflig, *leptokurtisch*.
Stärker ausgeprägter schmaler Gipfel im Vergleich zu den Rändern,
häufigere extreme Werte
- ▶ $k^* < 0$: flachgipflig, *platykurtisch*.
Wenig ausgeprägter breiterer Gipfel im Vergleich zu den Rändern,
seltener extreme Werte

Wölbung und Extremwerte

Klassisches theoretisches Schaubild:

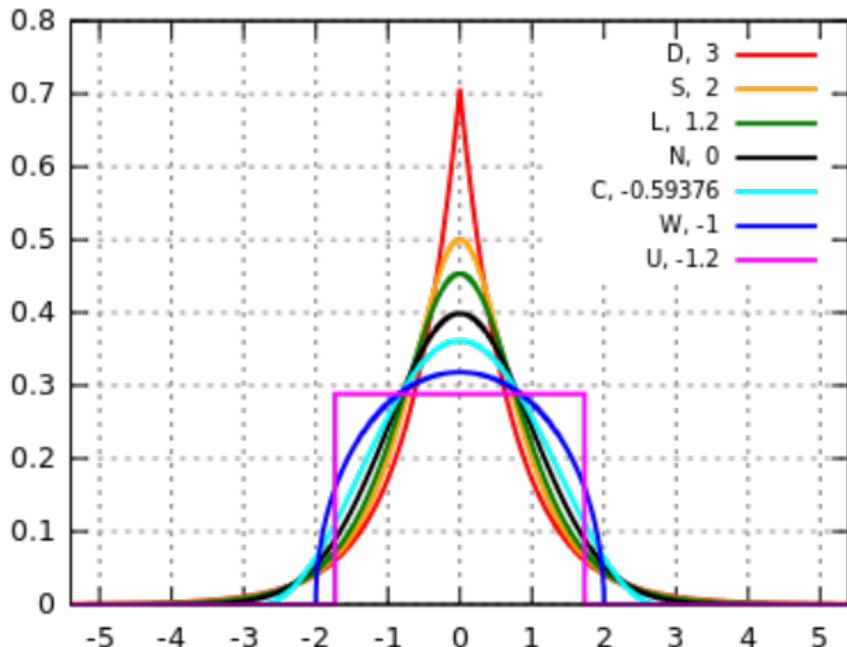
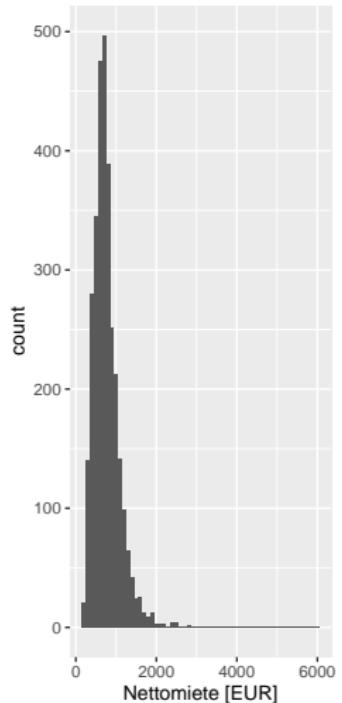
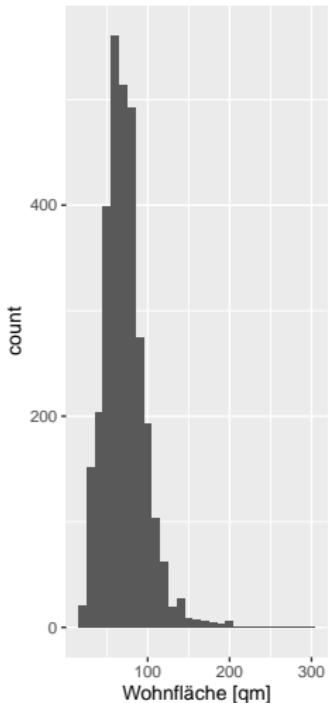
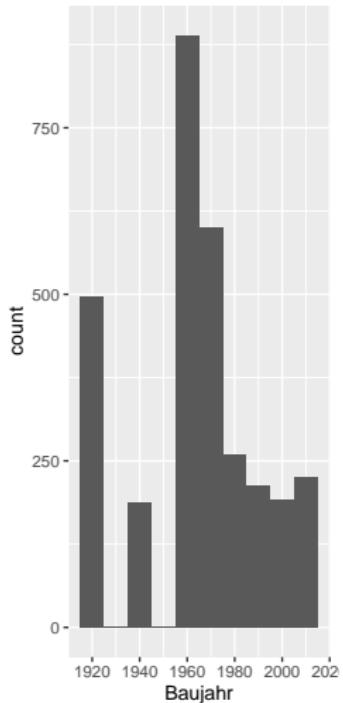


Abb: Wikimedia Commons

Wölbung und Extremwerte

Praktische Beispiele:



Momente von Zufallsvariablen

Def.: Moment einer ZV

Das k -te Moment einer ZV X ist $E(X^k)$

Def.: Zentriertes Moment einer ZV

Das k -te zentrierte Moment einer ZV X ist $E((X - E(X))^k)$

- Der Erwartungswert ist das *erste Moment* einer Verteilung
- Die Varianz ist das *zweite zentrierte Moment* einer Verteilung

Höhere Momente von Zufallsvariablen

Theoretische Analoga zu empirischem Momentenkoeffizient der Schiefe und Wölbung:

- ▶ Die **Schiefe** einer Verteilung ist das dritte (zentrale) Moment ihrer standardisierten Werte: $E \left[\left(\frac{X - E(X)}{\sigma(X)} \right)^3 \right]$
- ▶ Die **Wölbung** (Kurtosis) einer Verteilung ist das vierte (zentrale) Moment ihrer standardisierten Werte: $E \left[\left(\frac{X - E(X)}{\sigma(X)} \right)^4 \right]$
- ▶ Alle geraden zentralen Momente sind im Endeffekt "Varianzen", die zunehmend mehr Gewicht auf die extremen Ränder der Verteilung legen
- ▶ Alle ungeraden zentralen Momente (außer dem ersten) messen die Asymmetrie von Verteilungen, mit immer größerem Gewicht auf Asymmetrie in den extremen Rändern der Verteilung

Theorie & Empirie

Theorie (ZV)		Empirie (beob. Merkmale)
$E(X) := \int x f(x) dx$	\leftrightarrow	$\bar{x} := 1/n \sum_{i=1}^n x_i = \sum_{j=1}^k a_j f_j$
$\text{Var}(X) := \int (x - E(X))^2 f(x) dx$	\leftrightarrow	$s_x^2 := 1/(n-1) \sum_{i=1}^n (x_i - \bar{x})^2$
Median $x_{\text{med}} : F_X(x_{\text{med}}) = 0.5$	\leftrightarrow	$\tilde{x}_{\text{med}} := x_{(\lceil n/2 \rceil)}$
Modus $x_{\text{mod}} := \arg \max(f_X(x))$	\leftrightarrow	$\tilde{x}_{\text{mod}} : h(x_{\text{mod}}) \geq h(a_j) \forall j = 1, \dots, k$

etc etc

Kennwerte & Verteilungseigenschaften

Univariate statistische Kennwerte

Statistische Kennwerte: Lagemaße

Lagemaße für Zufallsvariablen

Erwartungswert einer Zufallsvariablen

Statistische Kennwerte: Streuungsmaße

Varianz einer Zufallsvariable

Statistische Kennzahlen: Verteilungseigenschaften

Statistische Kennzahlen: Konzentrationsmaße

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 Marktsegment

Lorenzkurve

Grundidee:

Es sollen folgende Aussagen grafisch dargestellt werden:

- ▶ Die “Ärmsten”/“Kleinsten” x% besitzen einen Anteil von y% der gesamten Merkmalssumme.
- ▶ Die “Reichsten”/“Größten” x% besitzen einen Anteil von y% der gesamten Merkmalssumme.

Lorenzkurve

Definition:

- ▶ Das Merkmal darf nur *positive* Ausprägungen annehmen.
- ▶ Die Gesamtsumme aller Merkmalswerte ist $\sum_{i=1}^n x_i = \sum_{i=1}^n x_{(i)}$.
- ▶ Die Lorenzkurve verbindet Punktepaare bestehend aus den *Teilsummen* der *nach Größe geordneten* Beobachtungswerte $0 \leq x_{(1)} \leq \dots \leq x_{(n)}$ und dem *relativen Anteil* der Individuen, die diese Teilsumme besitzen.

Lorenzkurve

- ▶ Es wird festgelegt: $u_0 = 0$ und $v_0 = 0$.
- ▶ Die horizontale Achse wird in *gleiche Längen* aufgeteilt, deren Anzahl der Individuen (Merkmalsausprägungen) entspricht:

$$u_j = \frac{j}{n}, \quad j = 1, \dots, n.$$

- ▶ Die Werte auf der vertikalen Achse werden wie folgt berechnet:

$$v_j = \frac{\sum_{i=1}^j X_{(i)}}{\sum_{i=1}^n X_{(i)}}, \quad j = 1, \dots, n,$$

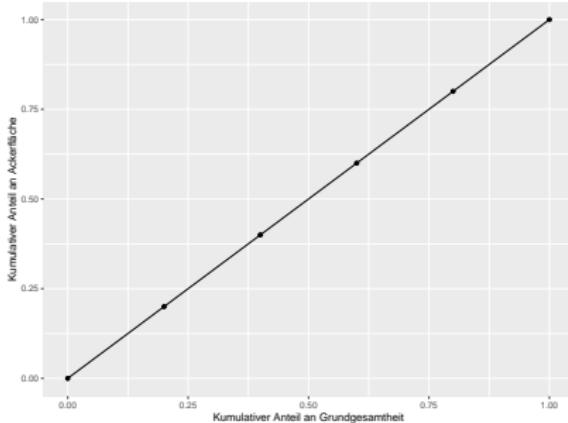
also: Quotienten aus der Teilsumme und der Gesamtsumme.

- ▶ Die so errechneten Koordinatenpunkte (u_j, v_j) werden in den Graphen eingetragen und mit Linien verbunden.

Lorenzkurve

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

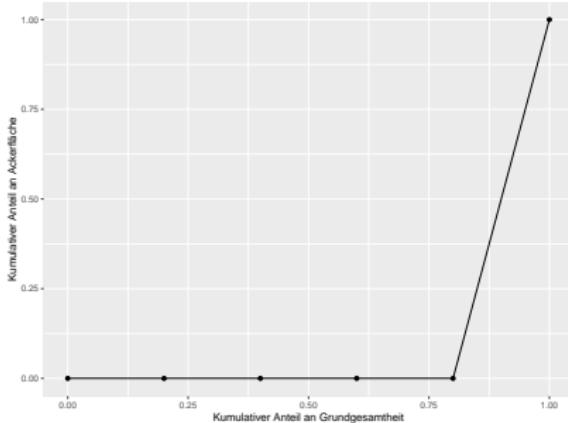
j	$x_{(j)}$	u_j	v_j
0	-	0	0
1	20	1 5 <hr/> 100	20 <hr/> 100
2	20	2 5 <hr/> 100	40 <hr/> 100
3	20	3 5 <hr/> 100	60 <hr/> 100
4	20	4 5 <hr/> 100	80 <hr/> 100
5	20	5 5 <hr/> 100	100 <hr/> 100



Lorenzkurve

Beispiel: 4 Bauern besitzen nichts, 1 besitzt alles:

j	$x_{(j)}$	u_j	v_j
0	-	0	0
1	0	5	$\frac{0}{100}$
2	0	25	$\frac{0}{100}$
3	0	35	$\frac{0}{100}$
4	0	45	$\frac{0}{100}$
5	100	55	$\frac{100}{100}$



Lorenzkurve

Erscheinungsbild von Lorenzkurven:

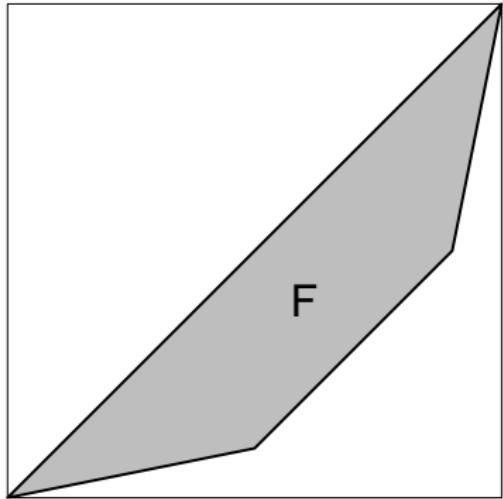
- ▶ Die Koordinate $(u_0; v_0)$ ist *immer* $(0; 0)$.
- ▶ Die Koordinate $(u_n; v_n)$ ist *immer* $(1; 1)$.
- ▶ Der konstruierte Polygonzug verläuft *immer unterhalb* (im Grenzfall auf) der Winkelhalbierenden.
- ▶ Der konstruierte Polygonzug ist (*streng*) *monoton steigend*.
- ▶ Die Steigung des nächsten Polygonsegments ist entweder *gleich groß* oder *größer* als die Steigung des vorherigen Polygonsegments.

Gini-Koeffizient

Der **Gini-Koeffizient** bzw. das **Lorenz'sche 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:

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

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

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.

Eigenschaften des Gini-Koeffizient

- ▶ Sehr unterschiedliche Lorenzkurven führen zum selben Gini-Wert, z.B: $G^+ = 0.5$ für
 - ▶ Verteilung 1: 50% ohne Anteile, restliche 50% haben gleiche Anteile an Gesamtsumme
 - ▶ Verteilung 2: untere 75% teilen sich gleichmäßig 25% der Gesamtsumme, obere 25% gleichmäßig 75%
- ▶ Unempfindlich gegen relatives Wachstum: G^+ -Wert bleibt unverändert falls alle x -Werte um den selben Faktor wachsen/schrumpfen
- ▶ “Empfindlich” gegen Ausreißer am oberen Ende
 - ⇒ niedrige Gini-Werte für große Grundgesamtheiten unwahrscheinlicher selbst wenn Verteilung ähnlich zu der in kleinerer Grundgesamtheit, da größere Grundgesamtheiten häufiger extremere Ausreißer enthalten
(z.B. US-Amerikaner \ni Jeff Bezos \gg Susanne Klatten \in Deutsche)
- ▶ Häufig Anwendung von verallgemeinerten Lorenzkurven bzw. Ginikoeffizienten auch auf Merkmale mit teils negativen Ausprägungen (z.B. Schulden als negatives Vermögen, s. übernächste Folie) - dann keine Monotonie etc mehr.

Herfindahl-Index

$x_1, \dots x_n$ 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

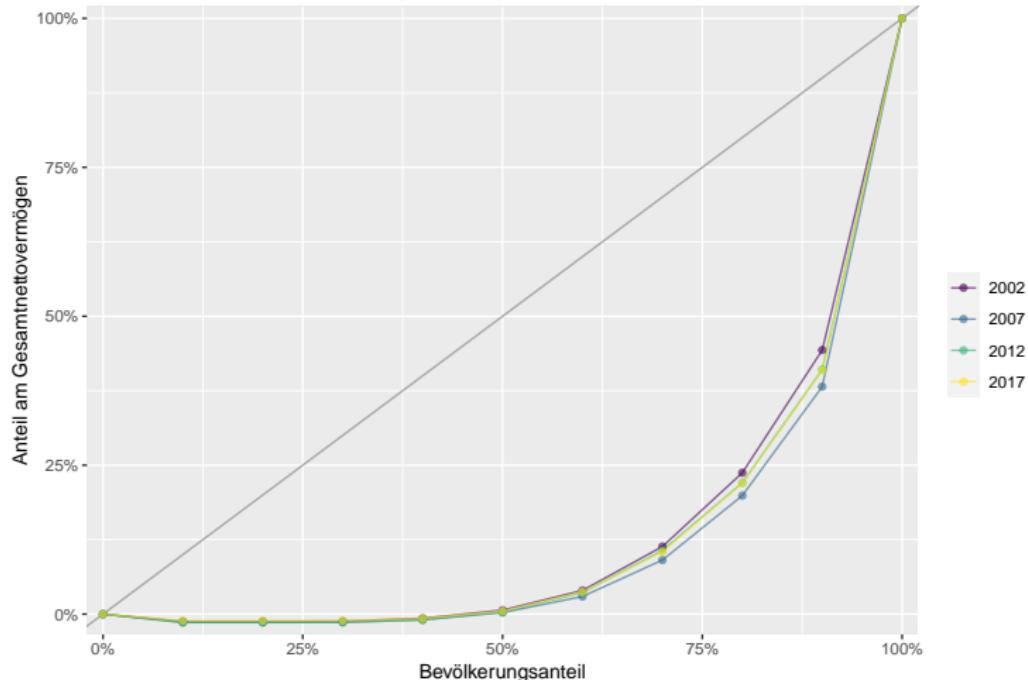
$$H := \sum_{i=1}^n p_i^2$$

Der Wertebereich ist von $\frac{1}{n}$ (identische x_i) bis 1 (Monopol)

Konzentrationsmessung: Deutsche Vermögensverteilung

Lorenzkurve der individuellen Nettovermögen in Deutschland

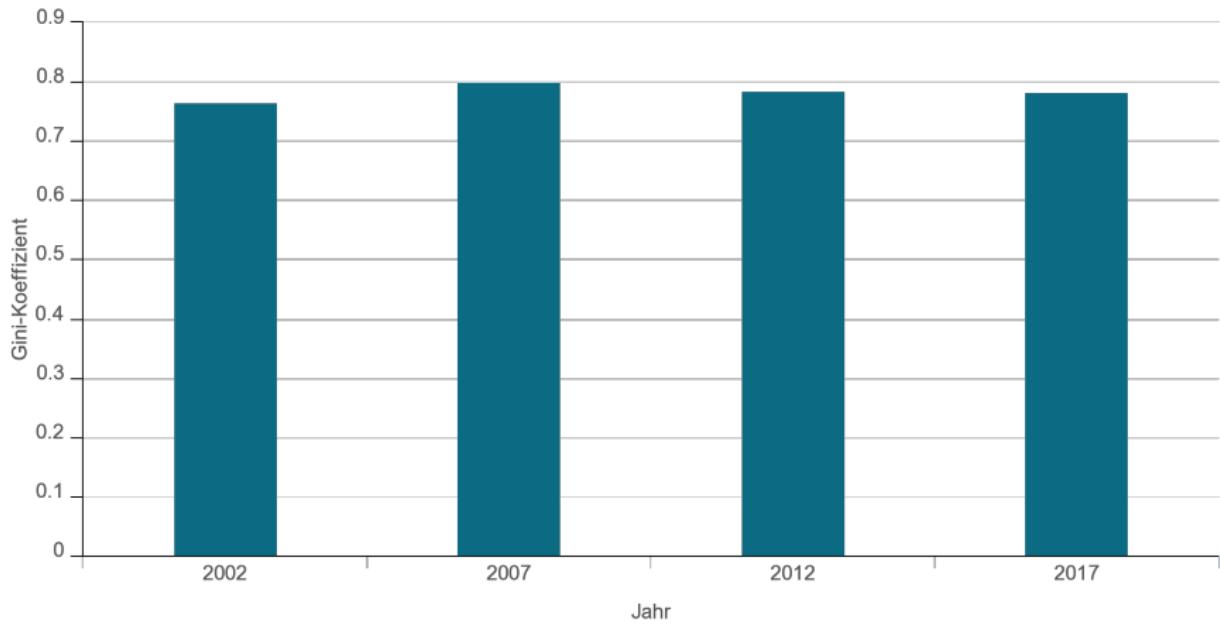
Gini-Koeffizienten: ca. 0.77 – 0.80



Daten: www.armuts-und-reichtumsbericht.de / BM Arbeit & Soziales, 4.10.2021 (nur Dezile)

Gini-Index Vermögen Deutschland

Vermögensverteilung individuelle Nettovermögen nach SOEP

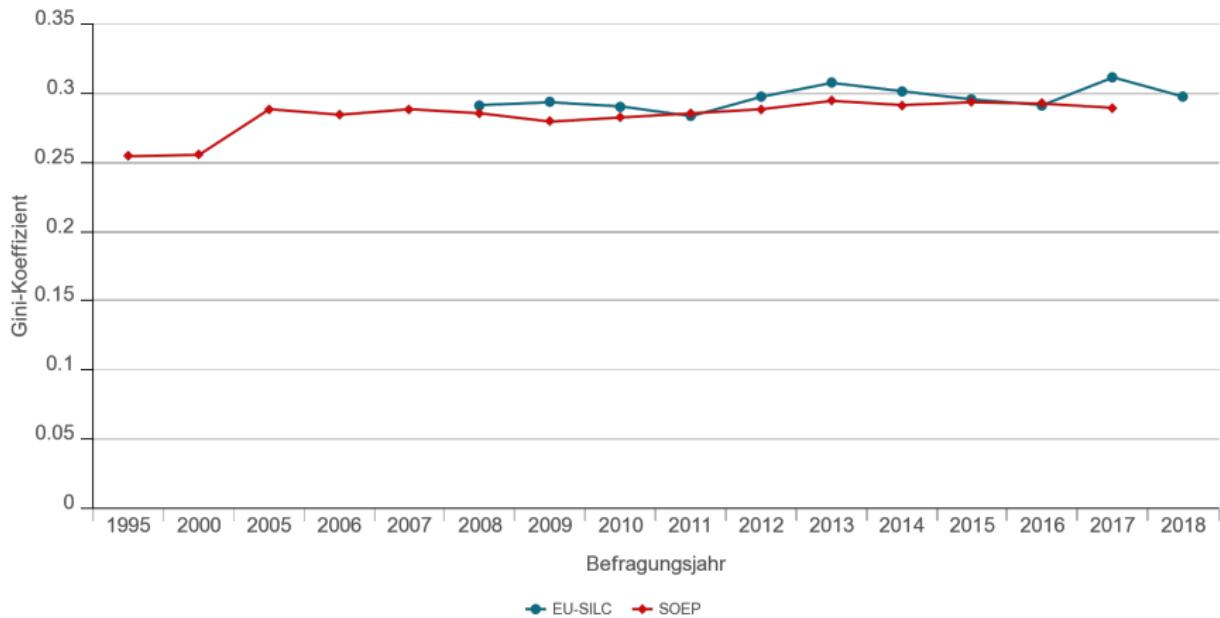


© Bundesministerium für Arbeit und Soziales

Armut- & Reichtumsbericht, BM Arbeit & Soziales

Gini-Index Einkommen Deutschland

Einkommensverteilung Gini-Koeffizient



© Bundesministerium für Arbeit und Soziales

Armut- & Reichtumsbericht, BM Arbeit & Soziales

Wichtige parametrische Verteilungen

Diskrete parametrische Verteilungen

Stetige parametrische Verteilungen

Dichtetransformationssatz

Wichtige parametrische Verteilungen

Diskrete parametrische Verteilungen

Stetige parametrische Verteilungen

Dichtetransformationssatz

Parametrische Verteilungen

Statistik benutzt viele “gängige” Klassen von Verteilungen,

- ▶ die bestimmte Arten von häufigen, idealtypischen Zufallsvorgängen (zB “Alles gleich wahrscheinlich”, “Ziehen mit/ohne Zurücklegen”) formalisieren und
- ▶ die meist von weiteren Parametern, die das konkrete Setting beschreiben, abhängen.

Bernoulli-Verteilung

Das einfachste Beispiel ist die **Bernoulli-Verteilung**.

Eine Bernoulli-verteilte ZV kann nur die Werte 0 und 1 annehmen:

$$P(X = 1) = f(1) = \pi$$

$$P(X = 0) = f(0) = 1 - \pi$$

$$f_X(x) = \pi^x(1-\pi)^{1-x} \cdot I(x \in \{0, 1\})$$

$\pi \in [0, 1]$ ist der einzige Parameter der Bernoulli-Verteilung.

Notation:

$$X \sim \mathcal{B}(\pi)$$

Beispiel: Ergebnis eines Münzwurfs ist $\mathcal{B}(\pi = 0.5)$ -verteilt.

Diskrete Gleichverteilung

Die allgemeine **diskrete Gleichverteilung** hat einen endlichen Träger $T = \{x_1, x_2, \dots, x_k\}$, wobei

$$P(X = x) = f(x) = \frac{1}{k} \cdot I(x \in T)$$

Häufig sind alle natürlichen Zahlen zwischen $a \in \mathbb{N}$ und $b \in \mathbb{N}$ Element des Trägers T . Die Grenzen a und b sind dann die Parameter der **diskreten Gleichverteilung**.

Notation:

$$X \sim \mathcal{U}_D(a, b)$$

Beispiel: Augenzahl beim fairen Würfelwurf ist $\mathcal{U}_D(a = 1, b = 6)$ -verteilt.

In R: `sample()`

Geometrische Verteilung

Ein Zufallsvorgang, bei dem mit Wahrscheinlichkeit π ein Ereignis A eintritt, wird *unabhängig* voneinander so oft wiederholt, bis zum ersten Mal A eintritt.

Sei X die ZV "Anzahl der Versuche bis zum ersten Mal A eintritt".
Dann ist $T = \mathbb{N}^+$ und die Wahrscheinlichkeitsfunktion von X lautet:

$$f(x) = \underbrace{(1 - \pi)^{x-1}}_{(x-1)\text{-mal } \bar{A}} \cdot \underbrace{\pi}_{1\text{-mal } A} \cdot I(x \in T)$$

$\pi \in (0, 1)$ ist der Parameter der geometrischen Verteilung.

Notation:

$$X \sim \mathcal{G}(\pi)$$

Alternative Definition

Sei $Y := \text{"Anzahl der Versuche bevor das erste mal } A \text{ eintritt"}$, d.h. $Y = X - 1$.
Dann ist

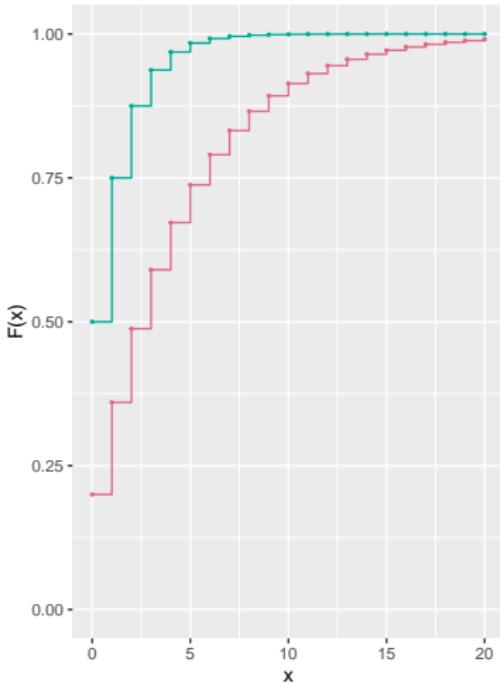
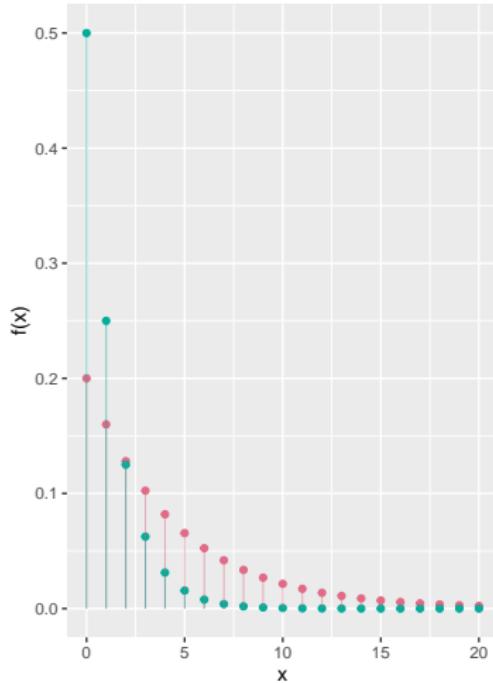
$$T = \{0, 1, 2, \dots\} = \mathbb{N}_0^+$$
$$f(y) = (1 - \pi)^y \pi \cdot I(y \in T)$$

Für diese Definition gibt es folgende Implementationen in R:

- ▶ `dgeom(x, prob=π)` berechnet Wahrscheinlichkeitsfunktion $f_Y(x)$
- ▶ `pgeom(q, prob=π)` wertet Verteilungsfunktion für q aus: $F_Y(q)$
- ▶ `rgeom(n, prob=π)` generiert n Zufallszahlen aus der Verteilung
- ▶ `qgeom(p, prob=π)` berechnet Quantile der geom. Verteilung: $\approx F_Y^{-1}(p)$

Visualisierung $\mathcal{G}(\pi)$

Geometrische Verteilung



$\pi = 0.2$ in rot; $\pi = 0.5$ in grün

Binomialverteilung

Bei einer Folge von unabhängigen Bernoulli-Experimenten $X_i, i = 1, \dots, n$ interessiert man sich häufig nur für die **Anzahl** $X := \sum_{i=1}^n X_i$, wie oft $X_i = 1$ aufgetreten ist.

Diese ZV X heißt **binomialverteilt** mit Parametern $n \in \mathbb{N}, \pi \in [0, 1]$. Sie hat Träger $T = \{0, 1, \dots, n\}$ sowie die Wahrscheinlichkeitsfunktion:

$$P(X = x) = f(x) = \binom{n}{x} \cdot \pi^x (1 - \pi)^{n-x} \cdot I(x \in T)$$

Notation:

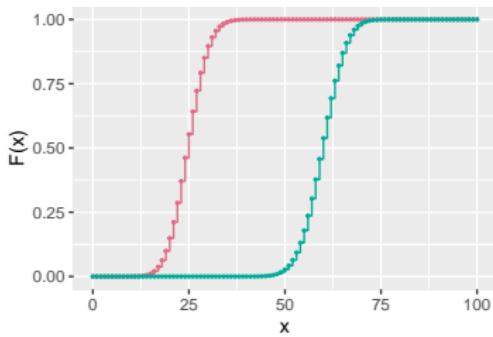
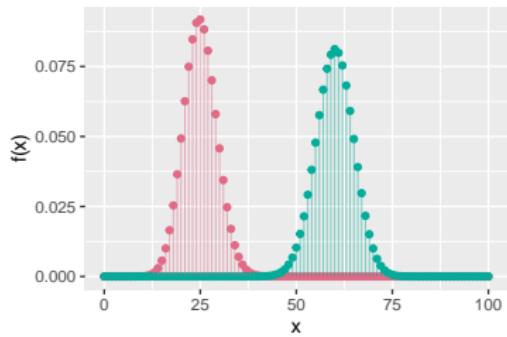
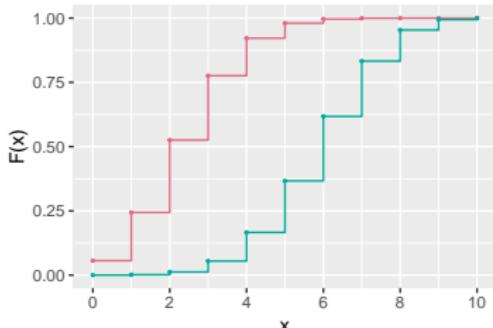
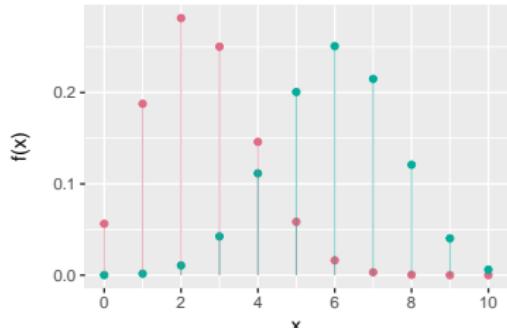
$$X \sim \mathcal{B}(n, \pi)$$

Es gilt $\mathcal{B}(1, \pi) = \mathcal{B}(\pi)$.

in R: `[dpqr]binom(size=n, prob=π, ...)`

Visualisierung $\mathcal{B}(n, \pi)$

Binomial Verteilung



oben: $n = 10$; unten: $n = 100$. $\pi = 0.25$ in rot; $\pi = 0.6$ in grün

Beispiele

Das **Urnenmodell**:

Zufälliges **Ziehen mit Zurücklegen** einer Stichprobe von n Kugeln aus einer Urne mit N Kugeln, darunter M markierte.

Sei X : "Anzahl der markierten Kugeln in der Stichprobe".

Dann gilt

$$X \sim \mathcal{B}(n, M/N).$$

Hypergeometrische Verteilung

Häufig wird jedoch **ohne Zurücklegen** gezogen, d.h.

“Auswahlwahrscheinlichkeiten” ändern sich von Ziehung zu Ziehung
(Beispiel: Kartenspiele).

Die Verteilung von X (Anzahl der markierten Kugeln) nennt man dann
hypergeometrisch. Die *hypergeometrische Verteilung* hat den Träger

$$T = \{\max(0, n - (N - M)), \dots, \min(n, M)\}$$

und die Wahrscheinlichkeitsfunktion

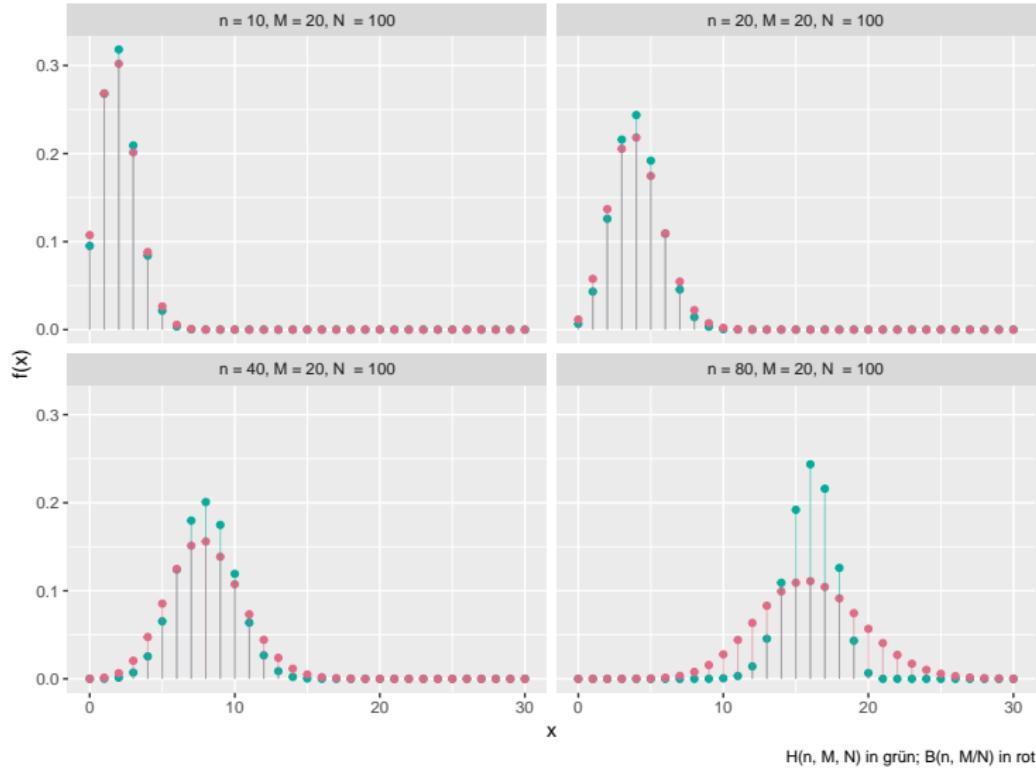
$$f(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \cdot I(x \in T)$$

Notation:

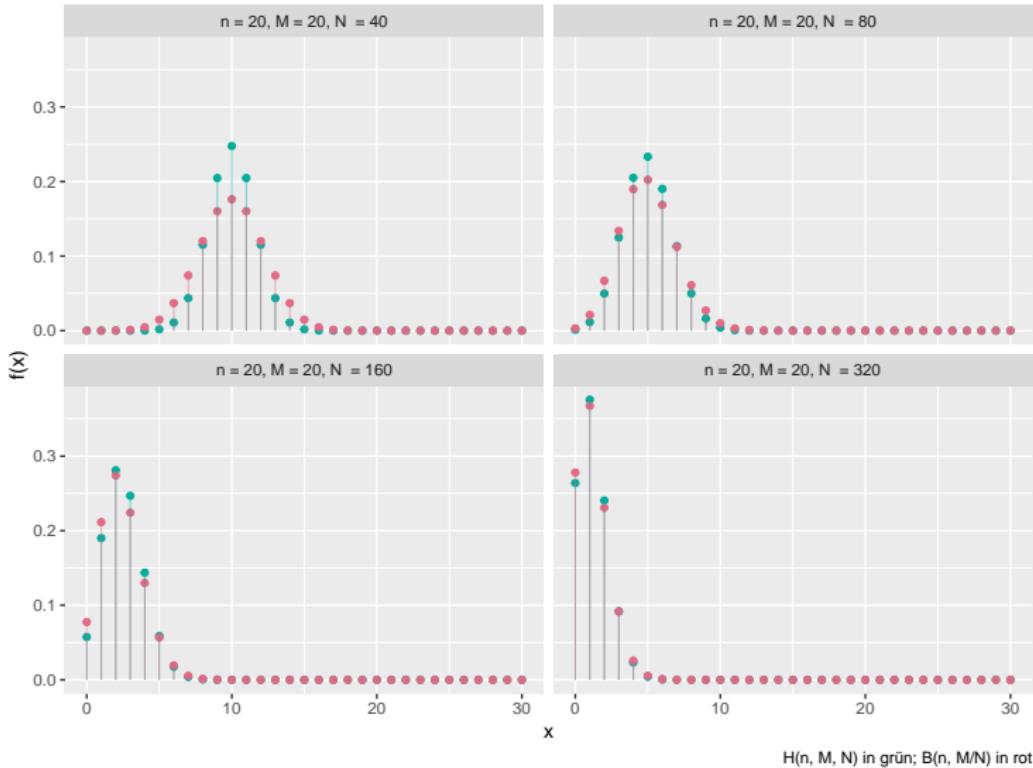
$$X \sim \mathcal{H}(n, N, M)$$

Funktionen in R: [dpqr]hyper(m = M, n = N - M, k = n, ...)

Ziehen mit und ohne Zurücklegen I



Ziehen mit und ohne Zurücklegen II



Approximation der hypergeometrischen Verteilung

Für N "groß" und n "klein" lässt sich die hypergeometrische Verteilung also gut durch die Binomialverteilung approximieren:

$$\mathcal{H}(n, N, M) \quad \stackrel{d}{\approx} \quad \mathcal{B}\left(n, \pi = \frac{M}{N}\right)$$

Poisson-Verteilung

Häufig gibt es zufällige Vorgänge, bei denen es (... zumindest theoretisch ...) keine natürliche obere Grenze für die interessierende Anzahl an Ereignissen in einem gegebenen Zeitintervall gibt, z.B.:

- ▶ Anzahl an Telefonanrufen in einem “Call-Center” pro Stunde
- ▶ Anzahl an Blitzschlägen pro Woche in Oberbayern

Die einfachste Verteilung für solche Phänomene ist die Poisson-Verteilung (nach Siméon Denis Poisson [1781-1840]).

Poisson-Verteilung

Eine Zufallsvariable X mit Träger $T = \mathbb{N}_0^+$ und Wahrscheinlichkeitsfunktion

$$f(x) = \exp(-\lambda) \frac{\lambda^x}{x!} \cdot I(x \in T)$$

folgt einer **Poisson-Verteilung**.

Der Parameter $\lambda \in \mathbb{R}^+$ ist die durchschnittliche **Rate** oder die **Intensität**, mit der die interessierenden Ereignisse in dem zugrundeliegenden Zeitintervall auftreten.

Notation:

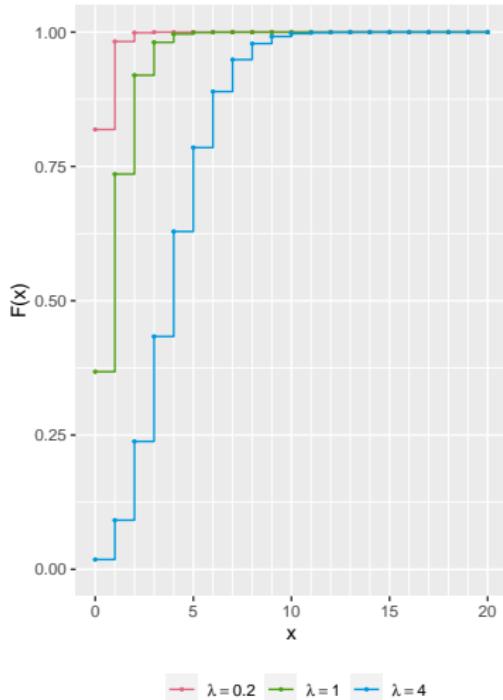
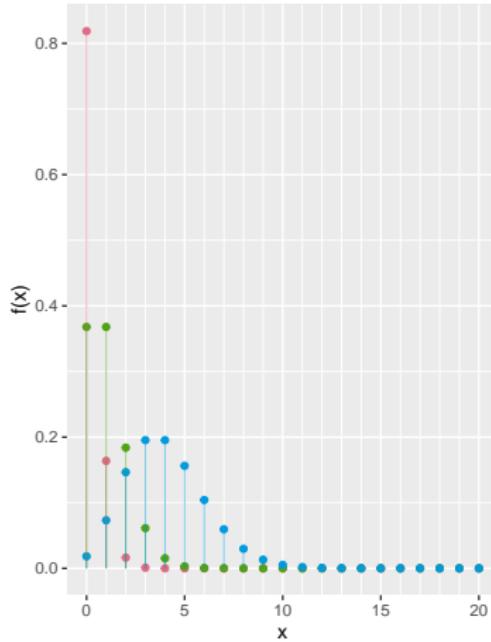
$$X \sim \mathcal{P}(\lambda)$$

Beachte: Wartezeiten zwischen Ereignissen, deren Gesamtzahl $\mathcal{P}(\lambda)$ -verteilt ist, sind $\mathcal{E}(\lambda)$ -verteilt!

Funktionen in R: [dpqr]pois()

Visualisierung $\mathcal{P}(\lambda)$

Poisson Verteilung



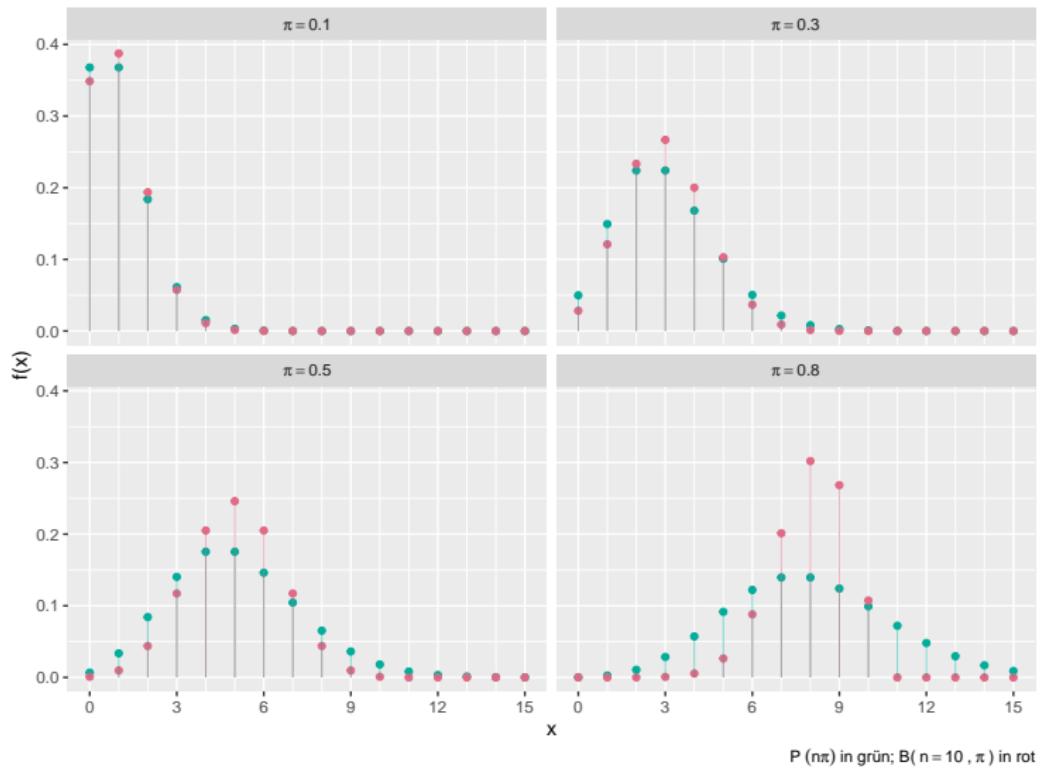
Approximation der Binomialverteilung

Die Binomialverteilung $B(n, \pi)$ kann für “großes n ” und “kleines π ” gut durch die Poisson-Verteilung mit $\lambda = n\pi$ approximiert werden.

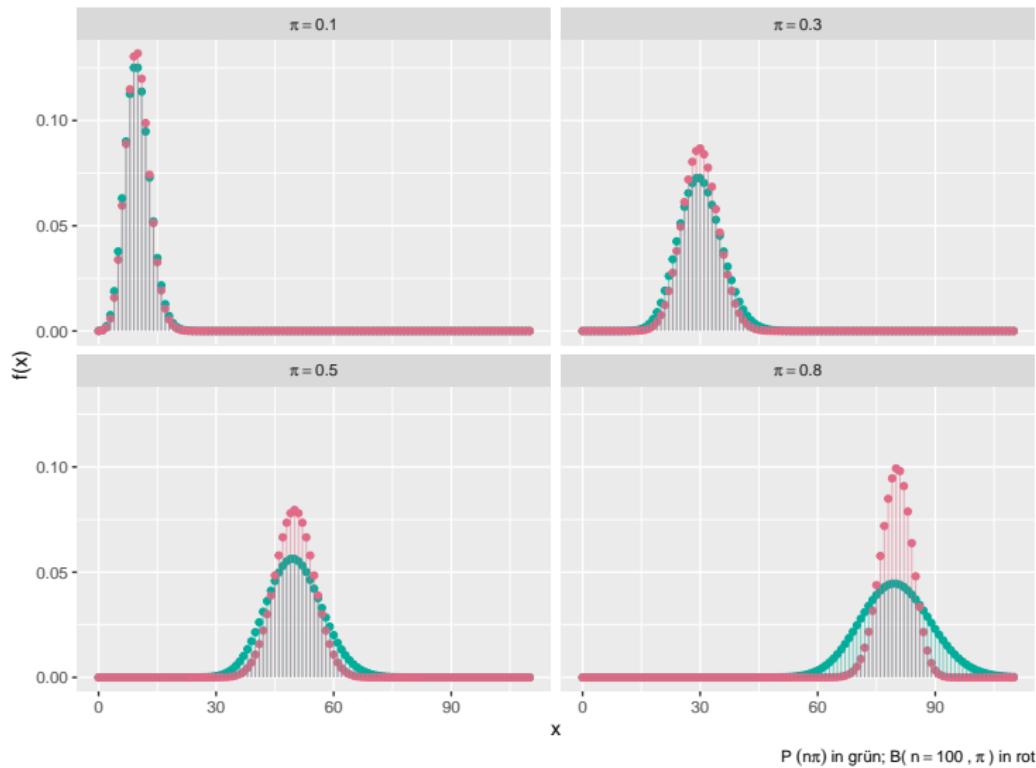
$$\mathcal{B}(n, \pi) \underset{d}{\approx} \mathcal{P}(\lambda = n\pi)$$

Je größer n ist und vor allem je kleiner π , desto besser ist die Approximation.

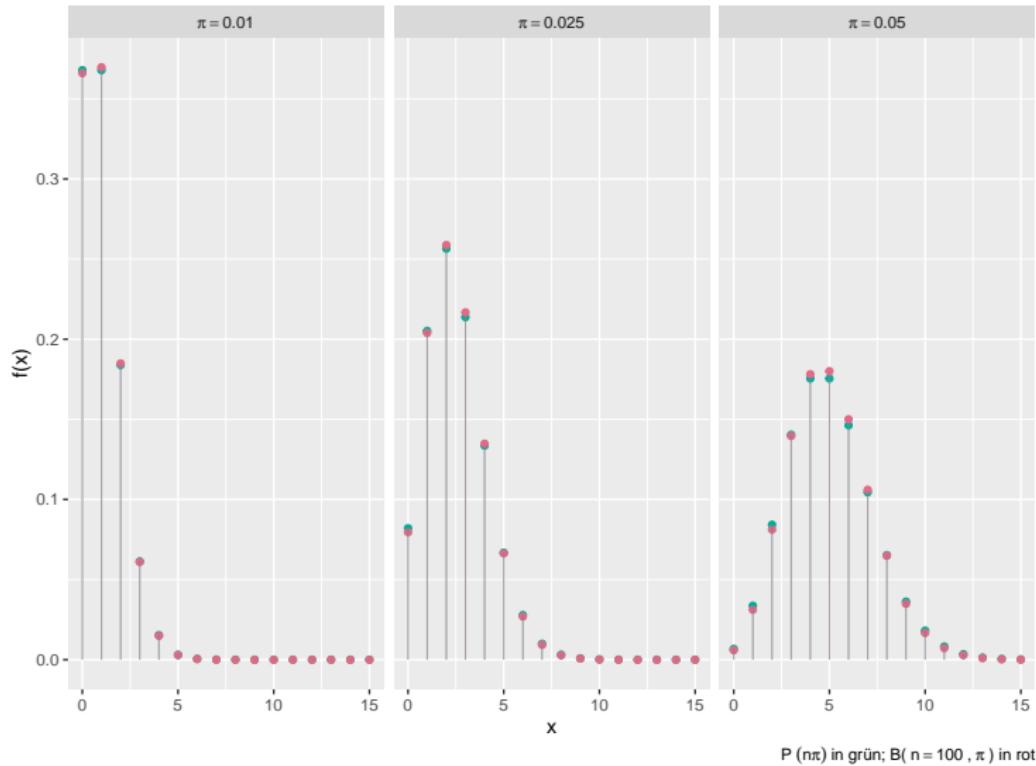
Vergleich Binomial/Poissonverteilung ($n = 10$)



Vergleich Binomial/Poissonverteilung ($n = 100$)



Vergleich Binomial/Poissonverteilung ($n = 100$)



Einige diskrete Verteilungen

Name & Symbol	Parameter	Träger	$E(X)$	$\text{Var}(X)$
Binomialverteilung $\mathcal{B}(n, \pi)$	$n \in \mathbb{N}^+; \pi \in [0, 1]$	$\{0, 1, \dots, n\}$	$n\pi$	$n\pi(1 - \pi)$
Geometrische Verteilung $\mathcal{G}(\pi)$	$\pi \in (0, 1]$	\mathbb{N}^+	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$
Poissonverteilung $\mathcal{P}(\lambda)$	$\lambda \in \mathbb{R}^+$	\mathbb{N}_0^+	λ	λ
Hypergeometrische V. $\mathcal{H}(n, N, M)$	$n, N, M \in \mathbb{N}^+;$ $M \leq N; n \leq N$	$\{\max(0, n-(N-M)), \dots, \min(n, M)\}$	$n \frac{M}{N}$	$n \frac{M}{N} \frac{N-M}{N} \frac{N-n}{N-1}$

Wichtige parametrische Verteilungen

Diskrete parametrische Verteilungen

Stetige parametrische Verteilungen

Dichtetransformationssatz

Wichtige stetige Verteilungen

Im Folgenden werden wir nun wichtige stetige parametrische Verteilungen kennenlernen. Diese hängen wie parametrische diskrete Verteilungen von einem oder mehreren **Parametern** ab.

Zur Charakterisierung geben wir meist die **Dichtefunktion** und den **Träger** an.

Stetige Gleichverteilung

Die **stetige Gleichverteilung** hat

- ▶ Parameter $a \in \mathbb{R}$ und $b \in \mathbb{R}$ ($a < b$)
- ▶ Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

- ▶ Träger $T = [a, b]$.

Notation:

$$X \sim \mathcal{U}[a, b]$$

Funktionen in R: [dprq]unif()

Exponentialverteilung

Eine stetige Zufallsvariable X

- mit *nichtnegativem* Träger $T = \mathbb{R}_0^+$
- und Dichtefunktion

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

- mit Parameter $\lambda \in \mathbb{R}^+$

heißt *exponentialverteilt*. Die Verteilungsfunktion ist

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Notation:

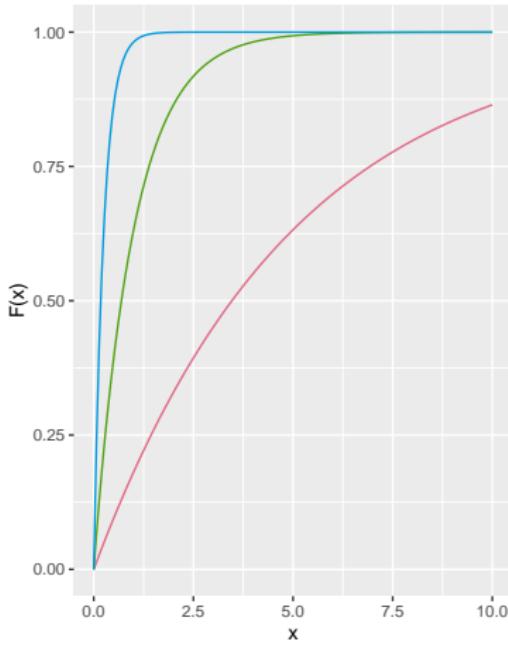
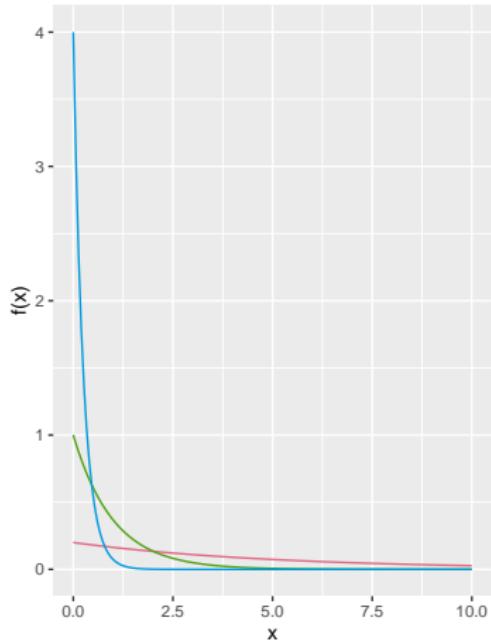
$$X \sim \mathcal{E}(\lambda)$$

in R: [dpqr]exp(rate = λ , ...)

Wenn die Wartezeit auf bzw. zwischen Ereignissen $\mathcal{E}(\lambda)$ -verteilt ist, dann ist die Anzahl der Ereignisse in einem Zeitintervall der Länge 1 $\mathcal{P}(\lambda)$ -verteilt.

Visualisierung Exponentialverteilung

Exponential Verteilung



$\lambda = 0.2$ $\lambda = 1$ $\lambda = 4$

Gammaverteilung

Die **Gammaverteilung** ist eine Verallgemeinerung der Exponentialverteilung mit:

- ▶ Parameter $\alpha \in \mathbb{R}^+$ und $\beta \in \mathbb{R}^+$
- ▶ Dichtefunktion

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & x > 0 \\ 0 & \text{sonst} \end{cases}$$

- ▶ nichtnegativem Träger $T = \mathbb{R}^+$

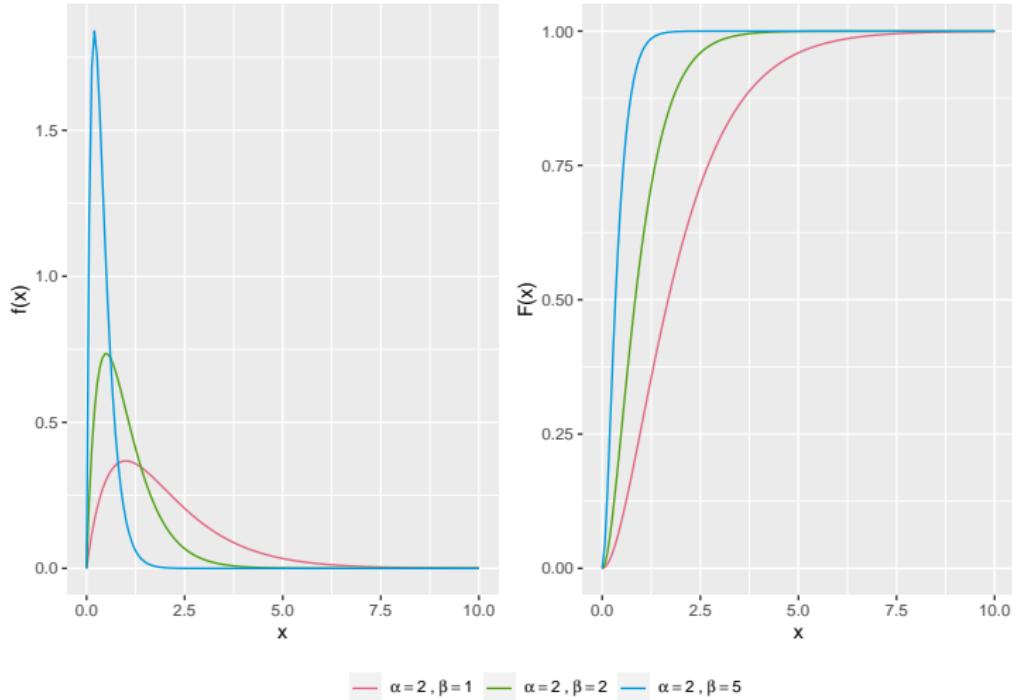
$\Gamma(\alpha)$ bezeichnet die **Gammafunktion** $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$, wobei $\Gamma(x+1) = x!$ für $x = 0, 1, 2, \dots$ gilt.

Notation: $X \sim \mathcal{G}(\alpha, \beta)$

in R: [dpqr]gamma(shape = alpha, rate = beta, ...)

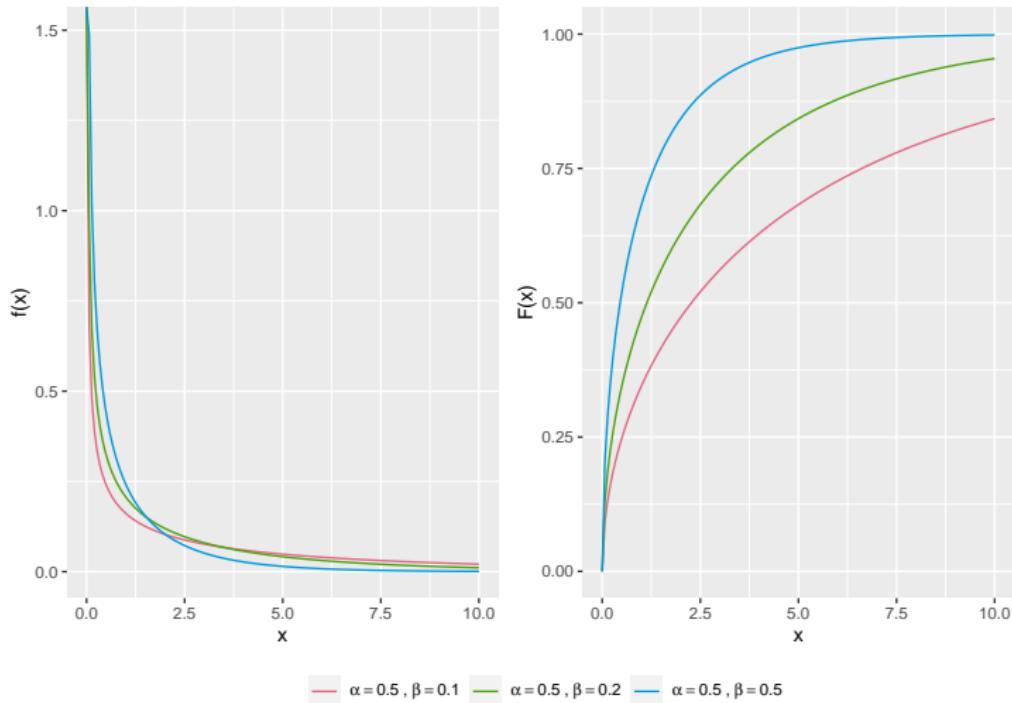
Visualisierung Gammaverteilung

Gamma Verteilung



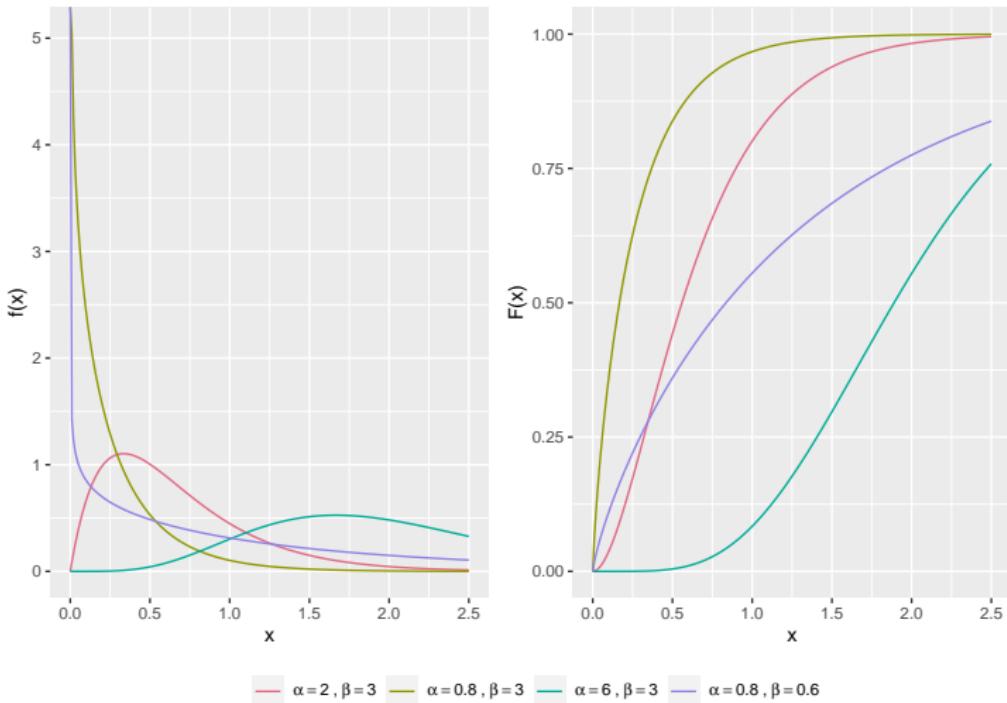
Visualisierung Gammaverteilung

Gamma Verteilung



Visualisierung Gammaverteilung

Gamma Verteilung



Eigenschaften der Gammaverteilung

- ▶ Für $\alpha = 1$ ergibt sich die Exponentialverteilung mit Parameter $\lambda = \beta$:
 $\mathcal{E}(\lambda) \equiv \mathcal{G}(\alpha = 1, \beta = \lambda)$
- ▶ Für $\alpha = \frac{d}{2}$ mit $d \in \mathbb{N}$ und $\beta = \frac{1}{2}$ ergibt sich die sogenannte **χ^2 -Verteilung** mit d Freiheitsgraden: $X \sim \chi^2(d)$:
 $\chi^2(d) \equiv \mathcal{G}(\alpha = \frac{d}{2}, \beta = \frac{1}{2})$
- ▶ in R:
 - ▶ Gammaverteilung: [dpqr]`gamma(..., shape =α, rate =β)`
 - ▶ χ^2 -Verteilung: [dpqr]`chisq(..., df=d)`

Normalverteilung

Eine Zufallsvariable X

- ▶ mit Träger $T = \mathbb{R}$
- ▶ Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right) \quad \text{für } x \in \mathbb{R}$$

- ▶ mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}_+$

heißt **normalverteilt**.

Für $\mu = 0$ und $\sigma^2 = 1$ nennt man die Zufallsvariable **standardnormalverteilt**.

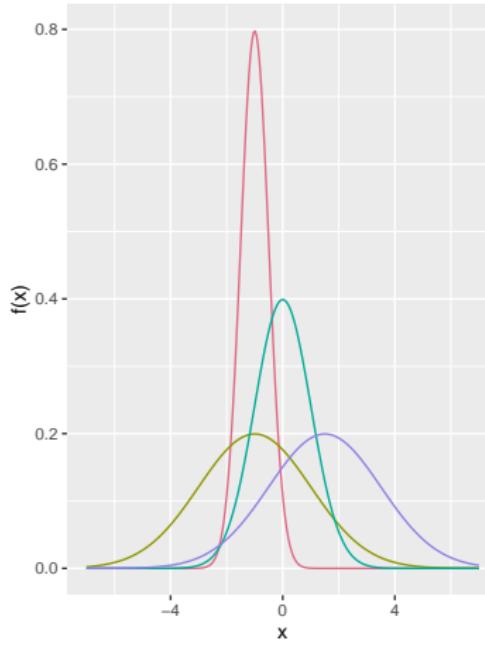
Notation:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

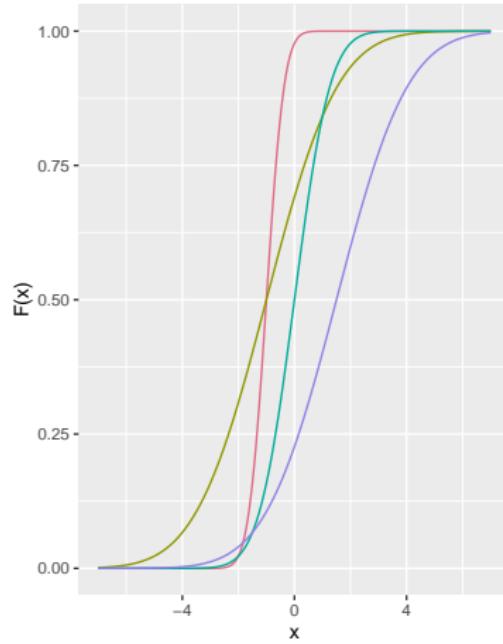
In R: [dpqr]norm(m = μ , sd = $\sqrt{\sigma^2}$, ...) (NB: σ nicht σ^2 !)
("Gaußsche Glockenkurve")

Visualisierung Normalverteilung

Normal–Verteilung



■ $m = -1, \sigma^2 = 0.25$ ■ $m = -1, \sigma^2 = 4$ ■ $m = 0, \sigma^2 = 1$ ■ $m = 1.5, \sigma^2 = 4$



Mehr zur Normalverteilung

- ▶ Verschieben und Skalieren einer normalverteilten Zufallsvariable erzeugt eine neue normalverteilte Zufallsvariable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies (a + bX) \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$$

- ▶ Summen (und Differenzen) normalverteilter Zufallsvariablen sind normalverteilt:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \implies \sum_i X_i \sim \mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$$

- ▶ Das Integral der Normalverteilungsdichte ist nicht analytisch zugänglich, d.h. $F(x) = \int_{-\infty}^x f(u) du$ hat keine geschlossene Form (d.h. man findet keine Stammfunktion und braucht numerische Integration).
Software-Implementation oft über “error function Erf”

Betaverteilung

Eine Zufallsvariable X

- mit Träger $T = (0, 1)$
- Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

- und Parametern $\alpha \in \mathbb{R}_+$ und $\beta \in \mathbb{R}_+$

heißt **betaverteilt**.

Notation:

$$X \sim \mathcal{B}e(\alpha, \beta)$$

Betafunktion $B(\alpha, \beta)$ so definiert, dass Dichtefunktion die

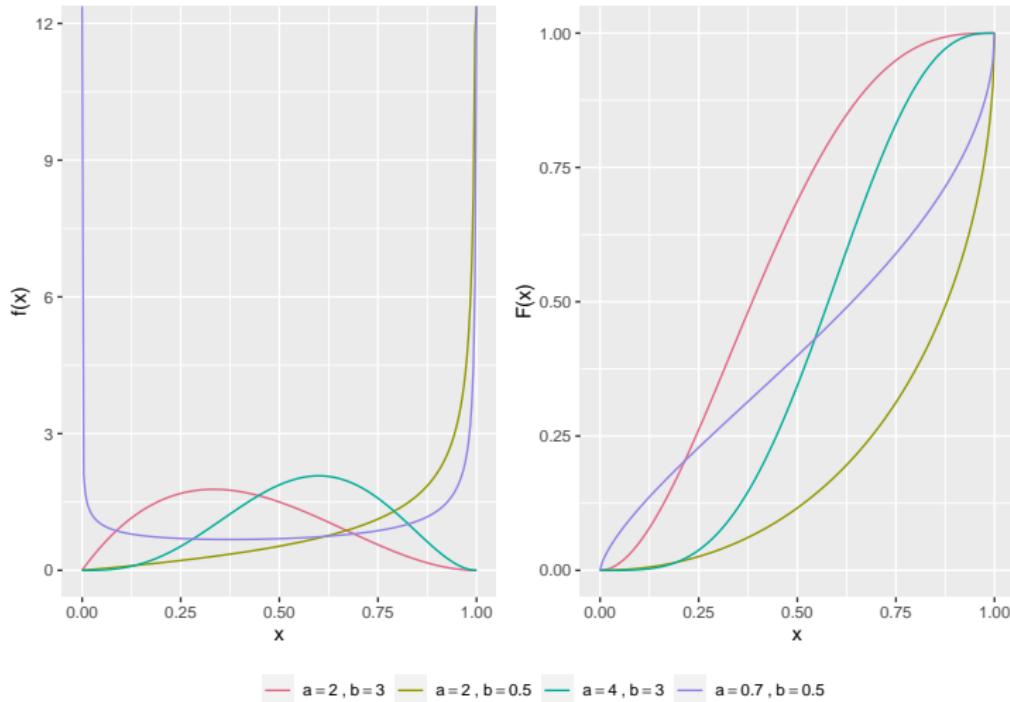
Normierungseigenschaft $\int_0^1 f(x) dx = 1$ besitzt:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx.$$

In R: [dpqr]beta(shape1 = α , shape2 = β , ...)

Visualisierung Betaverteilung

Beta-Verteilung



Cauchy-Verteilung

Eine Zufallsvariable X

- mit Träger $T = \mathbb{R}$
- und Dichte- bzw. Verteilungsfunktion

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$$
$$F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}$$

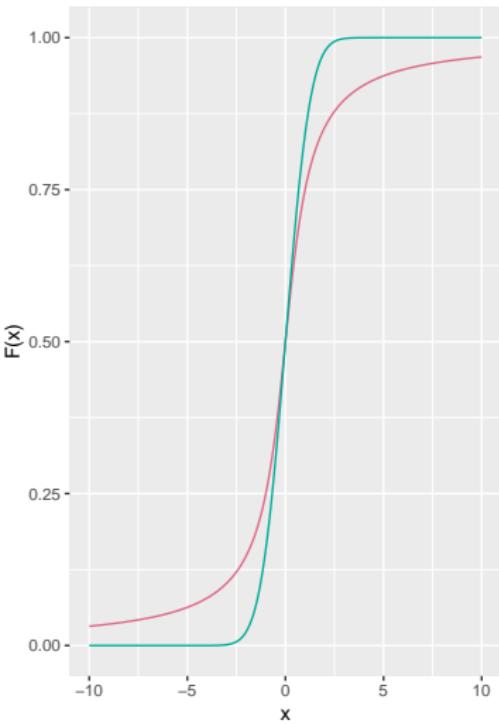
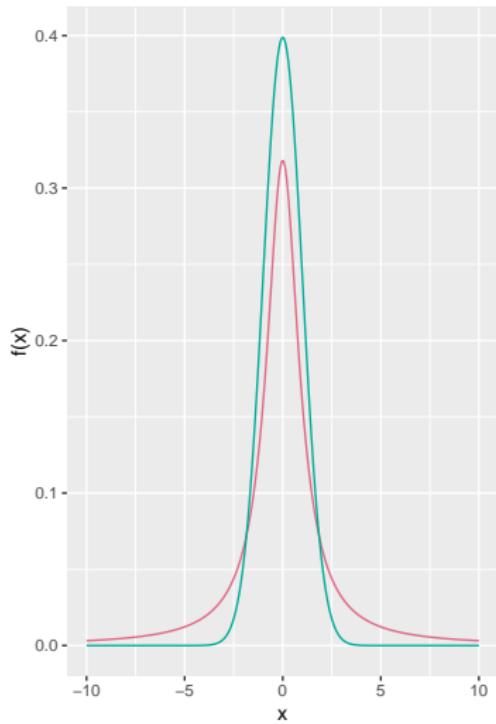
heißt *Cauchy*-verteilt.

Notation:

$$X \sim \mathcal{C}$$

Die Cauchy-Verteilung hat “heavy tails”, d.h. sehr viel Wahrscheinlichkeitsmasse verteilt sich auf extreme Werte.

Vergleich Cauchy-Verteilung / Normalverteilung



$N(0, 1)$ in grün; Cauchy in rot

Einige stetige Verteilungen

Name & Symbol

Parameter

Träger

$E(X)$

$\text{Var}(X)$

Stetige Gleichverteilung

$\mathcal{U}(a, b)$

$a, b \in \mathbb{R}; a \leq b$

$[a, b] \subset \mathbb{R}$

$\frac{a+b}{2}$

$\frac{(b-a)^2}{12}$

Exponentialverteilung

$\mathcal{E}(\lambda)$

$\lambda \in \mathbb{R}^+$

\mathbb{R}^+

$\frac{1}{\lambda}$

$\frac{1}{\lambda^2}$

Gammaverteilung

$\mathcal{G}(\alpha, \beta)$

$\alpha, \beta \in \mathbb{R}^+$

\mathbb{R}^+

$\frac{\alpha}{\beta}$

$\frac{\alpha}{\beta^2}$

Normalverteilung

$\mathcal{N}(\mu, \sigma^2)$

$\mu \in \mathbb{R}; \sigma^2 \in \mathbb{R}^+$

\mathbb{R}

μ

σ^2

Betaverteilung

$\mathcal{B}e(\alpha, \beta)$

$\alpha, \beta \in \mathbb{R}^+$

$(0, 1)$

$\frac{\alpha}{\alpha+\beta}$

$\frac{\alpha \cdot \beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Erwartungswert, Varianz und höhere Momente der Cauchyverteilung sind **nicht definiert**, die entsprechenden Integrale divergieren!

Wichtige parametrische Verteilungen

Diskrete parametrische Verteilungen

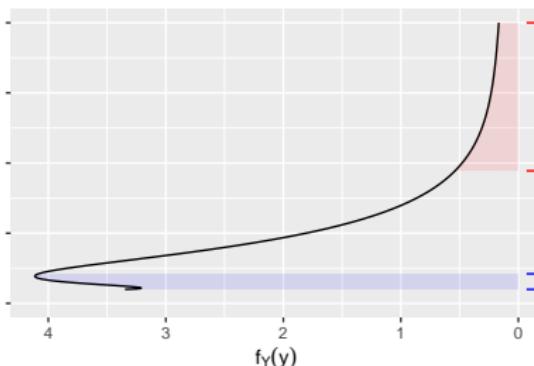
Stetige parametrische Verteilungen

Dichtetransformationssatz

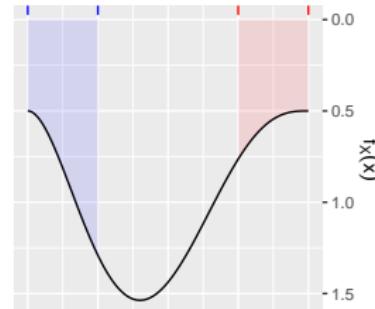
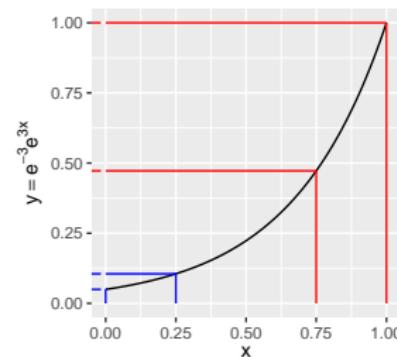
Transformationssatz für Dichten: Idee

Was ist die Dichte von $Y = g(X)$?

Dichte der transformierten ZV $Y = g(X)$



Transformation: $g(x) = e^{-3}e^{3x}$



Der Transformationssatz für Dichten

Transformationssatz für Dichten:

Sei X eine stetige Zufallsvariable mit Dichte $f_X(x)$ und sei Y die transformierte Zufallsvariable $Y = g(X)$ mit Träger

$$T_Y = g(T_X) = \{y : (\exists x \in T_X : g(x) = y)\}.$$

Für eine *streng monotone und differenzierbare* Funktion g gilt:

$$f_Y(y) = f_X(g^{-1}(y)) \left| (g^{-1})'(y) \right|$$

→ berechne Dichten von $Y = \exp(X)$, $Y = \sqrt{X}$, etc....

Formal:

Zufallsvariable Y ist die Verkettung der Abbildungen $X : \Omega \rightarrow T_X$ und $g : T_X \rightarrow T_Y$:

$$Y(\omega) = (g \circ X)(\omega) = g(X(\omega))$$

Transformationssatz für Dichten: Idee

Intuition für $f_Y(y) = f_X(g^{-1}(y)) \left| (g^{-1})'(y) \right|$:

- ▶ Es muss gelten: $P(X \in [a, b]) = P(Y = g(X) \in [g(a), g(b)])$, also
 $\int_a^b f_X(x) dx = \int_{g(a)}^{g(b)} f_Y(y) dy$.
- ▶ Außerdem gilt $(g^{-1})'(y) = \frac{1}{g'(g^{-1}(y))}$, also:
- ▶ Flache Steigung von $g(x)$: $|g'(x)| \ll 1 \iff |(g^{-1})'(y)| \gg 1$
 \implies lange Intervalle in T_X werden zu kurzen Intervallen in T_Y
 \implies Dichte muss dort größer werden damit die selbe Wahrscheinlichkeit in das kleinere Intervall passt.
- ▶ Steile Steigung von $g(x)$: $|g'(x)| \gg 1 \iff |(g^{-1})'(y)| \ll 1$
 \implies kurze Intervalle in T_X werden zu langen Intervallen in T_Y
 \implies Dichte muss dort kleiner werden damit die selbe Wahrscheinlichkeit über ein längeres Intervall verteilt wird.

Beispiel: Lineare Transformation

Wie lautet die Dichte von $Y = g(X) = aX + b$ ($a \neq 0$)?

$$\begin{aligned}g^{-1}(y) &= \frac{y-b}{a} \\ \left| (g^{-1})'(y) \right| &= \left| \frac{1}{a} \right| \\\implies f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)\end{aligned}$$

Beispiel: Das Quadrat einer Standardnormalverteilung

Wie lautet die Dichte von $Y = X^2$, falls $X \sim \mathcal{N}(0, 1)$?

Betrachte zunächst $Z = |X|$, dann hat Z offensichtlich die Dichte

$$f(z) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \quad \text{für } z \geq 0 \text{ und } 0 \text{ sonst}$$

Nun ist $X^2 = Y = Z^2 = g(Z)$ und g monoton wachsend auf dem Wertebereich \mathbb{R}^+ von Z . Es ergibt sich wegen $y = z^2 \iff z = \sqrt{y}$:

$$\begin{aligned} (g^{-1})'(y) &= \frac{1}{2}y^{-\frac{1}{2}} \\ \implies f(y) &= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}y\right) \quad \text{für } y \geq 0 \text{ und } 0 \text{ sonst} \end{aligned}$$

Das ist die Dichte einer $\mathcal{G}(0.5, 0.5)$,
also einer χ^2 -Verteilung mit Freiheitsgrad 1:

$$X^2 = Y \sim \chi^2(1)$$

Beispiel: Erzeugung exponentialverteilter Zufallsvariablen

Betrachte $X \sim \mathcal{U}[0, 1]$ und $Y = -\log(X)$, also $g(x) = -\log(x)$.
Die Umkehrfunktion und deren Ableitung lauten:

$$g^{-1}(y) = \exp(-y) \quad (g^{-1})'(y) = -\exp(-y)$$

Durch Anwendung des Transformationssatzes für Dichten erhält man

$$f_Y(y) = 1 \cdot |-\exp(-y)| = \exp(-y)$$

Es gilt: $Y \sim \mathcal{E}(\lambda = 1)$!

Dies ist also eine einfache Art, exponentialverteilte Zufallsvariablen zu erzeugen, und allgemeiner liefert $Y = -\frac{1}{\lambda} \log(X)$ Zufallszahlen $Y \sim \mathcal{E}(\lambda)$.

Anwendung: Inversions-Methode

Allgemeiner kann man die **Inversions-Methode** zur Erzeugung von Realisationen x_i aus einer beliebigen stetigen Verteilung mit Verteilungsfunktion $F_X(x)$ verwenden:

1. Erzeuge Realisationen u_1, \dots, u_n mit $U_i \stackrel{\text{iid}}{\sim} \mathcal{U}[0, 1]$.
2. Berechne $x_i = F_X^{-1}(u_i)$, $i = 1, \dots, n$

Die x_i sind dann Zufallszahlen aus der gewünschten Verteilung mit CDF $F(x)$.

Beweis:

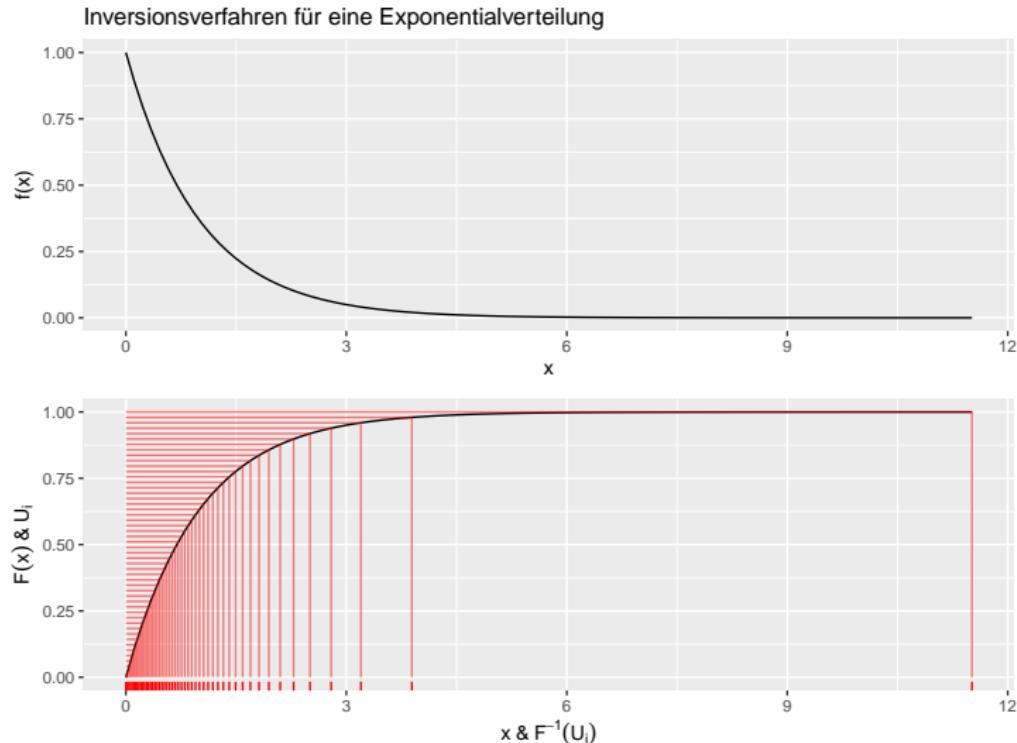
$g() := F_X^{-1}()$, also ist $g^{-1}() = F_X()$.

Ausserdem ist $f_U(x) = 1 \forall x \in [0, 1]$.

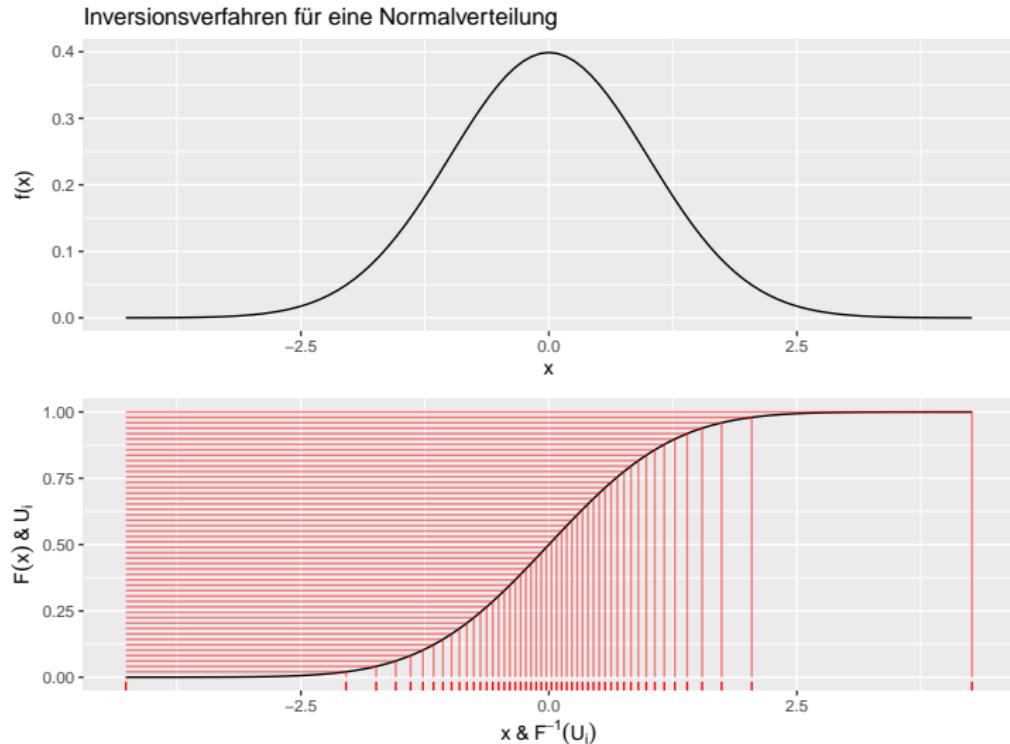
Damit ist die Dichte der X_i :

$$f(x) = \underbrace{f_U(F_X(x))}_{=1} \cdot \underbrace{F'_X(x)}_{f_X(x)} = f_X(x)$$

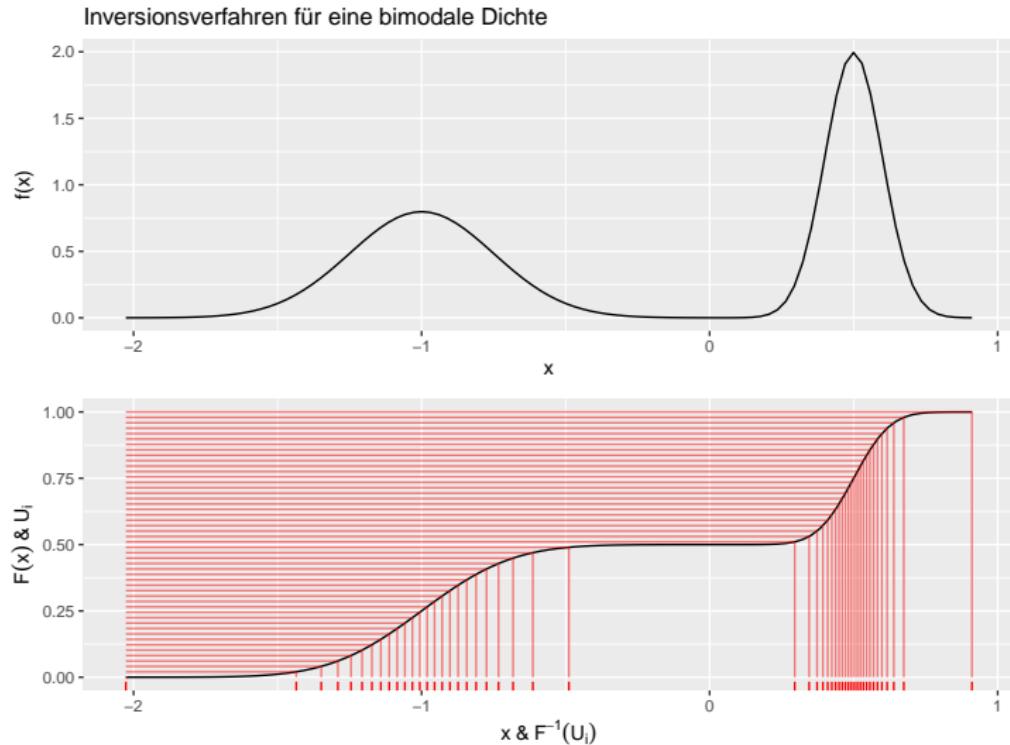
Intuition: Inversions-Methode



Intuition: Inversions-Methode



Intuition: Inversions-Methode



Beispiel: Zufallszahlen aus der Cauchy-Verteilung

Dichte- und Verteilungsfunktion der Cauchy-Verteilung sind:

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$$
$$F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}$$

Die inverse Verteilungsfunktion ist somit:

$$F^{-1}(y) = \tan\left[\pi\left(y - \frac{1}{2}\right)\right]$$

Zufallszahlen aus der Cauchy-Verteilung lassen sich also leicht erzeugen, indem man U_1, \dots, U_n aus $\sim \mathcal{U}[0, 1]$ erzeugt und $\tan\left[\pi\left(U_i - \frac{1}{2}\right)\right]$ berechnet.

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Gemeinsame Verteilungsfunktion

Def.: Gemeinsame Verteilungsfunktion

Die *gemeinsame Verteilungsfunktion* zweier Zufallsvariablen X und Y auf dem Wahrscheinlichkeitsraum (Ω, P) ist die Funktion

$$\begin{aligned}F_{X,Y}(x,y) &:= P(X \leq x \wedge Y \leq y) \\&= P(\{\omega \in \Omega : X(\omega) \leq x \wedge Y(\omega) \leq y\})\end{aligned}$$

Beachte:

- $F_{X,Y}(x,y)$ ist eine Funktion von $\mathbb{R} \times \mathbb{R}$ nach $[0, 1]$
- Definition gilt für stetige und diskrete ZV

Gemeinsame Wahrscheinlichkeitsfunktion

Def.: Gemeinsame Wahrscheinlichkeitsfunktion

Für diskrete ZVn X und Y mit Trägern T_X, T_Y ist die *gemeinsame Wahrscheinlichkeitsfunktion* von X und Y

$$\begin{aligned} f_{X,Y}(x,y) &= P(X = x \wedge Y = y) \\ &= P(X = x, Y = y) \quad \forall x \in T_X, y \in T_Y \end{aligned}$$

Für solche *diskreten* ZVn X und Y gilt dann auch:

$$F_{X,Y}(x,y) = \sum_{\{u: u \leq x\}} \sum_{\{v: v \leq y\}} f_{X,Y}(u,v)$$

Gemeinsame Dichtefunktion

Def.: Gemeinsame Dichtefunktion

Die *gemeinsame Dichtefunktion* $f_{X,Y}(x,y)$ von stetigen X und Y wird über ihre *gemeinsame Verteilungsfunktion* $F_{X,Y}(x,y)$ definiert:

$$F_{X,Y}(x,y) = \int_{v=-\infty}^y \int_{u=-\infty}^x f_{X,Y}(u,v) du dv \quad \forall x, y \in \mathbb{R}$$

Eigenschaften von $f(x, y)$ und $F(x, y)$

- Im Folgenden meist $f(x, y)$ statt $f_{X,Y}(x, y)$, $F(x, y)$ statt $F_{X,Y}(x, y)$
- Für die gemeinsame Dichte stetiger ZV gilt überall dort wo $f(x, y)$ stetig ist:

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$$

- Genau wie univariate Dichtefunktionen ist die gemeinsame Dichtefunktion $f(x, y)$ normiert:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$$

- ... und ihr Integral ergibt Wahrscheinlichkeiten für beliebige Teilmengen $A \subseteq (\mathbb{R} \times \mathbb{R})$:

$$P((X, Y) \in A) = \int_A f(x, y) d(x, y)$$

Transformationsregel für $E(g(X, Y))$

Seien X und Y zwei ZVn mit gemeinsamer Wahrscheinlichkeits-/Dichtefunktion $f_{X,Y}(x, y)$.

Sei $g(x, y)$ eine reellwertige Funktion.

Dann gilt für $Z = g(X, Y)$

$$E(Z) = E(g(X, Y)) = \begin{cases} \sum_x \sum_y g(x, y) f_{X,Y}(x, y) & X, Y \text{ diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & X, Y \text{ stetig} \end{cases}$$

Beispiel: für $Z = X \cdot Y$ gilt daher

$$E(X \cdot Y) = \sum_x \sum_y x \cdot y \cdot f_{X,Y}(x, y)$$

bzw.

$$E(X \cdot Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y) dx dy.$$

Randverteilungen

Die Dichten der *Randverteilungen* von X und Y sind gegeben durch:

$$f_X(x) = \begin{cases} \sum_{y \in T_Y} f_{X,Y}(x,y) & X, Y \text{ diskret} \\ \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy & X, Y \text{ stetig} \end{cases}$$

bzw.

$$f_Y(y) = \begin{cases} \sum_{x \in T_X} f_{X,Y}(x,y) & X, Y \text{ diskret} \\ \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx & X, Y \text{ stetig} \end{cases}$$

Die **gemeinsame Verteilung** von X und Y enthält i. A. mehr Information als in den **Randverteilungen** von X und Y steckt
~~ stochastische (Un-)Abhängigkeit

Allgemeine Zufallsvektoren

Allgemeiner kann man mehr als 2 Zufallsvariablen X_1, \dots, X_n zu einem **Zufallsvektor $\mathbf{X} = (X_1, \dots, X_n)$** der Dimension n zusammenfassen.

Dieser hat dann Wahrscheinlichkeits-/Dichtefunktion

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1, \dots, X_n}(x_1, \dots, x_n)$$

und Verteilungsfunktion

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1 \wedge \dots \wedge X_n \leq x_n).$$

Die Randverteilungen der einzelnen Komponenten sind entsprechend

$$f_{X_i}(x_i) = \begin{cases} \sum_{\mathbf{x}: x_i = x_i} f_{\mathbf{X}}(\mathbf{x}) & \text{diskretes } \mathbf{X} \\ \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f_{\mathbf{X}}(u_1, \dots, u_{i-1}, x_i, u_{i+1}, \dots, u_n) du_1 \dots du_{i-1} du_{i+1} \dots du_n & \text{stetiges } \mathbf{X} \end{cases} .$$

Beispiel: Betrug beim Münzwurf

Ein Lehrer bittet seine Schüler, eine (faire) Münze zweimal zu werfen, und das Ergebnis ("Kopf" = 0, "Zahl" = 1) für jeden Wurf zu notieren. Sei X das Ergebnis des ersten Wurfes und Y das Ergebnis des zweiten Wurfes.

- ▶ Ein gewissenhafter Schüler folgt genau den Anweisungen des Lehrers und notiert das Ergebnis X_G und Y_G . Ein fauler Schüler wirft nur eine Münze und notiert das erzielte Ergebnis zweimal: X_F und Y_F .
- ▶ Berechne die gemeinsame Wahrscheinlichkeitsfunktion von (X_G, Y_G) und von (X_F, Y_F) .

Beispiel: Trinomialverteilung

Ein Experiment, bei dem ein von drei möglichen Ereignissen mit Wahrscheinlichkeit π_1 , π_2 und π_3 ($\pi_1 + \pi_2 + \pi_3 = 1$) auftritt, wird unabhängig voneinander n -mal wiederholt. Sei \mathbf{X} ein drei-dimensionaler Zufallsvektor, dessen i -te Komponente angibt, wie oft das i -te Ereignis eingetreten ist.

Beispiel:

In einer Population mit Häufigkeiten π_1 , π_2 und π_3 der Genotypen aa , ab und bb wird eine Stichprobe vom Umfang n gezogen. Die Anzahlen X_1 , X_2 und X_3 der drei Genotypen ist dann **trinomialverteilt**.

Beispiel: Trinomialverteilung

Ein drei-dimensionaler diskreter Zufallsvektor \mathbf{X} heißt **trinomialverteilt**, falls er Träger

$$T_X = \{\mathbf{x} = (x_1, x_2, x_3) : x_i \in \{0, 1, \dots, n\} \wedge x_1 + x_2 + x_3 = n\}$$

und Wahrscheinlichkeitsfunktion

$$f_{\mathbf{x}}(\mathbf{x}) = f_{x_1, x_2, x_3}(x_1, x_2, x_3) = \frac{n!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3}$$

besitzt.

Multinomialverteilung

Man schreibt kurz: $\mathbf{X} \sim \mathcal{M}_3(n, \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3))$

Hierbei steht \mathcal{M}_3 für die **Multinomialverteilung** der Dimension 3.

Allgemein:

Ein $\mathcal{M}_d(n, \boldsymbol{\pi})$ -verteilter Zufallsvektor \mathbf{X} hat

- Träger $T_{\mathbf{X}} = \{(x_1, x_2, \dots, x_d) : x_j \in \{0, 1, \dots, n\} \forall j \wedge \sum_{j=1}^d x_j = n\}$
- und Parameter $\boldsymbol{\pi} = (\pi_1, \dots, \pi_d)$ mit $\pi_j \in [0, 1] \forall j$ und $\sum_{j=1}^d \pi_j = 1$.

Die Multinomialverteilung ist eine Verallgemeinerung der Binomialverteilung auf unabhängige Wiederholungen von identischen Zufallsexperimenten mit mehr als zwei möglichen Ergebnissen.

Für ihre Randverteilungen gilt: $X_i \sim \mathcal{B}(n, \pi_i)$.

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Unabhängige Zufallsvariablen

Def: Unabhängige Zufallsvariablen

Zufallsvariablen X, Y sind genau dann *stochastisch unabhängig*, wenn $\forall x \in T_X, y \in T_Y$ (X, Y diskret) bzw. $\forall x, y \in \mathbb{R}$ (X, Y stetig) gilt:

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)$$

und damit

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

(bzw. $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ im diskreten Fall.)

Vektoren von unabhängigen Zufallsvariablen

Allgemein: ZVn X_1, X_2, \dots, X_n heißen **unabhängig**, falls

$$f_{X_1, \dots, X_n}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

d.h. für diskrete X_i :

$$P\left(\bigcap_i (X_i = x_i)\right) = \prod_{i=1}^n P(X_i = x_i)$$

für alle x_1, x_2, \dots, x_n aus den entsprechenden Trägern gilt.

- Zufallsvariablen sind stochastisch unabhängig, wenn ihre gemeinsame Verteilung dem Produkt ihrer Randverteilungen entspricht.
Ihre gemeinsame Verteilung enthält dann also keine zusätzliche Information gegenüber den Randverteilungen.
- Analog zu stoch. Unabhängigkeit von Ereignissen A, B :

$$A \perp B \iff P(A \cap B) = P(A)P(B)$$

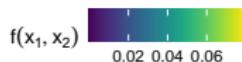
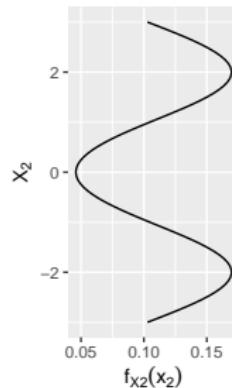
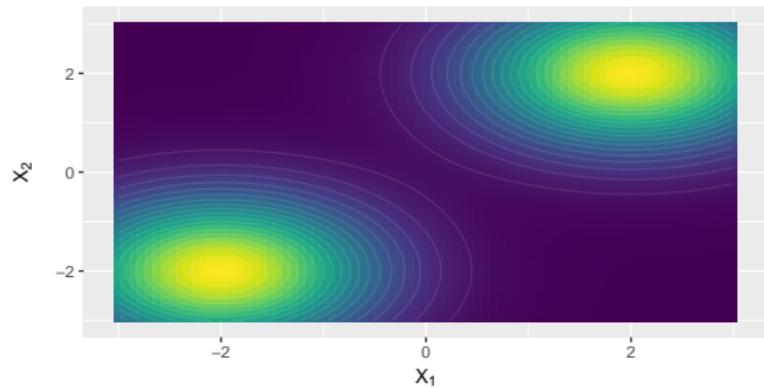
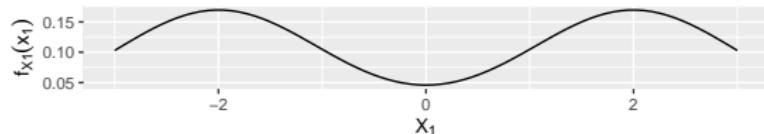
$$X \perp Y \iff P((X(\omega) \leq x) \cap (Y(\omega) \leq y)) = P(X(\omega) \leq x)P(Y(\omega) \leq y) \quad \forall x, y$$

$$\iff F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \forall x, y$$

$$\iff f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \forall x, y$$

Beispiel: Randverteilung & Gemeinsame Verteilung

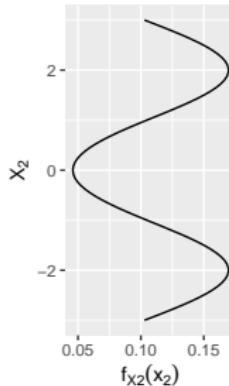
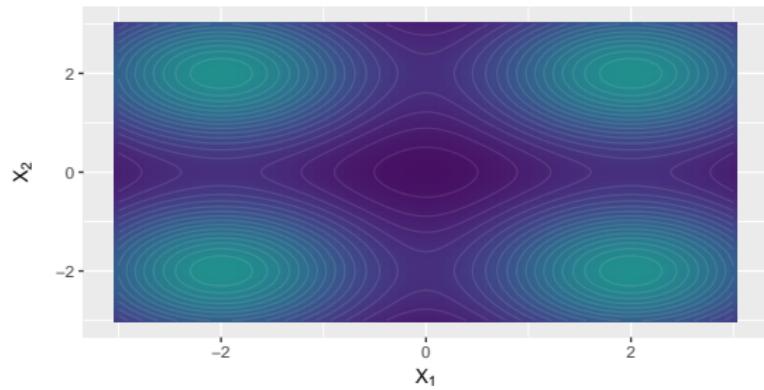
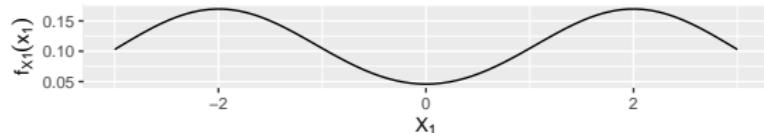
Abhängige Zufallsvariablen



$$f(x_1, x_2) \neq f_{X_1}(x_1)f_{X_2}(x_2)$$

Beispiel: Randverteilung & Gemeinsame Verteilung

Unabhängige Zufallsvariablen



$$f(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

Beispiel: Bernoulli-Kette

Sind X_1, X_2, \dots, X_n jeweils Bernoulli-verteilt mit Parameter π und unabhängig, so heißt $\mathbf{X} = (X_1, X_2, \dots, X_n)$ **Bernoulli-Kette**.

Beispiel:

$n = 3, \pi = \frac{1}{6}$; wegen der Unabhängigkeit gilt dann z.B.

$$\begin{aligned} P(\mathbf{X} = (1, 0, 0)) &= P(X_1 = 1, X_2 = 0, X_3 = 0) \\ &= P(X_1 = 1) \cdot P(X_2 = 0) \cdot P(X_3 = 0) \\ &= \frac{1}{6} \left(\frac{5}{6} \right)^2 = \frac{25}{216} \end{aligned}$$

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Faltungen

Sind X und Y diskrete ZV mit Wahrscheinlichkeitsfunktionen $f_X(x)$ und $f_Y(y)$, so gilt für die Summe $Z = X + Y$:

$$\begin{aligned} P(X + Y = z) &= \sum_x P(X = x, X + Y = z) \\ &= \sum_x P(X = x, Y = z - x) \end{aligned}$$

falls $X \perp Y$ gilt außerdem:

$$\begin{aligned} &= \sum_x P(X = x) \cdot P(Y = z - x) \\ &= \sum_x f_X(x) \cdot f_Y(z - x) \end{aligned}$$

Faltungen

Def: Faltung

Die *Faltung* der stochastisch unabhängigen, diskreten Zufallsvariablen X und Y bezeichnet die Wahrscheinlichkeitsverteilung ihrer Summe $Z = X + Y$:

$$\begin{aligned} P(X + Y = z) = f_Z(z) &= \sum_x f_X(x) \cdot f_Y(z - x) \\ &= \sum_y f_X(z - y) \cdot f_Y(y) \end{aligned}$$

Für *stetige* unabhängige ZVn X, Y mit $Z = X + Y$ gilt analog die **Faltung**

$$f_Z(z) = \int f_X(x)f_Y(z - x)dx = \int f_X(z - y)f_Y(y)dy$$

Bsp: Faltung zweier geometrischen ZV

Seien $X \sim \mathcal{G}(\pi)$ und $Y \sim \mathcal{G}(\pi)$ unabhängig.

Berechne Träger und Wahrscheinlichkeitsfunktion der Summe $Z = X + Y$.

Wie kann man Z interpretieren?

Negative Binomialverteilung

Die Verteilung der Faltung $X = X_1 + \dots + X_n$ mit $X_i \stackrel{\text{iid}}{\sim} \mathcal{G}(\pi)$; $i = 1, \dots, n$ ist eine **Negative Binomialverteilung** auf $T_X = \{n, n+1, n+2, \dots\}$ mit Parametern $n \in \mathbb{N}^+$ und $\pi \in (0, 1)$ und Wahrscheinlichkeitsfunktion

$$f(x) = \binom{x-1}{n-1} \pi^n (1-\pi)^{x-n} I(x \geq n)$$

Wir schreiben: $X \sim \mathcal{NB}(n, \pi)$

Funktionen in R: [dpqr]nbinom(size, prob)

Beachte: Abweichende Definition in R – Träger ist \mathbb{N}_0^+ !

Bsp: Faltung zweier Poisson-ZV

Sind $X \sim \mathcal{P}(\lambda_1)$ und $Y \sim \mathcal{P}(\lambda_2)$ unabhängig, so ist die *Faltung* von X und Y wieder Poisson-verteilt mit Parameter $\lambda_1 + \lambda_2$:

$$(X + Y) \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

Beweis: $P(X + Y = z) = \sum_{x=0}^z f_X(x) \cdot f_Y(z - x)$

$$\begin{aligned} &= \sum_{x=0}^z \frac{\lambda_1^x}{x!} \exp(-\lambda_1) \frac{\lambda_2^{z-x}}{(z-x)!} \exp(-\lambda_2) \\ &= \sum_{x=0}^z \frac{\lambda_1^x \lambda_2^{z-x}}{x!(z-x)!} \exp(-(\lambda_1 + \lambda_2)) \\ &= \left(\sum_{x=0}^z \frac{z!}{x!(z-x)!} \lambda_1^x \lambda_2^{z-x} \right) \frac{\exp(-(\lambda_1 + \lambda_2))}{z!} \\ &= \underbrace{(\lambda_1 + \lambda_2)^z}_{\sum_{k=0}^n \binom{n}{k} a^{n-k} b^k = (a+b)^n} \frac{\exp(-(\lambda_1 + \lambda_2))}{z!}. \end{aligned}$$

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Bedingte Verteilungen von diskreten ZVn

Def: Bedingte Verteilungs- & Dichtefunktion diskreter ZV

Die bedingte Verteilungsfunktion $F_{X|Y}(x|y)$ und die bedingte Wahrscheinlichkeits- oder Dichtefunktion $f_{X|Y}(x|y)$ einer diskreten ZV X gegeben $Y = y$, sind definiert für alle y mit $P(Y = y) > 0$:

$$F_{X|Y}(x|y) = P(X \leq x | Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

$$f_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$= \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Folgerungen

Es gilt immer:

$$\begin{aligned}f_{X,Y}(x,y) &= f_{X|Y}(x|y) \cdot f_Y(y) \\&= f_{Y|X}(y|x) \cdot f_X(x)\end{aligned}$$

Daraus folgt:

X und Y sind genau dann unabhängig wenn

$$\begin{aligned}f_{X|Y}(x|y) &= f_X(x) \\ \text{oder} \quad f_{Y|X}(y|x) &= f_Y(y)\end{aligned}$$

für alle x und y gilt.

Beispiel

Betrachte zwei unabhängige ZV $X \sim \mathcal{P}(\lambda_1)$ und $Y \sim \mathcal{P}(\lambda_2)$. Sei $Z = X + Y$. Dann gilt

$$X|Z=z \sim \mathcal{B}\left(z, \pi = \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$$

Beweis:

$$\begin{aligned} P(X=x|Z=z) &= \frac{P(X=x \cap Y=z-x)}{P(Z=z)} \\ &= \frac{\frac{\lambda_1^x}{x!} \exp(-\lambda_1) \frac{\lambda_2^{z-x}}{(z-x)!} \exp(-\lambda_2)}{\frac{(\lambda_1+\lambda_2)^z}{z!} \exp(-(\lambda_1+\lambda_2))} \\ &= \frac{z!}{x!(z-x)!} \frac{\lambda_1^x \lambda_2^{z-x}}{(\lambda_1+\lambda_2)^z} = \binom{z}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \underbrace{\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{z-x}}_{=\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)} \end{aligned}$$

Bedingte Verteilungen von stetigen ZVn

Betrachte die Zufallsvariablen X und Y mit gemeinsamer Dichte $f_{X,Y}(x,y)$.
Wir interessieren uns für die bedingte Verteilung von X gegeben $Y = y$.

“Problem”:

Es gilt $P(Y = y) = 0$ für alle y . Daher ist

$$P(X \leq x | Y = y) = \frac{P(X \leq x \quad \wedge \quad Y = y)}{P(Y = y)}$$

nicht definiert.

Bedingte Verteilungen von stetige ZVn

“Lösung:”

Betrachte W.keit bedingt auf sehr kleine Intervalle mit Länge δ – es gilt
näherungsweise: $P(y - \frac{\delta}{2} \leq Y \leq y + \frac{\delta}{2}) = \int_{y-\delta/2}^{y+\delta/2} f_Y(u)du \approx f_Y(y) \delta$

$$\begin{aligned}\implies P(X \leq x | Y = y) &\approx P(X \leq x | y - \frac{\delta}{2} \leq Y \leq y + \frac{\delta}{2}) = \frac{P\left((X \leq x) \wedge \left(y - \frac{\delta}{2} \leq Y \leq y + \frac{\delta}{2}\right)\right)}{P(y - \frac{\delta}{2} \leq Y \leq y + \frac{\delta}{2})} \\ &= \frac{\int_{-\infty}^x \left(\int_{y-\delta/2}^{y+\delta/2} f_{X,Y}(u,v)dv \right) du}{\int_{y-\delta/2}^{y+\delta/2} f_Y(u)du} \\ &\approx \frac{\int_{-\infty}^x f_{X,Y}(u,y) \delta du}{f_Y(y) \delta} \\ &= \int_{-\infty}^x \underbrace{\frac{f_{X,Y}(u,y)}{f_Y(y)}}_{\text{Dichte von } X|Y=y} du\end{aligned}$$

Bedingte Verteilungen von stetigen ZVn

Def: Bedingte Verteilungs- & Dichtefunktion stetiger ZV

Die *bedingte Verteilungsfunktion* einer stetigen ZV X , gegeben $Y = y$ ist

$$F_{X|Y}(x|y) := \int_{-\infty}^x \frac{f_{X,Y}(u,y)}{f_Y(y)} du$$

für alle y mit $f_Y(y) > 0$.

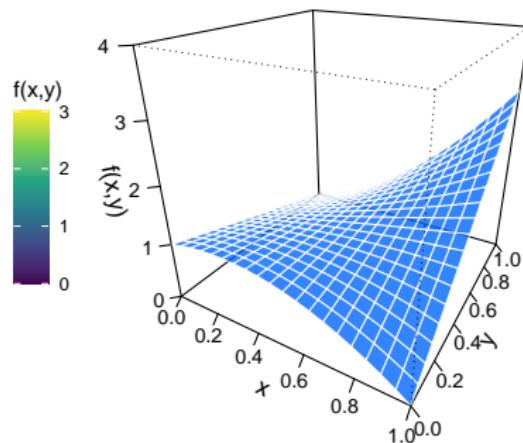
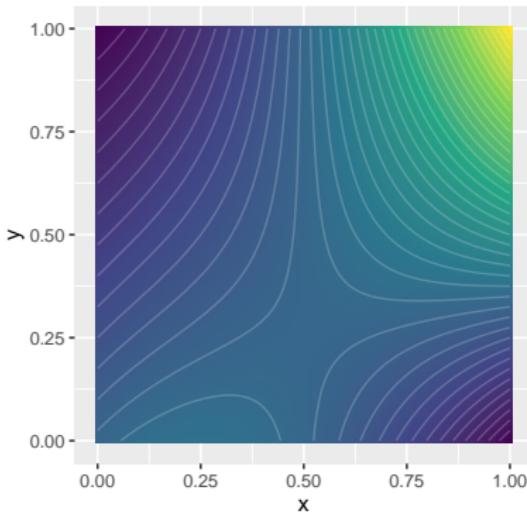
Die *bedingte Dichte* von X gegeben $Y = y$ ist somit

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Beispiel: Gemeinsame, bedingte und Rand-Dichten

Sei $(X, Y) \in [0, 1] \times [0, 1]$ ein stetiger Zufallsvektor mit gemeinsamer Dichte

$$f_{X,Y}(x, y) = (1 + x - y - 2x^2 + 4x^2y) \cdot I(x \in [0, 1]) \cdot I(y \in [0, 1]) :$$



- ▶ Was ist die Rand-Dichte von X ?
- ▶ Was ist die bedingte Dichte von $Y|X = x$?

Verteilung einer diskreten und einer stetigen ZV

Das Konzept von gemeinsamer und bedingter Verteilung lässt sich problemlos auch auf zwei Zufallsvariablen verallgemeinern, von denen eine diskret und eine stetig ist.

Bsp:

Bedingt binomialverteilte Zufallsvariable Y , deren Erfolgswahrscheinlichkeit π selbst eine Zufallsvariable X ist:

$$\begin{aligned} \text{Sei } X &\sim \mathcal{B}(a, \beta) \\ \text{und } Y|X &\sim \mathcal{B}(n, \pi = X) \end{aligned}$$

Beispiel: Die gemeinsame Verteilung

Die gemeinsame Verteilung von X und Y ist

$$\begin{aligned}f(x, y) &= f(y|x) \cdot f(x) \\&= \binom{n}{y} x^y (1-x)^{n-y} \cdot \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \\&= \frac{1}{B(\alpha, \beta)} \binom{n}{y} x^{y+\alpha-1} (1-x)^{n-y+\beta-1}\end{aligned}$$

für $x \in [0, 1]$ und $y \in \{0, 1, \dots, n\}$.

Beispiel: Die bedingte Verteilung von $X|Y = y$

Für die bedingte Dichte $f(x|y)$ folgt:

$$f(x|y) = \frac{f(x,y)}{f(y)} \stackrel{(*)}{\propto} x^{y+\alpha-1}(1-x)^{n-y+\beta-1}$$

also: $X|Y = y \sim \mathcal{B}e(\alpha + y, \beta + n - y)$

Bei (*) haben wir ausgenutzt, dass der Nenner $f(y)$ in

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

nicht von x abhängt, also für $Y = y$ konstant ist.

Beispiel: Die Randverteilung von Y

Damit folgt für $f(y) = f(x, y)/f(x|y)$:

$$\begin{aligned}f(y) &= \binom{n}{y} \frac{B(y + \alpha, n - y + \beta)}{B(\alpha, \beta)} \\&= \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + n)}}_{\text{hängt nicht von } y \text{ ab}} \binom{n}{y} \Gamma(\alpha + y) \Gamma(\beta + n - y)\end{aligned}$$

für $y = 0, \dots, n$.

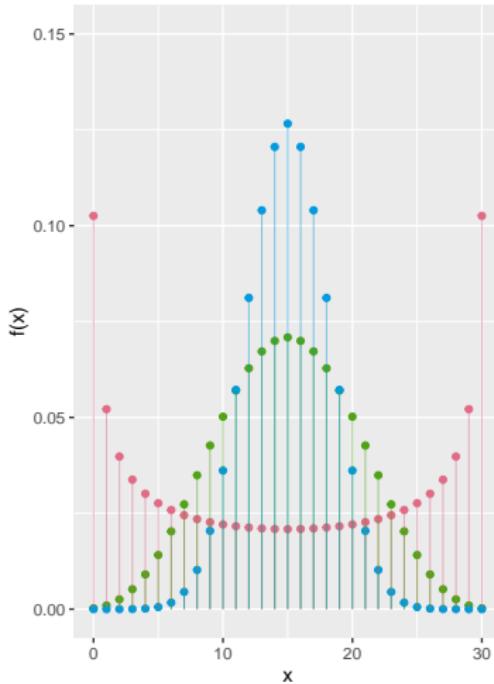
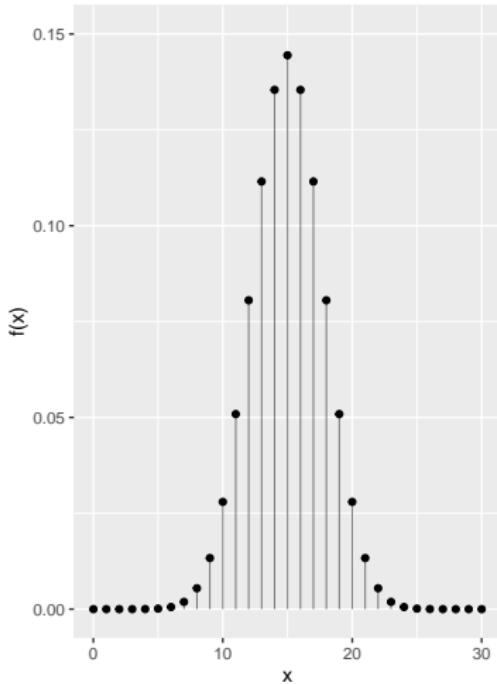
Diese Verteilung heißt “**Beta-Binomialverteilung**” mit Parameter $n \in \mathbb{N}$, $\alpha, \beta \in \mathbb{R}^+$:

$$Y \sim \mathcal{BB}(n, \alpha, \beta)$$

Inhaltlich: Verteilung der Anzahl Erfolge in n Versuchen bei unabhängigen Bernoulliexperimenten mit variierender Erfolgswahrscheinlichkeit (genauer: Beta-verteilter Erfolgsw.keit).

Beta-Binomial-Verteilung

$B(n = 30, p = 0.5)$ und Beta-Binomial-Verteilungen



$\alpha, \beta = 0.5$ in rot; $\alpha, \beta = 5$ in grün; $\alpha, \beta = 50$ in blau

Zufallsvektoren & multivariate Verteilungen

Zufallsvektoren

Stochastische Unabhängigkeit von Zufallsvariablen

Faltungen

Bedingte Verteilungen und Dichten

Bedingte Momente

Bedingte Momente

Def: Bedingte Momente von Zufallsvariablen

Völlig analoge Definitionen basierend auf den bedingten Dichten – es gilt:

$$E(X|Z = z) := \begin{cases} \sum_{x \in T_X} x P(X = x|Z = z) & X \text{ diskret} \\ \int_{-\infty}^{\infty} x f_{X|Z}(x|Z = z) dx & X \text{ stetig} \end{cases}$$

$$\begin{aligned} \text{Var}(X|Z = z) &:= E[(X - E(X|Z = z))^2 | Z = z] \\ &= \begin{cases} \sum_{x \in T_X} (x - E(X|Z = z))^2 P(X = x|Z = z) & X \text{ diskret} \\ \int_{-\infty}^{\infty} (x - E(X|Z = z))^2 f_{X|Z}(x|Z = z) dx & X \text{ stetig} \end{cases} \end{aligned}$$

Beachte:

- $E(X|Z = z)$ ist ein *konkreter Zahlenwert*, der eine Eigenschaft der bedingten Verteilung von X für eine feste Bedingung $Z = z$ beschreibt.
- $E(X|Z)$ ist als Funktion der Werte von Z selbst wieder eine *Zufallsvariable*.

Satz vom iterierten Erwartungswert

Satz vom iterierten Erwartungswert

Für beliebige Funktionen g und Zufallsvariablen X, Z gilt:

$$E(E(g(X)|Z)) = E(g(X))$$

Es gilt also auch: $E(E(X|Z)) = E(X)$

Beweis (diskreter Fall):

$$\begin{aligned} E_Z(E_{X|Z}(g(X)|Z)) &= \sum_{z \in T_z} E_{X|Z}(g(X)|Z = z)P(Z = z) = \sum_{z \in T_z} \sum_{x \in T_x} g(x)P(X = x|Z = z)P(Z = z) \\ &= \sum_{x \in T_x} \sum_{z \in T_z} g(x)P(X = x, Z = z) = \sum_{x \in T_x} g(x)P(X = x) \\ &= E(g(X)) \end{aligned}$$

Satz von der totalen Varianz

Satz von der totalen Varianz

Für beliebige Zufallsvariablen X, Z gilt:

Die Varianz von X ist die Summe der erwarteten bedingten Varianz von X gegeben Z und der Varianz der bedingten Erwartung von X gegeben Z :

$$\text{Var}(X) = E(\text{Var}(X|Z)) + \text{Var}(E(X|Z))$$

- Analog zur *Streuungszerlegung* in geschichteten Stichproben mit schichtweisen Mittelwerten \bar{x}_j und schichtweisen Varianzen s_{xj}^2 , $j = 1, \dots, r$ –
“Gesamtvarianz = Varianz innerhalb der Schichten + Varianz zwischen den Schichten”:

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

Satz von der totalen Varianz - Beweis

Es gilt: $\text{Var}(X|Z) = E(X^2|Z) - (E(X|Z))^2$, also auch

$$\begin{aligned} E_Z \left(\text{Var}_{X|Z} (X|Z) \right) &= E_Z \left(E_{X|Z} (X^2|Z) \right) - E_Z \left((E_{X|Z} (X|Z))^2 \right) \\ &= E_X (X^2) - E_Z \left((E_{X|Z} (X|Z))^2 \right) \\ &= E_X (X^2) - E_X (X)^2 + E_X (X)^2 - E_Z \left((E_{X|Z} (X|Z))^2 \right) \\ &= \text{Var}_X (X) - \left(E_Z \left((E_{X|Z} (X|Z))^2 \right) - \underbrace{E_X (X)^2}_{= E_Z (E_{X|Z} (X|Z))^2} \right) \\ &= \text{Var}_X (X) - \text{Var}_Z (E_{X|Z} (X|Z)) \\ \implies \text{Var}(X) &= E(\text{Var}(X|Z)) + \text{Var}(E(X|Z)) \end{aligned}$$

Schätzung & Grenzwertsätze

Ausblick: Parameterschätzung

Das Gesetz der großen Zahlen

Fundamentalsatz der Statistik

Der zentrale Grenzwertsatz

Schätzung & Grenzwertsätze

Ausblick: Parameterschätzung

Das Gesetz der großen Zahlen

Fundamentalsatz der Statistik

Der zentrale Grenzwertsatz

Schätzung

Bisher kennengelernt:

- ▶ Kennwerte und Methoden zur Beschreibung von Daten
- ▶ Theoretische Eigenschaften von Zufallsvariablen

Offensichtlich:

Enger Zusammenhang – Beobachtete Kennwerte haben theoretische Entsprechungen (z.B. $\bar{x} \leftrightarrow E(X)$)

Verbindung:

Daten als Realisierungen von Zufallsvariablen

Nächster Schritt:

Rückschluss von Daten auf *Verteilungsparameter* der zugrundeliegenden Zufallsvariablen

⇒ **Schätzung**

Ausblick: Schätzung

Daten: x_1, \dots, x_n

Modellannahme: Daten sind Realisierungen einer ZV X mit unbekannten Verteilungsparametern $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$.

3 Fragestellungen

1. Welcher Wert von $\boldsymbol{\theta}$ "passt" am besten zu den beobachteten Daten?
 \Rightarrow *Punktschätzung* der Parameter
2. Wie präzise wird dieser "beste" Wert durch die Daten bestimmt?
 Welcher Bereich von Parameterwerten ist mit den beobachteten Daten "ähnlich gut verträglich"?
 \Rightarrow *Intervallschätzung*, Quantifizierung von Unsicherheit
3. Ist ein bestimmter vorgebener Wert von $\boldsymbol{\theta}$ "kompatibel" mit den beobachteten Daten?
 \Rightarrow *statistische Tests*

→ Stoff späterer Vorlesungen

Ausblick: Schätzung

- ▶ Ein **Schätzer** (= estimator) definiert eine *neue Zufallsvariable* als Funktion von anderen Zufallsvariablen .
Theoretische Perspektive: $\bar{X}_n(\omega) = \frac{1}{n} \sum_{i=1}^n X_i(\omega)$
- ▶ Ein Schätzer ist eine *Rechenvorschrift* um aus beobachteten Daten einen konkreten numerischen Wert (= **Schätzung**, estimate) zu bestimmen.
Empirische Perspektive: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- ▶ Die Rechenvorschrift wird so konstruiert, dass die Verteilung der resultierenden Zufallsvariable möglichst präzise und/oder unverzerrte Schlüsse auf die **zu schätzende** Verteilungseigenschaft/-parameter (= estimand) zulässt.
z.B. $\bar{x} \approx E(X)$ für großes n

Im Folgenden:

Eigenschaften des *arithmetischen Mittels* und der *ECDF* eines beobachteten Merkmals als Schätzer für den *Erwartungswert* bzw. für die *Verteilungsfunktion* der zugrundeliegenden Verteilung.

Ausblick: Schätzung



estimand



estimate

Ingredients

150g unsalted butter, plus extra for greasing
150g plain chocolate, broken into pieces
150g plain flour
½ tsp baking powder
½ tsp bicarbonate of soda
200g light muscovado sugar
2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.
2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

estimator

Abb.: Simon Grund (@simongrund89)

Schätzung & Grenzwertsätze

Ausblick: Parameterschätzung

Das Gesetz der großen Zahlen

Fundamentalsatz der Statistik

Der zentrale Grenzwertsatz

Das Gesetz der großen Zahlen (GGZ)

Schwaches Gesetz der großen Zahlen

Für das arithmetische Mittel $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ von unabhängig und identisch verteilten (“i.i.d.”) Zufallsvariablen aus einer Verteilung mit Erwartungswert $E(X_i) = \mu_X$ und Varianz $\text{Var}(X_i) = \sigma_X^2 < \infty$ gilt:

$$E(\bar{X}_n) = \mu_X \text{ und } \text{Var}(\bar{X}_n) = \frac{1}{n} \sigma_X^2 \xrightarrow{n \rightarrow \infty} 0$$

$$\implies P(|\bar{X}_n - \mu_X| \geq \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \epsilon > 0$$

Man schreibt: $\bar{X}_n \xrightarrow{P} \mu_X$

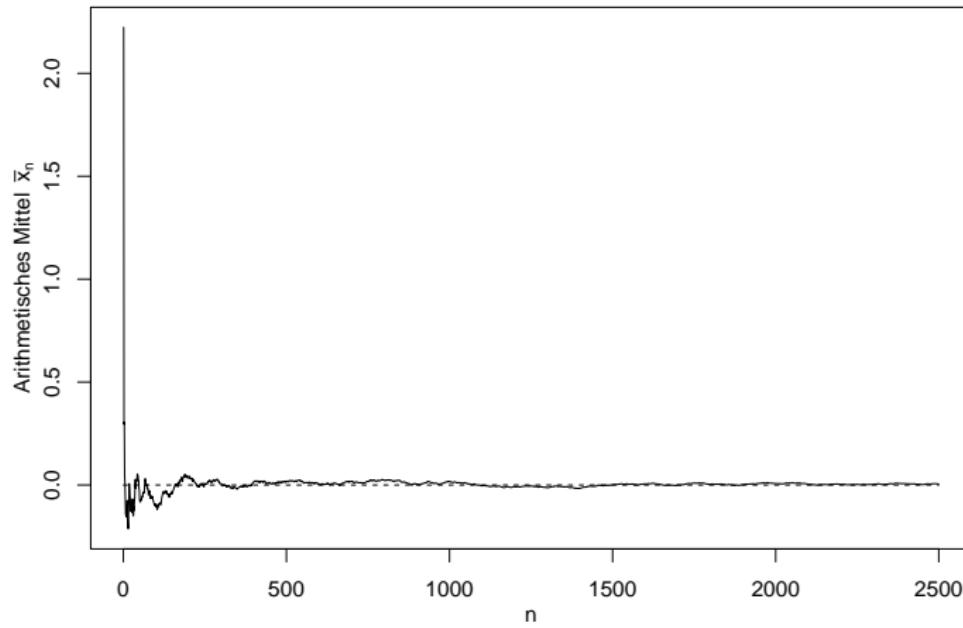
⇒ Das arithmetische Mittel beobachteter i.i.d. Daten konvergiert “in Wahrscheinlichkeit” gegen den Erwartungswert der datengenerierenden Verteilung.

Die Varianz dieses Schätzers wird bei wachsendem Stichprobenumfang beliebig klein.

Terminologie: *i.i.d.* := “independent and identically distributed”

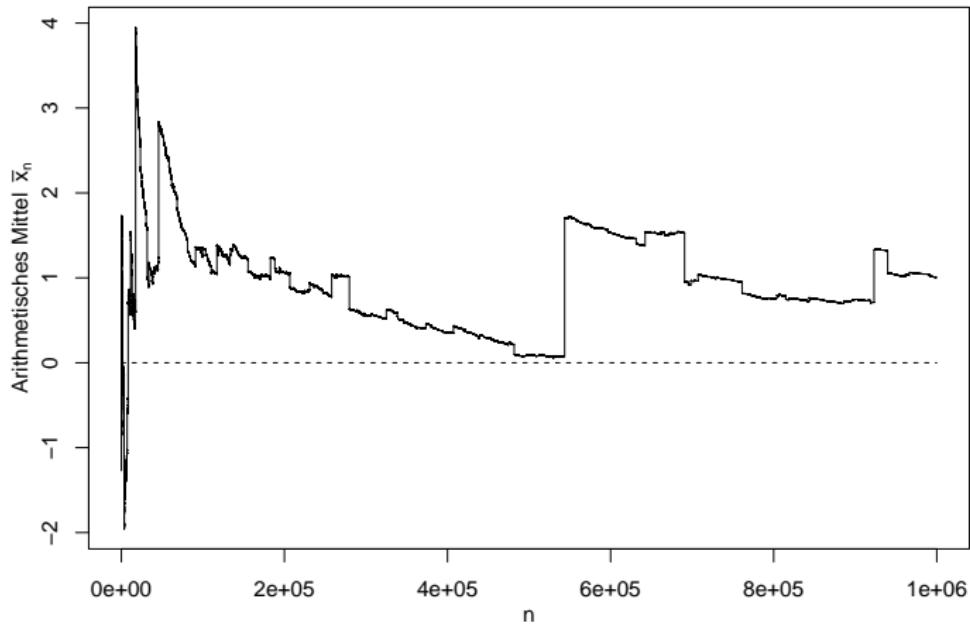
Beispiel: Normalverteilung

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \text{ mit steigendem } n \text{ für iid } X_i \sim N(0, 1)$$



Gegenbeispiel: Cauchyverteilung

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n x_i \text{ mit steigendem } n \text{ für Cauchy-verteilte } X_i$$



Schätzung & Grenzwertsätze

Ausblick: Parameterschätzung

Das Gesetz der großen Zahlen

Fundamentalsatz der Statistik

Der zentrale Grenzwertsatz

Punktweise Konvergenz der empirischen Verteilung

Weil für beliebige Zufallsvariablen X und Mengen B gilt:

$$E(I(X \in B)) = P(X \in B)$$

impliziert das GGZ für *iid* ZV $X_i; i = 1, \dots, n$ mit Verteilung $F_X(x)$ speziell auch:

$$\frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \xrightarrow{P} P(X \leq x) = F_X(x).$$

Also gilt für die *empirische* Verteilungsfunktion $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$

$$F_n(x) \xrightarrow{P} F_X(x) \quad \forall x$$

⇒ Die empirische Verteilungsfunktion konvergiert punktweise gegen die “wahre” Verteilungsfunktion der datengenerierenden Verteilung

Notation: Indikatorfunktion $I(x \in B) := \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases}$; oft auch $I_B(x); \mathbb{1}_B(x); \delta_x(B)(!!)$

Fundamentalsatz der Statistik

Fundamentalsatz der Statistik (Satz v. Glivenko-Cantelli)

Für iid ZV X_i , $i = 1, \dots, n$ mit Verteilungsfunktion $F_X(x)$ gilt für die empirische Verteilungsfunktion $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$:

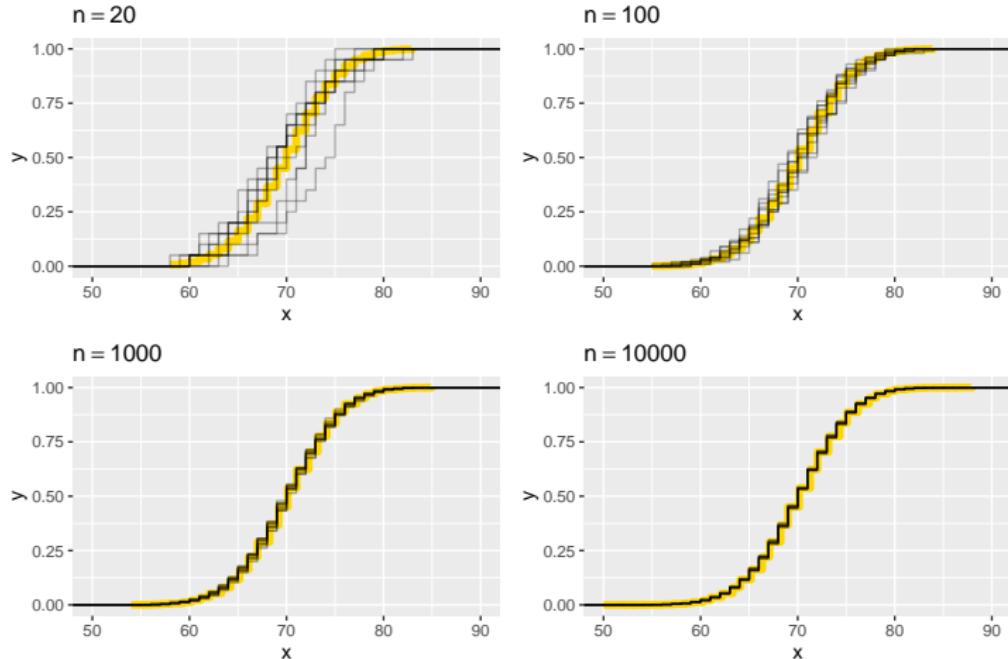
$$P\left(\sup_{x \in \mathbb{R}} (|F_n(x) - F_X(x)|) < \epsilon\right) \xrightarrow{n \rightarrow \infty} 1 \quad \forall \epsilon > 0 \quad \forall x$$

⇒ Die **maximale Abweichung zwischen ECDF und Verteilungsfunktion der datengenerierenden Verteilung wird für wachsenden Stichprobenumfang mit Sicherheit beliebig klein.**

- stärkere Aussage als punktweiser Konvergenz in Wahrscheinlichkeit: die **maximale Abweichung zwischen Schätzung F_n und Wahrheit F geht mit Wahrscheinlichkeit 1 gegen Null.**

Beispiel Glivenko-Cantelli: Binomialverteilung

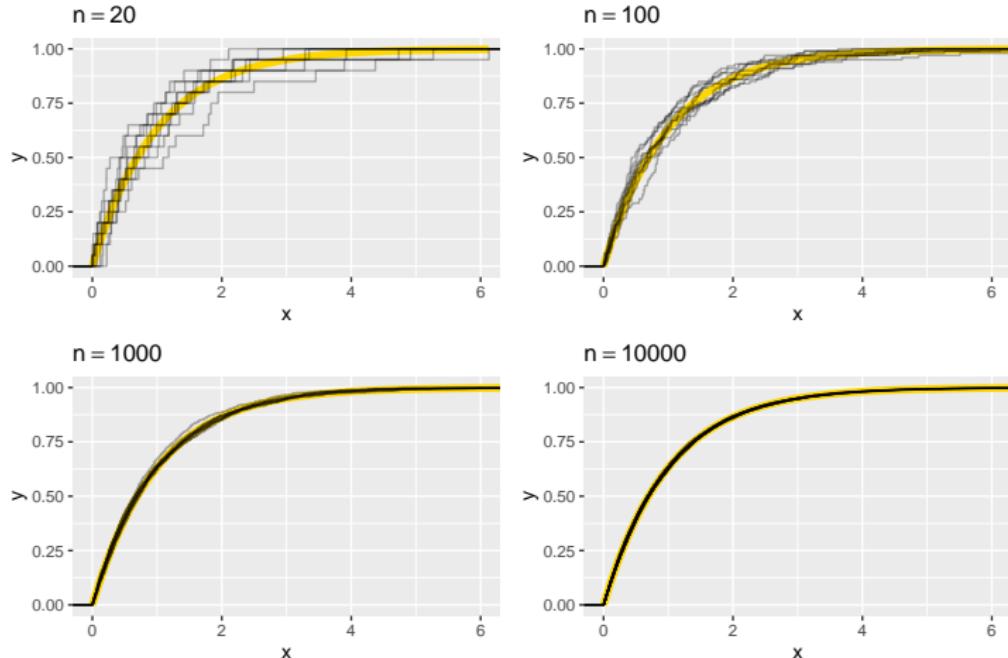
Binomialverteilung $B(100, 0.7)$: $F_n(x)$ & $F(x)$



ECDFs in schwarz, $F(x)$ in gold.
Jeweils 10 Stichproben mit n Beobachtungen.

Beispiel Glivenko-Cantelli: Exponentialverteilung

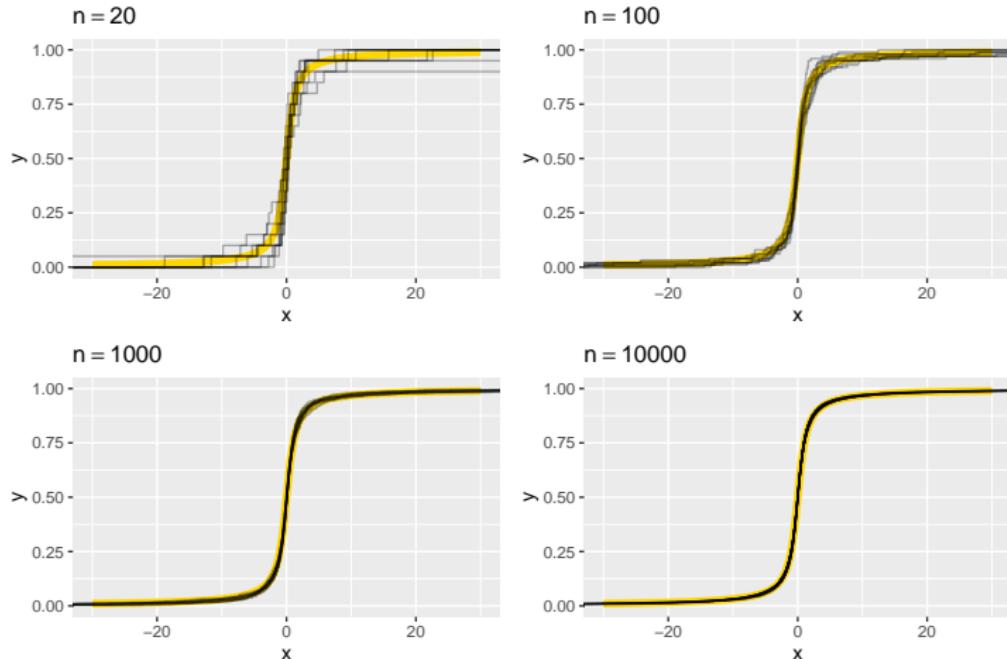
Exponentialverteilung $E(1)$: $F_n(x)$ & $F(x)$



ECDFs in schwarz, $F(x)$ in gold.
Jeweils 10 Stichproben mit n Beobachtungen.

Beispiel Glivenko-Cantelli: Cauchyverteilung

Cauchyverteilung: $F_n(x)$ & $F(x)$



ECDFs in schwarz, $F(x)$ in gold.
Jeweils 10 Stichproben mit n Beobachtungen.

Schätzung & Grenzwertsätze

Ausblick: Parameterschätzung

Das Gesetz der großen Zahlen

Fundamentalsatz der Statistik

Der zentrale Grenzwertsatz

Der zentrale Grenzwertsatz

- ▶ “Die Verteilung des arithmetischen Mittels von n iid Zufallsvariablen aus einer *beliebigen(!)* Verteilung konvergiert für wachsendes n gegen eine Normalverteilung.” (... unter nicht besonders strengen Bedingungen)
- ▶ Begründet die zentrale Rolle der Normalverteilung in der Stochastik und Statistik.

Zunächst müssen wir noch *standardisierte* Zufallsvariablen definieren.

Standardisierte Zufallsvariablen

Def.: Standardisierte Zufallsvariable

Jede Zufallsvariable X mit endlichem Erwartungswert $\mu_X = E(X)$ und endlicher Varianz $\sigma_X^2 = \text{Var}(X)$ kann man derart linear transformieren, dass sie Erwartungswert 0 und Varianz 1 besitzt:

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}$$

Dann gilt:

$$E(\tilde{X}) = \frac{1}{\sigma} (E(X) - \mu_X) = 0$$

$$\text{Var}(\tilde{X}) = \frac{1}{\sigma_X^2} \text{Var}(X) = 1$$

Standardisierung von summierten iid ZVn

Betrachte *iid* Zufallsvariablen X_1, X_2, \dots, X_n mit endlichem Erwartungswert $\mu_X = E(X_i)$ und endlicher Varianz $\sigma_X^2 = \text{Var}(X_i)$.

Für die Summe $Y_n = X_1 + X_2 + \dots + X_n$ gilt also

$$E(Y_n) = n \cdot \mu_X,$$
$$\text{Var}(Y_n) = n \cdot \sigma_X^2.$$

Für die *standardisierte Summe*

$$Z_n = \frac{Y_n - n\mu_X}{\sqrt{n}\sigma_X} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu_X}{\sigma_X}$$

gilt somit $E(Z_n) = 0$ und $\text{Var}(Z_n) = 1$.

Der zentrale Grenzwertsatz (ZGWS)

Zentraler Grenzwertsatz

Die Verteilungsfunktion $F_n(z)$ der standardisierten Summe Z_n von i.i.d. Zufallsvariablen mit existierendem Erwartungswert und endlicher Varianz konvergiert für $n \rightarrow \infty$ an jeder Stelle $z \in \mathbb{R}$ gegen die Verteilungsfunktion $F_{N(0,1)}(z)$ der Standardnormalverteilung:

$$F_n(z) \xrightarrow{n \rightarrow \infty} F_{N(0,1)}(z) \quad \forall z \in \mathbb{R}$$

$$\text{also } Z_n \stackrel{a}{\sim} \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

In der Praxis kann man also die Verteilung von Z_n für große n oft gut durch eine Standardnormalverteilung approximieren.

“ $\stackrel{a}{\sim}$ ” :≈ “asymptotisch/approximativ verteilt wie”

Englisch: central limit theorem CLT

Bemerkungen

- ▶ Solange Erwartungswert und Varianz existieren, gilt der ZGWS auch für
 - ▶ diskrete X_i ,
 - ▶ X_i mit anderem Träger $T_X \neq \mathbb{R}$ als die Normalverteilung,
 - ▶ und X_i mit schießen oder multimodalen Verteilungen!
- ▶ Unter zusätzlichen Bedingungen an höhere Momente von $|X_i|$ (Lyapunov-/Lindeberg-ZGWS) gilt der ZGWS sogar auch für *unabhängige, aber nicht identisch verteilte X_i .*
- ▶ Die Standardisierung ist nicht zwingend notwendig zur Formulierung des ZGWS. Alternativ kann man auch direkt die Summe $Y_n = X_1 + \dots + X_n$ betrachten. Dann gilt

$$Y_n \xrightarrow{a} \mathcal{N}(\mu = n\mu_X, \sigma^2 = n\sigma_X^2)$$

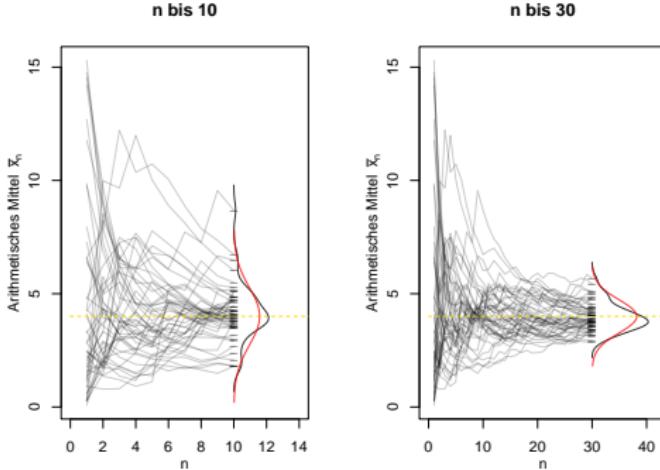
- ▶ Damit dann im speziellen auch (vgl. GGZ):

$$\bar{X}_n \xrightarrow{a} \mathcal{N}\left(\mu = \mu_X, \sigma^2 = \frac{\sigma_X^2}{n}\right)$$

Bemerkungen

- ▶ GGZ liefert asymptotische Aussage über EW und Varianz des arithmetischen Mittels.
- ▶ ZGWS liefert Aussage über die *komplette (asymptotische) Verteilung* des arithmetischen Mittels
- ▶ der ZGWS gilt sowohl für stetige als auch für diskrete ZV X_i .
- ▶ Anwendung: Die Verteilung der Summe/des Mittelwerts vieler (in etwa) *iid* Zufallsvariablen kann gut durch ein Normalverteilungsmodell approximiert werden.

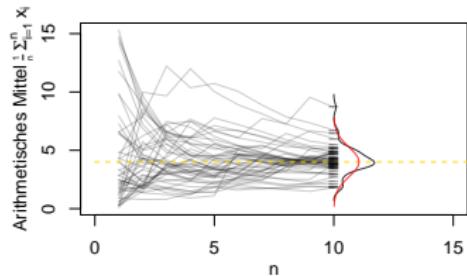
Veranschaulichung ZGWS und GGZ



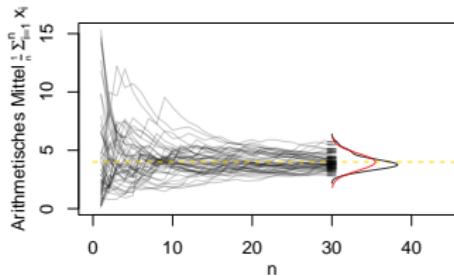
- ▶ Graue Linien: 50 Verläufe des arithmetischen Mittelwerts von jeweils n $\mathcal{E}(\lambda = 0.25)$ Zufallsvariablen ($E(X_i) = 4$, $\text{Var}(X_i) = 16$) für wachsendes n .
- ▶ Kleine Querstriche zeigen die Ergebnisse für das maximale n an, goldene Linie den wahren Erwartungswert.
- ▶ Schwarze Kurve ist beobachtete Dichte dieser 50 Mittelwerte, rote die vom ZGWS "vorhergesagte" $\mathcal{N}(\mu = 4, \sigma^2 = 16/n)$.

Veranschaulichung ZGWS und GGZ

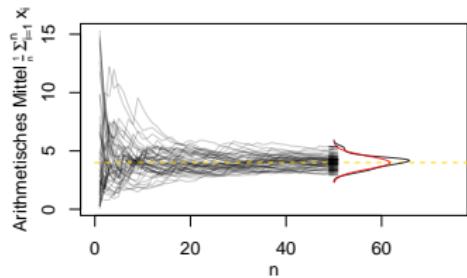
n bis 10



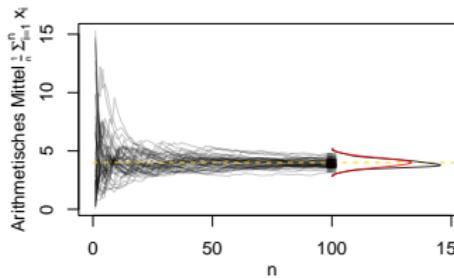
n bis 30



n bis 50



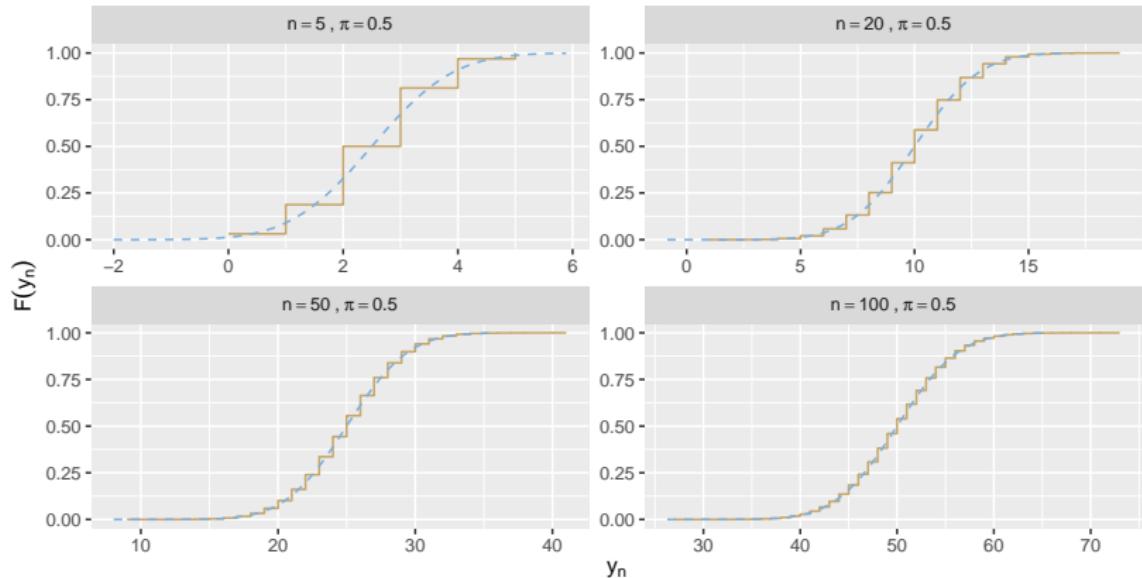
n bis 100



Diskreter Fall: Approximation der Binomialv.

Seien $X_i \stackrel{\text{iid}}{\sim} \mathcal{B}(\pi)$, $i = 1, \dots, n$. Dann ist $Y_n = \sum_{i=1}^n X_i \sim \mathcal{B}(n, \pi)$ und

$$\frac{Y_n - n\pi}{\sqrt{n\pi(1-\pi)}} \stackrel{a}{\sim} \mathcal{N}(0, 1) \iff Y_n \stackrel{a}{\sim} \mathcal{N}(\mu = n\pi, \sigma^2 = n\pi(1-\pi)).$$

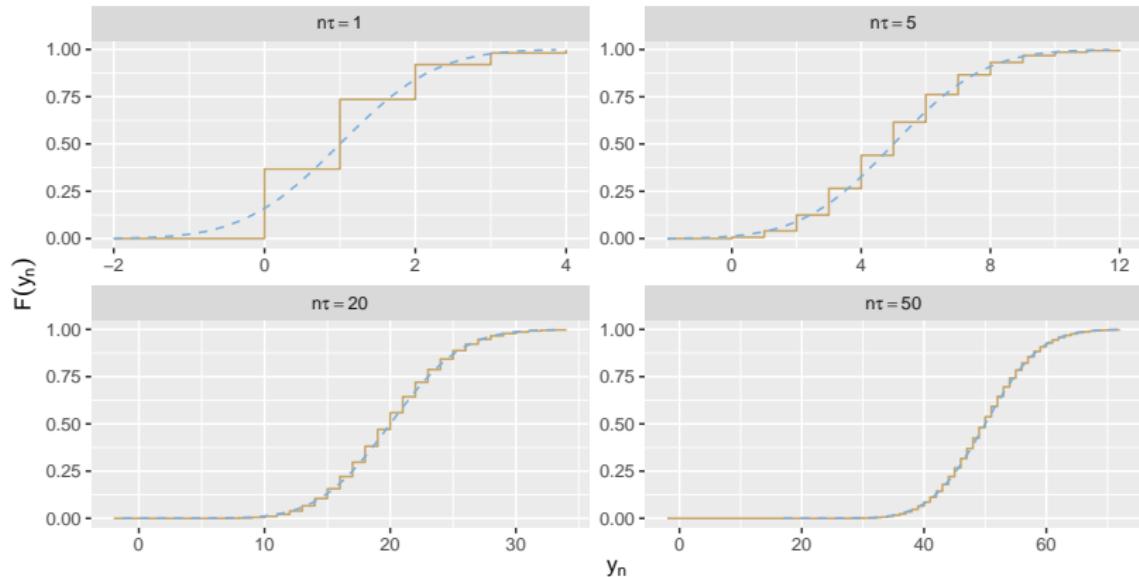


$\mathcal{N}(\mu = n\pi, \sigma^2 = n\pi(1-\pi))$ in blau (gestrichelt); $\mathcal{B}(n = 100, \pi = 0.5)$ in gold.

Diskreter Fall: Approximation der Poissonv.

Seien $X_i \stackrel{\text{iid}}{\sim} \mathcal{P}(\lambda = \tau)$, $i = 1, \dots, n$. Dann ist $Y_n = \sum_{i=1}^n X_i \sim \mathcal{P}(\lambda = n\tau)$ und

$$\frac{Y_n - n\tau}{\sqrt{n\tau}} \xrightarrow{a} \mathcal{N}(0, 1) \iff Y_n \xrightarrow{a} \mathcal{N}(n\tau, n\tau).$$

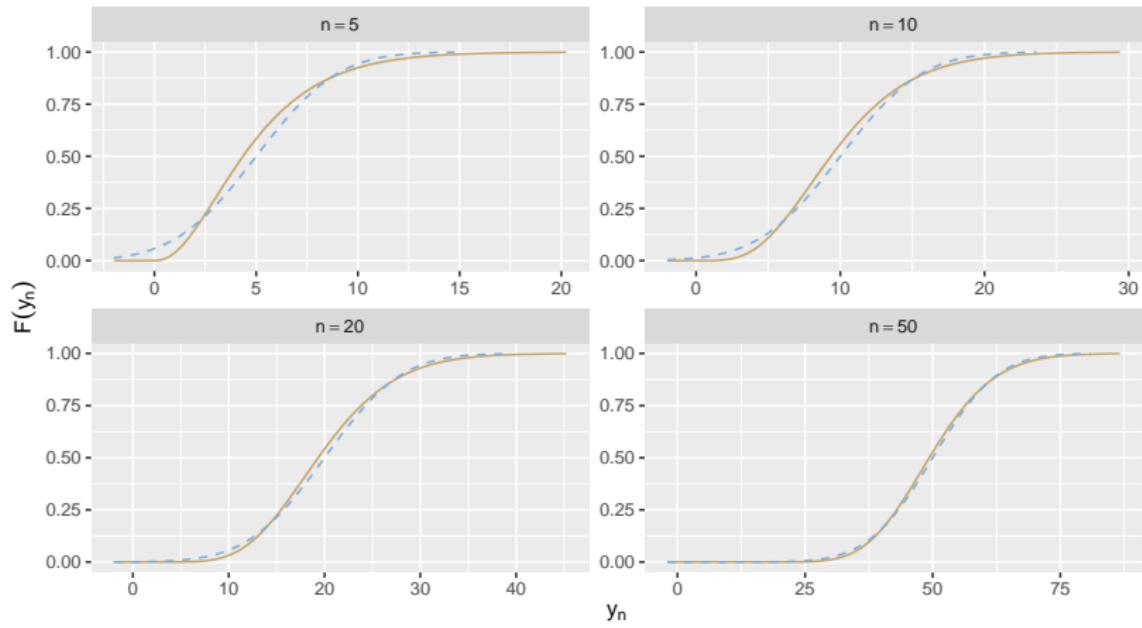


$N(m = n\tau, \sigma^2 = n\tau)$ in blau (gestrichelt); $P(\lambda = n\tau)$ in gold.

Stetiger Fall: Approximation der χ^2 -Verteilung

Seien $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $i = 1, \dots, n$. Dann ist $Y_n = \sum_{i=1}^n X_i^2 \sim \chi^2(d = n)$ und

$$\frac{Y_n - n}{\sqrt{2n}} \xrightarrow{a} \mathcal{N}(0, 1) \iff Y_n \xrightarrow{a} \mathcal{N}(n, 2n).$$



$N(m=n, \sigma^2=2n)$ in blau (gestrichelt); $\chi^2(n)$ in gold.

Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Kovarianz

Kennzahl für Stärke und Richtung des linearen Zusammenhang zweier metrischer Merkmale:

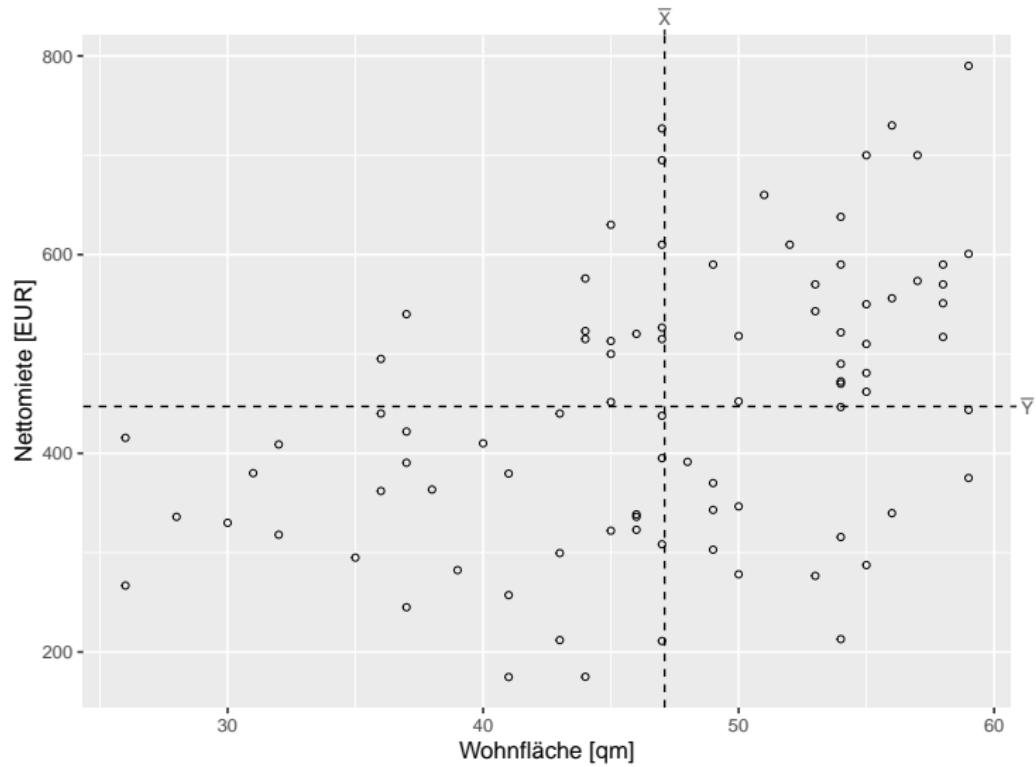
Daten: $(x_i, y_i), i = 1, \dots, n$

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

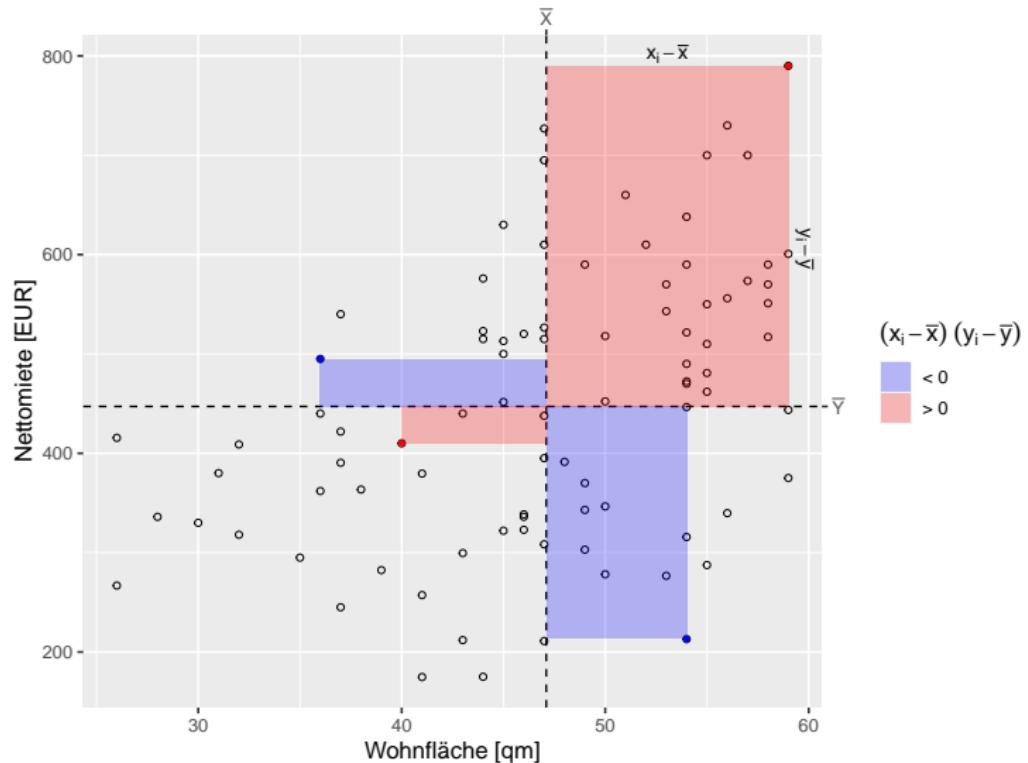
Beachte:

- ▶ Summand i positiv, falls Abweichungen von x_i und y_i zum jeweiligen Mittelwert das gleiche Vorzeichen haben, negativ falls unterschiedlich.
- ▶ Für S_{xx} ergibt sich die Varianz von X : $S_{xx} = S_x^2$
- ▶ Der Wert der Kovarianz hängt sowohl von den **Streuungen** der beiden Merkmal als auch von der **Stärke und Richtung ihres Zusammenhangs** ab.

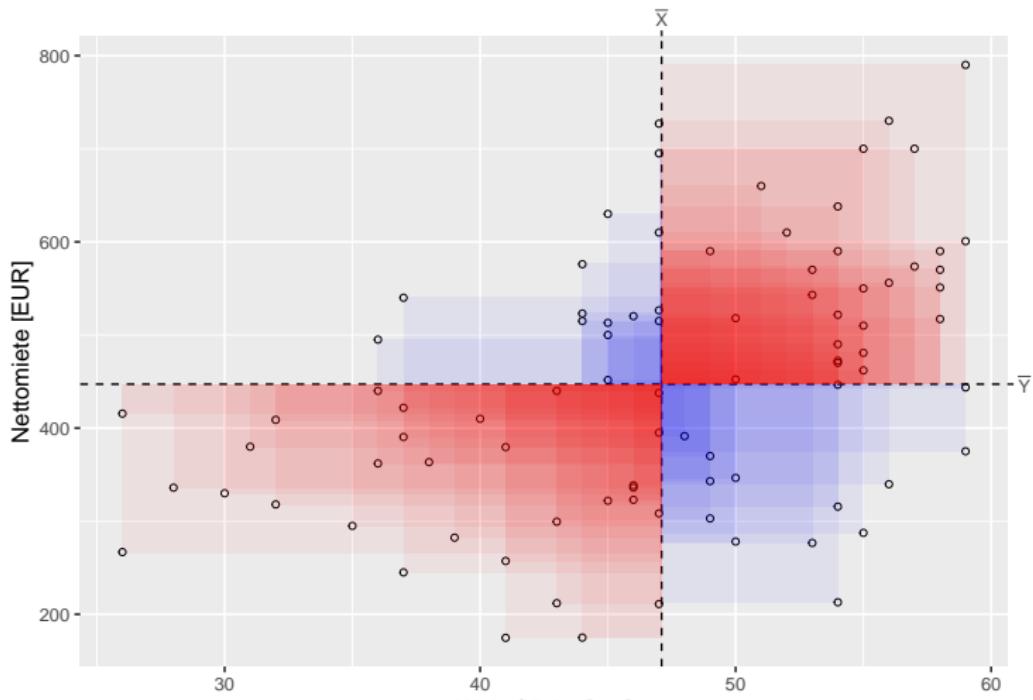
Kovarianz: Geometrische Intuition



Kovarianz: Geometrische Intuition



Kovarianz: Geometrische Intuition



$$S_x^2 = 70, S_y^2 = 19600, S_{xy} = 505, r_{xy} = 0.43$$

Bravais-Pearson-Korrelationskoeffizient

Der Bravais-Pearson-Korrelationskoeffizient r_{xy} ergibt sich aus den Daten $(x_i, y_i), i = 1, \dots, n$ durch

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

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

- dimensionslose Größe, hängt nicht mehr von S_x oder S_y ab:

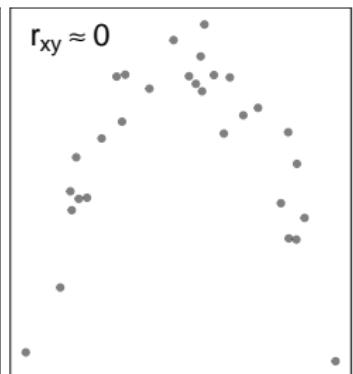
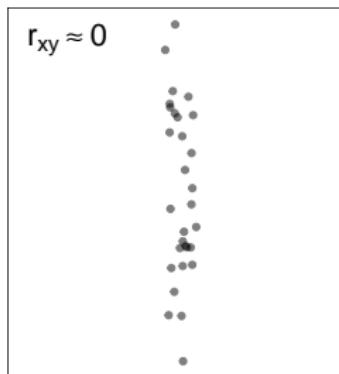
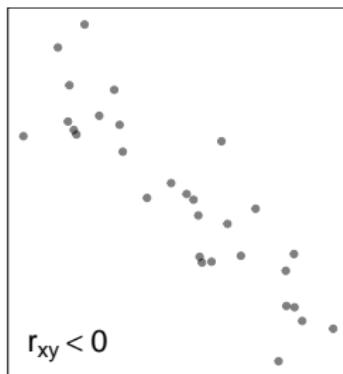
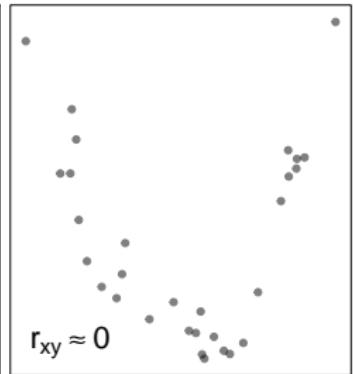
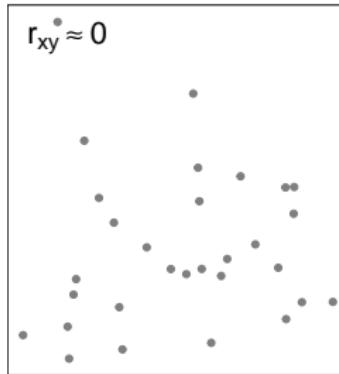
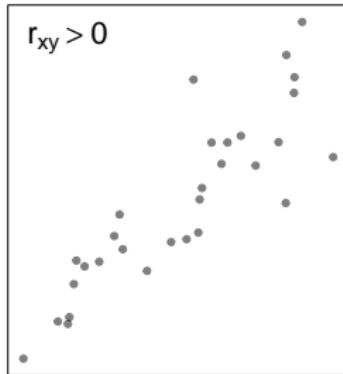
- $r_{xy} > 0$ positive Korrelation, gleichsinniger linearer Zusammenhang,
Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
- $r_{xy} < 0$ negative Korrelation, gegenläufiger linearer Zusammenhang,
Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
- $r_{xy} = 0$ keine Korrelation, unkorreliert, kein linearer Zusammenhang

Eigenschaften des Korrelationskoeffizienten

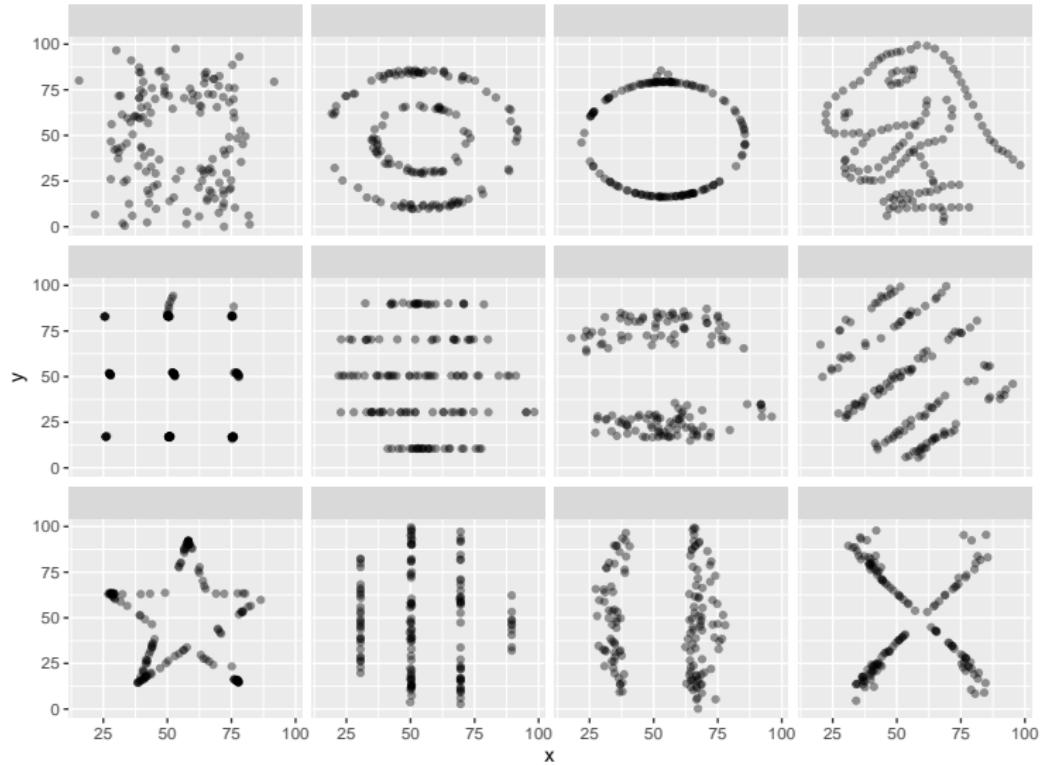
- ▶ Misst vor allem die Stärke des **linearen** Zusammenhangs
- ▶ Betrag der Korrelation unverändert bei linearen Transformationen
- ▶ Symmetrisch: $r_{xy} = r_{yx}$
- ▶ Positive Korrelation bedeutet:
“Je größer X, desto größer im Durchschnitt Y”
- ▶ Korrelation = +1 [-1] falls die Punkte genau auf einer Geraden mit positiver [negativer] Steigung liegen – **unabhängig von Steigung dieser Geraden** (außer sie ist 0 oder ∞).
- ▶ “Korrelation = 0” bedeutet “kein *linearer* Zusammenhang”, aber **nicht Unabhängigkeit!** (s. Bsp. unten)
- ▶ Korrelation (und Kovarianz) sind empfindlich gegenüber Ausreißern

Eigenschaften von r_{xy}

Misst nur Stärke des **linearen** Zusammenhangs:



Eigenschaften Bravais-Pearson-Korrelation



Alle diese Datensätze haben $r_{xy} = -0.06!$

(und $n = 142$, $\bar{x} = 54.3$, $S_x = 16.8$, $\bar{y} = 47.8$, $S_y = 26.9$)

Lineare Transformationen

- Bei exakten linearen Zusammenhängen gilt:

$$r_{xy} = +1 \text{ bzw. } -1 \iff Y = aX + b \text{ mit } a > 0 \text{ bzw. } a < 0$$

- Lineare Transformationen $\tilde{X} = a_X X + b_X$, $\tilde{Y} = a_Y Y + b_Y$, $a_X, a_Y \neq 0$:
 r_{xy} Korrelationskoeffizient zwischen X und Y
 \tilde{r}_{xy} Korrelationskoeffizient zwischen \tilde{X} und \tilde{Y}

$$\begin{aligned} \Rightarrow \quad \tilde{r}_{xy} = r_{xy} &\iff a_X, a_Y \text{ gleiches Vorzeichen} \\ \Rightarrow \quad \tilde{r}_{xy} = -r_{xy} &\iff a_X, a_Y \text{ verschiedene Vorzeichen.} \end{aligned}$$

Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Kovarianz- und Korrelationsmatrix

Bei mehr als zwei Merkmalen werden Kovarianzen und Korrelationen häufig in Form einer Matrix dargestellt.

Auf der Hauptdiagonalen stehen die Stichprobenkovarianzen bzw. die Korrelationen jedes Merkmals mit sich selbst, also die jeweilige Stichprobenvarianz bzw 1.

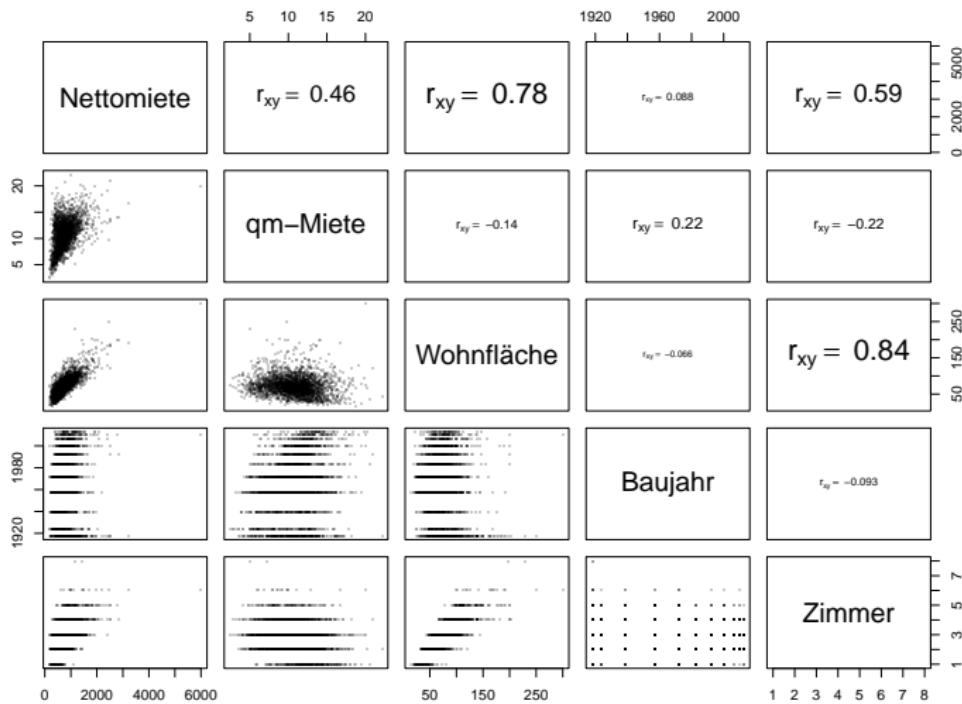
Die Matrix ist symmetrisch da $S_{XY} = S_{YX}$.

Bsp: Korrelationsmatrix der Merkmale X, Y, Z:

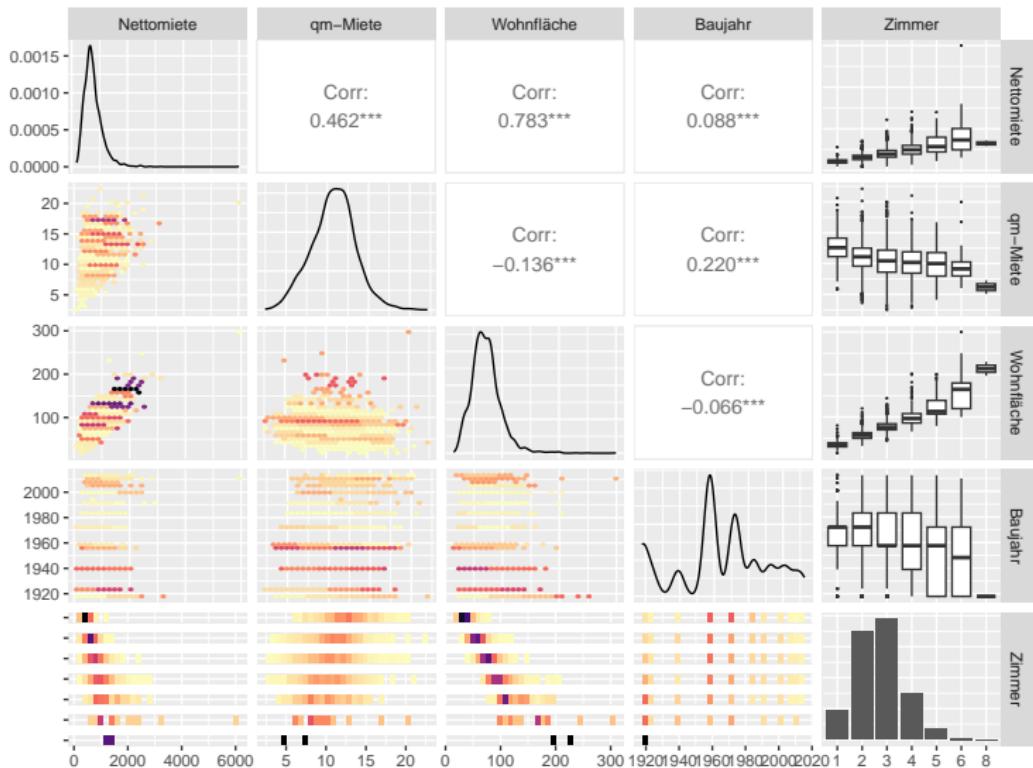
$$\begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix}$$

Analog für Zufallsvektoren mit entsprechenden *Erwartungswertvektoren* und *Kovarianzmatrizen*!

(Scatter)plotmatrix I



(Scatter)plotmatrix II



Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Kovarianzen und Korrelationen

Def.: Kovarianz und Korrelation

Die **Kovarianz** $\text{Cov}(X, Y)$ bzw die **Korrelation** $\rho(X, Y)$ zweier ZVn X, Y sind Maße für die Stärke und Richtung der linearen Abhängigkeit von ZVn X und Y . Es gilt:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$\text{und } \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

Letzteres unter der Voraussetzung, dass $\text{Var}(X) > 0$ und $\text{Var}(Y) > 0$.

Beachte: $\text{Cov}(X, X) = \text{Var}(X)$

Der Verschiebungssatz für die Kovarianz

Es gilt zudem:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Beweis:

$$\begin{aligned}\text{Cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - YE(X) - XE(Y) + E(X)E(Y)) \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y)\end{aligned}$$

Unkorreliertheit

X und Y heißen **unkorreliert**, wenn

$$\text{Cov}(X, Y) = 0 \quad \text{bzw.} \quad \rho(X, Y) = 0$$

d.h. wenn

$$E(XY) = E(X) \cdot E(Y)$$

gilt.

Beachte: Aus Unabhängigkeit folgt Unkorreliertheit, aber der Umkehrschluss gilt i.A. nicht!

X und Y sind **positiv/negativ korreliert** falls

$$\rho(X, Y) > 0 \quad \text{bzw.} \quad \rho(X, Y) < 0$$

Beispiel: Unabhängig vs Unkorreliert

Seien $X \sim \mathcal{B}(\pi = \frac{1}{2})$ und $Y \sim \mathcal{B}(\pi = \frac{1}{2})$ unabhängig.

Betrachte

$$Z_1 = X + Y = \begin{cases} 0 & \text{mit Wkeit } \frac{1}{4} \\ 1 & \text{mit Wkeit } \frac{1}{2} \\ 2 & \text{mit Wkeit } \frac{1}{4} \end{cases}; \quad Z_2 = X - Y = \begin{cases} -1 & \text{mit Wkeit } \frac{1}{4} \\ 0 & \text{mit Wkeit } \frac{1}{2} \\ 1 & \text{mit Wkeit } \frac{1}{4} \end{cases}$$

Dann sind Z_1 und Z_2 zwar unkorreliert aber nicht unabhängig:

$f_{Z_1, Z_2}(z_1, z_2)$	$z_1 = 0$	$z_1 = 1$	$z_1 = 2$	$f_{Z_2}(z_2)$
$z_2 = -1$	0	$1/4$	0	$1/4$
$z_2 = 0$	$1/4$	0	$1/4$	$1/2$
$z_2 = 1$	0	$1/4$	0	$1/4$
$f_{Z_1}(z_1)$	$1/4$	$1/2$	$1/4$	

$\implies f(z_1, z_2) \neq f(z_1)f(z_2)$ d.h. Z_1, Z_2 nicht unabhängig.

Aber $E(Z_1) = 1, E(Z_2) = 0; E(Z_1 Z_2) = 0$, also $\text{Cov}(Z_1, Z_2) = 0!$

Eigenschaften von Korrelationen

Für alle ZVn X und Y gilt:

$$-1 \leq \rho(X, Y) \leq 1$$

$|\rho(X, Y)| = 1$ gilt genau dann, wenn perfekte lineare Abhängigkeit zwischen X und Y besteht:

$$|\rho(X, Y)| = 1 \iff Y = a + b \cdot X \quad \text{für } a, b \in \mathbb{R}; b \neq 0$$

Lineare Transformationen

Sei $a, b, c, d \in \mathbb{R}$ mit $b \cdot d > 0$ und X, Y beliebige ZVn. Dann gilt:

$$\text{Cov}(a + bX, c + dY) = b \cdot d \cdot \text{Cov}(X, Y)$$

Daher gilt:

$$\rho(a + bX, c + dY) = \text{sgn}(b) \cdot \text{sgn}(d) \cdot \rho(X, Y)$$

d.h. die Korrelation ist **invariant** bzgl. linearer Transformationen

$$\text{sgn}(x) := \frac{x}{|x|} = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

Varianz der Summe zweier ZVn

Seien X und Y beliebige ZVn. Dann gilt für $X + Y$:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$$

Daher gilt speziell für *unabhängige* und damit unkorrelierte X und Y :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

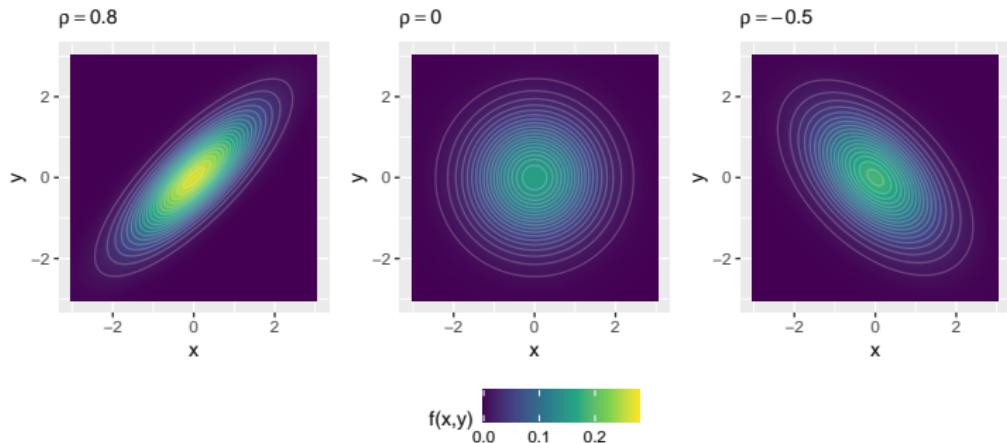
Anwendung: Bivariate Standardnormalverteilung

Die **bivariate Standardnormalverteilung** mit Parameter ρ ($|\rho| < 1$) hat Träger $T = \mathbb{R} \times \mathbb{R}$ und Dichtefunktion

$$f(x, y) = \frac{1}{2\pi\sqrt{(1 - \rho^2)}} \exp\left(-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

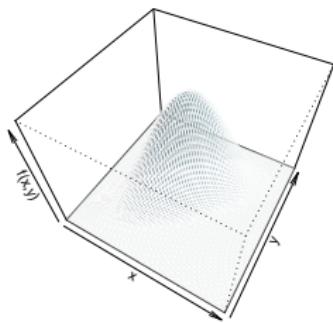
- ▶ Die Randverteilungen von X und Y sind (für jedes ρ) standard-normalverteilt.
- ▶ Die Korrelation zwischen X und Y ist gleich ρ .
- ▶ Aus Unkorreliertheit von X und Y folgt in diesem Spezialfall die Unabhängigkeit von X und Y .

Visualisierung: Bivariate Standardnormalverteilung

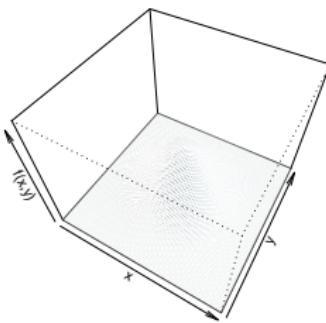


Visualisierung: Bivariate Standardnormalverteilung

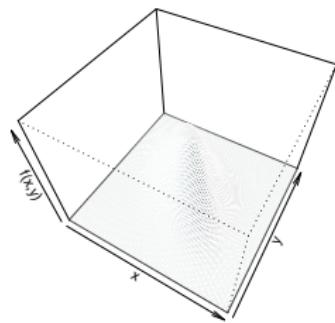
$\rho = 0.8$



$\rho = 0$



$\rho = -0.5$



Multivariate Normalverteilung

Die allgemeine bivariate Normalverteilung erhält man durch lineare Transformation der Komponenten der bivariaten Standardnormalverteilung:

$$X \rightarrow \mu_X + \sigma_X \cdot X; \quad Y \rightarrow \mu_Y + \sigma_Y \cdot Y$$

Insgesamt fünf Parameter: $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$.

Es gilt $\sigma_{XY} := \text{Cov}(X, Y) = \rho\sigma_X\sigma_Y$

Notation:

$$(X, Y) \sim \mathcal{N}_2 \left(\boldsymbol{\mu} = (\mu_X, \mu_Y)^T, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right)$$

Ausblick: Das Konzept ist auf beliebig hohe Dimensionen übertragbar. Die Parameter einer d -dimensionalen NV sind ihr Erwartungswertvektor $\boldsymbol{\mu}$ und ihre Kovarianzmatrix $\boldsymbol{\Sigma}$:

$$\begin{aligned} \mathbf{X} = (X_1, X_2, \dots, X_d)^T &\sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \implies f(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \end{aligned}$$

Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Spearman-/Rang-Korrelationskoeffizient

X, Y (mindestens) ordinal

Idee: Berechne **Korrelation nach Bravais-Pearson für die Ränge** statt für die Werte der Merkmale.

Der Korrelationskoeffizient nach Spearman ist definiert durch

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

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

Vorgehen

- Urliste der Größe nach sortieren
- \Rightarrow Ranglisten $rg(x_i), rg(y_i), i = 1, \dots, n$ vergeben (bei Bindungen: Durchschnittsränge)

z.B:

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

bei Bindungen (ties):

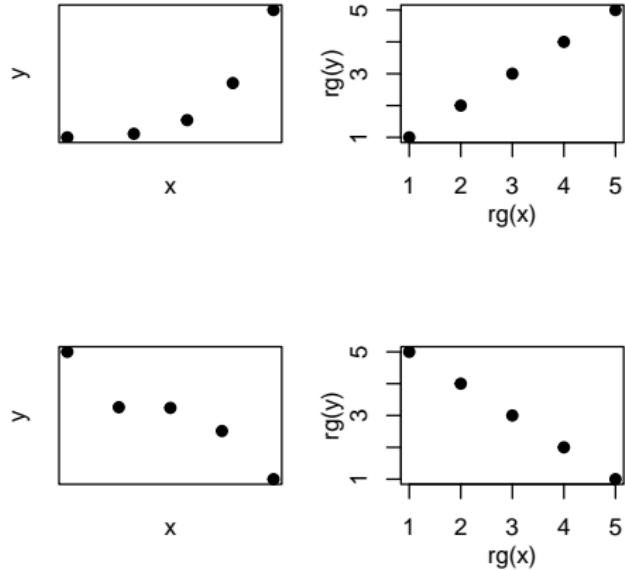
x_i	2.3	7.1	1.0	2.1	2.3
$rg(x_i)$	3.5	5	1	2	3.5

\Rightarrow Durchschnittsrang $\frac{3+4}{2} = 3.5$ vergeben.

Interpretation

- ▶ $r_{xy}^{SP} > 0 \iff$ gleichsinniger monotoner Zusammenhang,
Tendenz: x groß $\Leftrightarrow y$ groß, x klein $\Leftrightarrow y$ klein
- ▶ $r_{xy}^{SP} < 0 \iff$ gegensinniger monotoner Zusammenhang,
Tendenz: x groß $\Leftrightarrow y$ klein, x klein $\Leftrightarrow y$ groß
- ▶ $r_{xy}^{SP} \approx 0 \iff$ kein monotoner Zusammenhang

Extremfälle



$r_{xy}^{SP} = 1$ (oben) und $r_{xy}^{SP} = -1$ (unten)

Bemerkungen:

- Rechentechnische Vereinfachungen:

$$\begin{aligned}\bar{rg}_X &= \frac{1}{n} \sum_{i=1}^n rg(x_i) = \frac{1}{n} \sum_{i=1}^n i = (n+1)/2, \\ \bar{rg}_Y &= \frac{1}{n} \sum_{i=1}^n rg(y_i) = \frac{1}{n} \sum_{i=1}^n i = (n+1)/2.\end{aligned}$$

Rechentechnisch günstige Version von r_{xy}^{SP} :

- Voraussetzung: keine Bindungen
- Daten: (x_i, y_i) , $i = 1, \dots, n$, $x_i \neq x_j$, $y_i \neq y_j$ für alle i, j
- Rangdifferenzen: $d_i = rg(x_i) - rg(y_i)$

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

Monotone Transformationen

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

- g und h beide monoton wachsend oder beide monoton fallend
 $\implies r^{SP}(\tilde{X}, \tilde{Y}) = r^{SP}(X, Y)$
- g, h nicht beide wachsend oder beide fallend
 $\implies r^{SP}(\tilde{X}, \tilde{Y}) = -r^{SP}(X, Y)$

Paarvergleichsmaße: 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 $y_i > y_j$,
also: Rangfolge der X-Werte gleich der Y-Werte.

diskordant, falls $x_i < x_j$ und $y_i > y_j$
oder $x_i > x_j$ und $y_i < y_j$,
also: Rangfolge der X-Werte umgekehrt der Y-Werte.

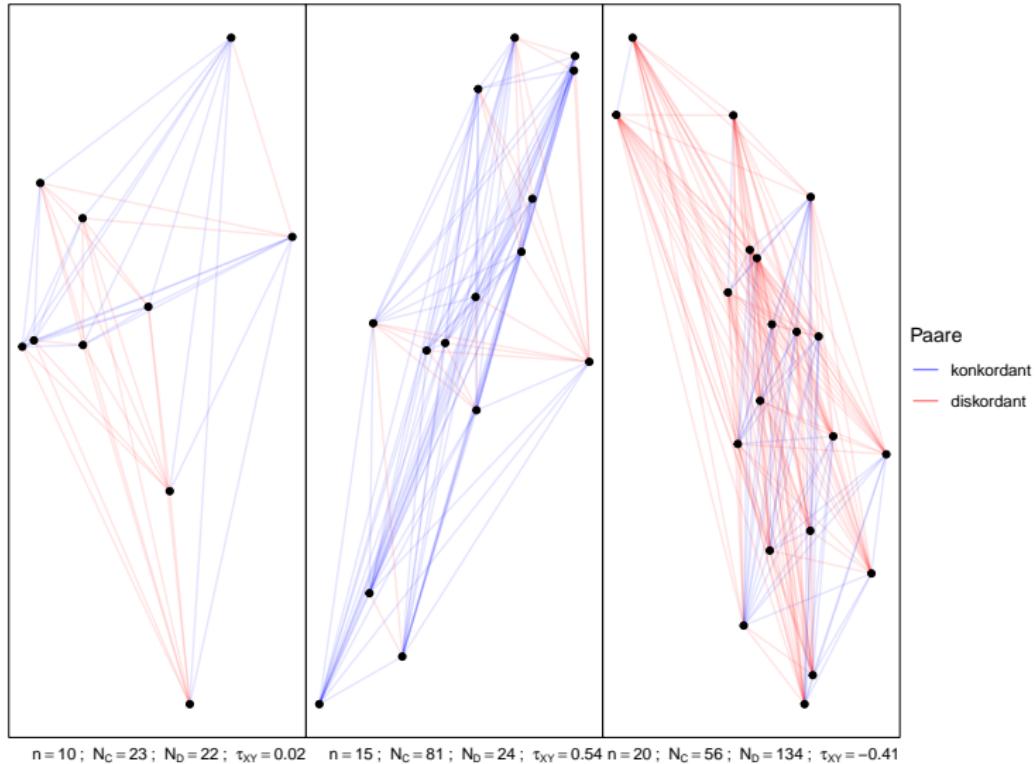
N_C : Anzahl der konkordanten Paare

N_D : Anzahl der diskordanten Paare

Insgesamt $n(n - 1)/2$ mögliche Paare.

$$\text{Kendall's Tau: } \tau_{xy} = \frac{N_C - N_D}{n(n - 1)/2}$$

Kendall's Tau: Veranschaulichung



Varianten

- Goodman & Kruskal γ -Koeffizient ignoriert Paare mit Bindungen:

$$\gamma_{xy} = \frac{N_C - N_D}{N_C + N_D}$$

- Somers' D wird typischerweise verwendet wenn Y binär ist
(\Rightarrow viele Paare mit Bindungen in Y)

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

Gemeinsamkeiten der Paarvergleichsmaße

Für Kendall's τ , Somers' D und Goodman & Kruskal's γ -Koeffizient gilt:

- ihr Wertebereich ist $[-1, 1]$
- sie setzen (mindestens) *ordinale* Variablen X, Y voraus

Paarvergleichsmaßen vs. Spearman's r_{xy}^{SP}

- ▶ r_{xy}^{SP} verwendet Abstände auf der Rang-Skala
- ▶ τ_{xy} und Varianten verwendet alle Paarvergleiche
- ▶ τ_{xy} ist in der Regel betragsmäßig kleiner als r_{xy}^{SP}

Distanzkovarianz & Distanzkorrelation

Modernes Zusammenhangsmaß für (fast) beliebige Zusammenhänge (nicht nur linear/monoton).

Basiert auf Produkten der (zentrierten) Distanzen der Beobachtungen untereinander, nicht nur auf Distanzen zu ihren Mittelwerten.

Distanzkovarianz dS_{xy} :

$$dS_{xy} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n D_{ij}^x D_{ij}^y$$

- ▶ zentrierte Distanzen $D_{ij}^x = d_{ij}^x - (\bar{d}_{i\cdot}^x + \bar{d}_{\cdot j}^x - \bar{d}^x)$ mit
- ▶ $d_{ij}^x = |x_i - x_j|$
- ▶ $\bar{d}_{i\cdot}^x = \frac{1}{n} \sum_{j=1}^n d_{ij}^x$ und $\bar{d}_{\cdot j}^x = \frac{1}{n} \sum_{i=1}^n d_{ij}^x$,
- ▶ $\bar{d}^x = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^x$
- ▶ analog für D_{ij}^y

Distanzkorrelation

Die **Distanzkorrelation** ist definiert als

$$dr_{xy} = \sqrt{\frac{dS_{xy}^2}{\sqrt{dS_{xx}^2 dS_{yy}^2}}}$$

wobei

- ▶ $dS_{xy}^2 = dS_{xy} dS_{xy}$
- ▶ dS_{xx} ist die Distanzkovarianz von X mit sich selbst.

Eigenschaften Distanzkorrelation

- ▶ $0 \leq dr_{xy} \leq 1$ – misst nur Stärke, nicht “Richtung” der Abhängigkeit.
- ▶ $dr_{xy} = 0 \iff X \text{ und } Y \text{ empirisch unabhängig (!)}$
- ▶ $dr_{xy} = 1$ für perfekt lineare Zusammenhänge (bei Verwendung der euklidischen Distanz)
- ▶ **Sehr** allgemein anwendbar:
 - ▶ beliebige Distanzmaße verwendbar als d_{ij}^x, d_{ij}^y
 - ▶ \implies auch für alle *multivariaten* oder *nicht-numerischen* Daten X, Y (z.B. Bilder, Audiosignale, Gensequenzen,) benutzbar, für die man Distanzen definieren kann.

(Székely, Rizzo, Bakirov; 2007)

Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

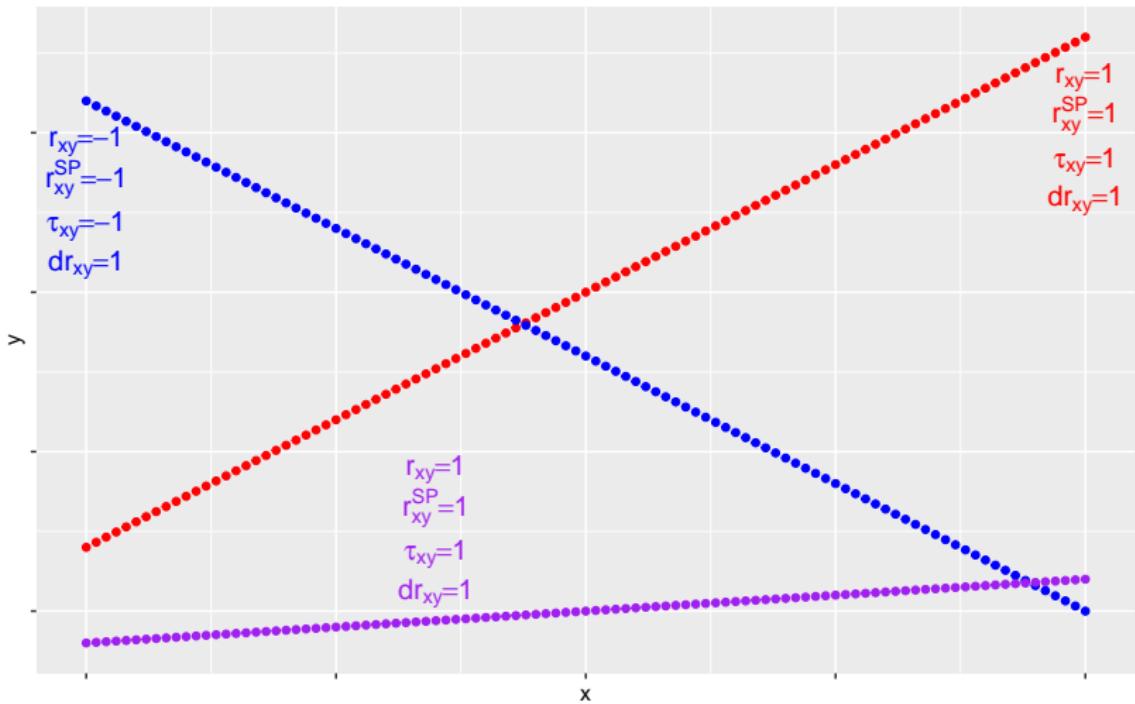
Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

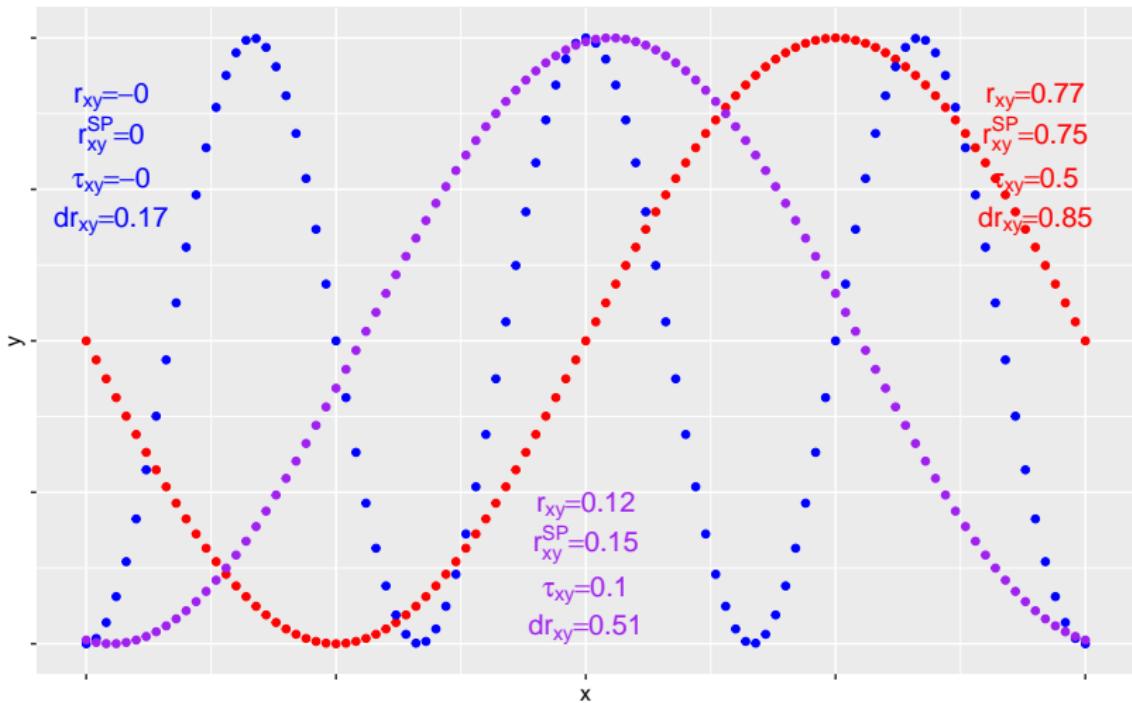
Beispiele: Deterministische Zusammenhänge

Lineare (unverrauschte) Funktion, $Y = a + b \cdot X$, 101 equidistante Stützstellen im Intervall [-1,1]



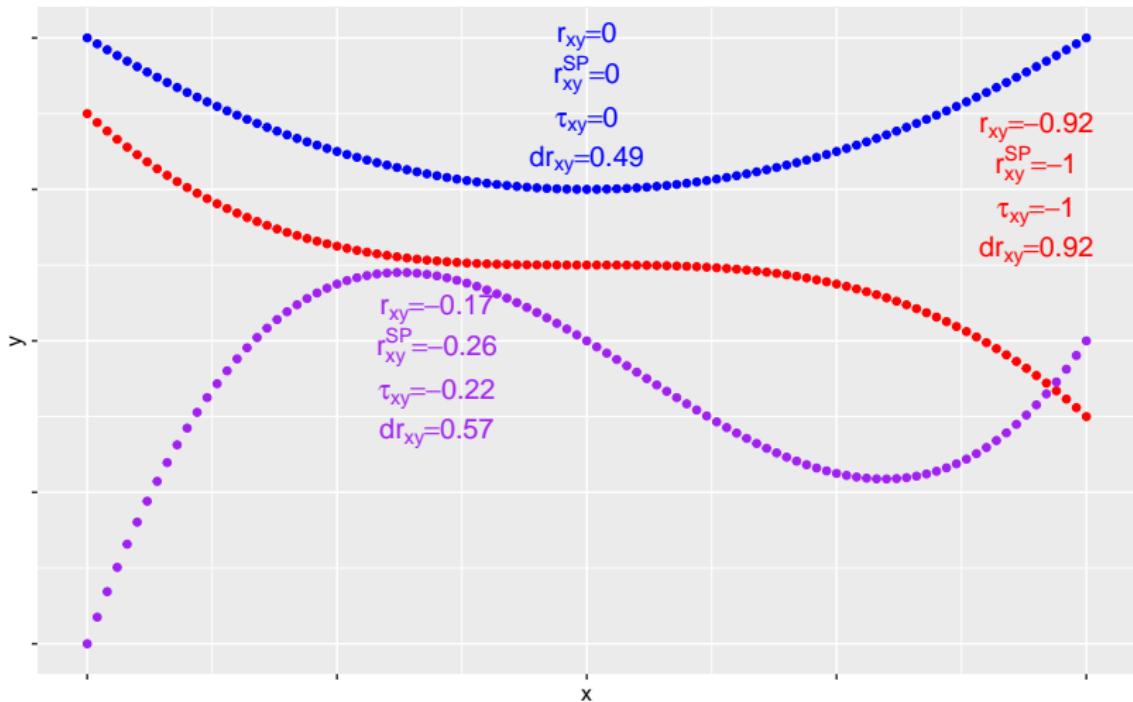
Beispiele: Deterministische Zusammenhänge

Periodische (unverrauschte) Funktionen, 101 equidistante Stützstellen im Intervall $[-1, 1]$



Beispiele: Deterministische Zusammenhänge

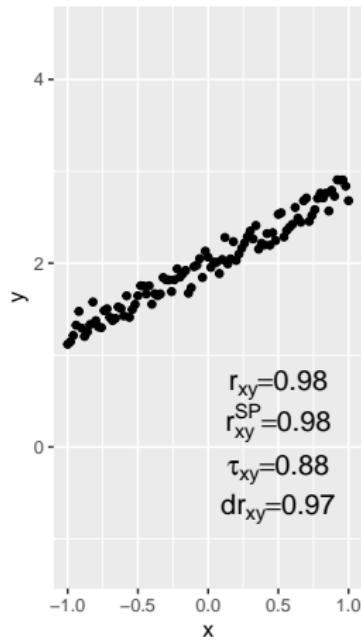
Quadratische und kubische (unverrauschte) Funktionen, 101 equidistante Stützstellen im Intervall [-1, 1]



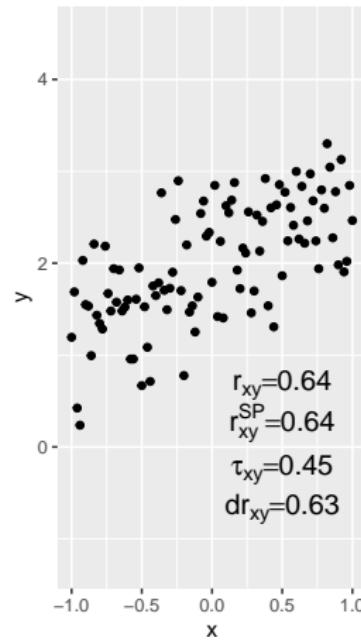
Beispiele: Exakte & verrauschte Zusammenhänge

Lineare, verrauschte Funktionen $Y = 2 + 0.8X + U$ mit zufälligem Fehler U .
101 equidistante Stützstellen im Intervall $[-1,1]$

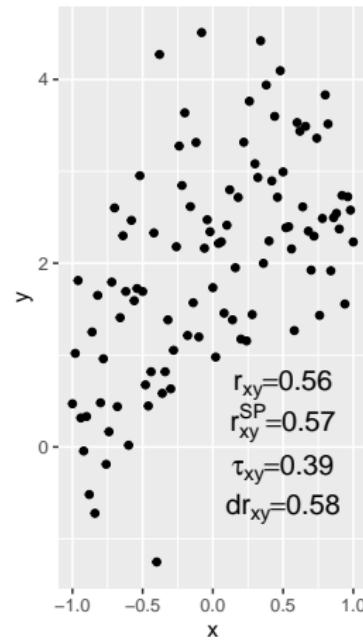
$$S_U = 0.1$$



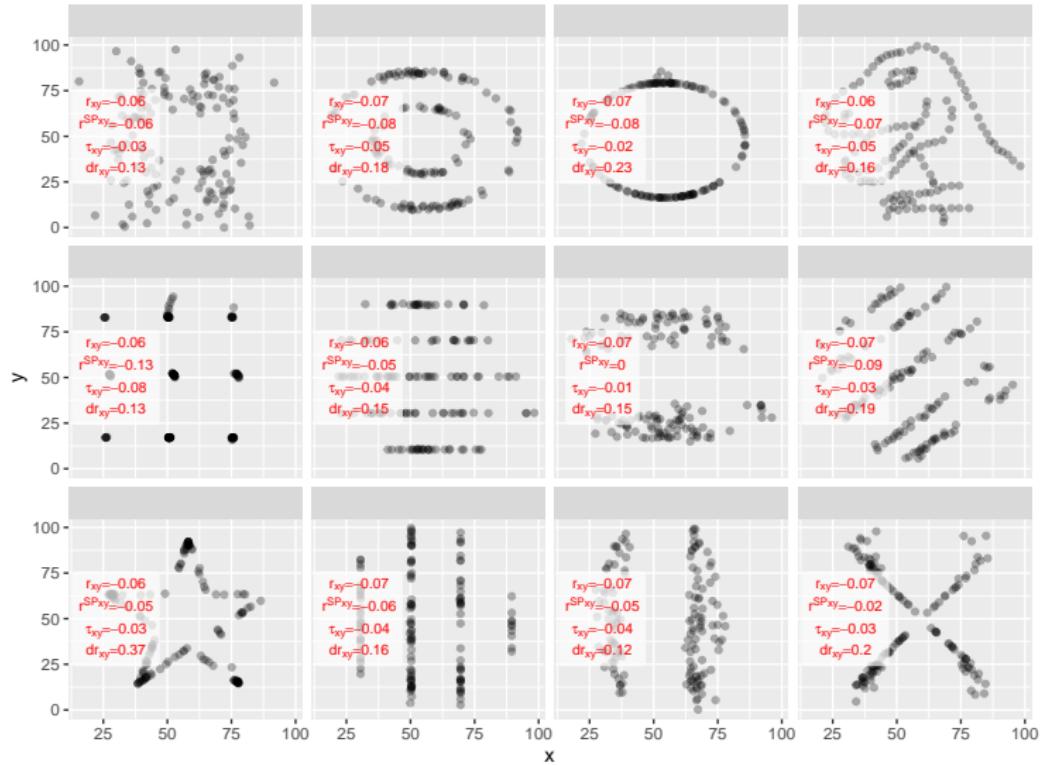
$$S_U = 0.5$$



$$S_U = 1$$



Noch mehr Beispiele



Zusammenhangsmaße für metrische Merkmale

Kovarianz und Korrelation beobachteter Merkmale

Darstellung multivariater Zusammenhänge

Kovarianz und Korrelation von Zufallsvariablen

Alternative Zusammenhangsmaße

Beispiele: Zusammenhänge metrischer Variablen

Zusammenhangsmaße für dichotome und ordinale/metrische Merkmale

Dichotome und ordinale/metrische Merkmale

Wichtiger Spezialfall:

Dichotomes Merkmal Y und ein mindestens ordinalskaliertes X.

Beispiele:

- ▶ Medizin: Diagnostische Tests
X: Biomarker (Stoffkonzentration in Blut o.ä.) oder diagnostischer Score
Y: krank vs. gesund; genesen vs. verstorben; etc...
- ▶ Marketing:
X: Kundeneigenschaften (z.B. bisher erzielter Umsatz mit diesem Kunden)
Y: Kaufentscheidung ja/nein; Vertragsverlängerung ja/nein; etc...
- ▶ Kredit-Score o.ä.
Y: Schufa-Eintrag (Kreditausfall) binnen 1.5 Jahre oder nicht

Sensitivität und Spezifität

Setting:

- $Y \in \{0, 1\}$ dichotom (Zielgröße)
- X mindestens ordinalskaliert (Einflussgröße)

$Y = 1 \iff \text{"positiver" Fall (z.B. "krank", "Kreditausfall", "Kündigung", ...)}$

$Y = 0 \iff \text{"negativer" Fall (z.B. "gesund", "Rückzahlung", "Verlängerung", ...)}$

Ziel:

Vorhersage/Diagnose \hat{y}_i nur auf Basis von x_i und Schwellenwert c :

$$\hat{y}_i = 1 \iff x_i \geq c$$

Fragestellung:

Wie kann ein geeigneter Schwellenwert c bestimmt werden?

Wie gut oder schlecht eignet sich Merkmal X insgesamt zur Einschätzung von Y ?

Sensitivität und Spezifität

	$y_i = 0$	$y_i = 1$	
Vorhersage $\hat{y}_i = 0$	“wahr negativ”	“falsch negativ”	# negative Vorhersagen
Vorhersage $\hat{y}_i = 1$	“falsch positiv”	“wahr positiv”	# positive Vorhersagen
	# negative	# positive	

Dilemma:

- ▶ Je größer c
 - ▶ desto weniger “positive” Vorhersagen insgesamt
 - ▶ tendenziell: desto weniger falsch positive aber auch weniger wahr positive
- ▶ Je kleiner c
 - ▶ desto weniger “negative” Vorhersagen insgesamt
 - ▶ tendenziell: desto weniger falsch negative aber auch weniger wahr negative

Sensitivität und Spezifität

Anteil wahr positiver Prognosen für echt positive = Sensitivität

$$TPR(c) = f(\hat{Y} = 1 | Y = 1) = f(X \geq c | Y = 1)$$

“Welcher Anteil der Kranken wurde entdeckt?”: “*Empfindlichkeit*” des Tests
true positive rate TPR

Anteil falsch positiver Prognosen für echt negative

$$FPR(c) = f(\hat{Y} = 1 | Y = 0) = f(X \geq c | Y = 0)$$

“Welcher Anteil der Gesunden wurde falsch diagnostiziert?”

false positive rate FPR

Anteil wahr negativer Prognosen für echt negative = Spezifität

$$TNR(c) = f(\hat{Y} = 0 | Y = 0) = 1 - f(X \geq c | Y = 0) = 1 - FPR(c)$$

“Welcher Anteil der Gesunden wurde korrekt diagnostiziert?”

true negative rate TNR

ROC-Kurve

Die ROC-Kurve

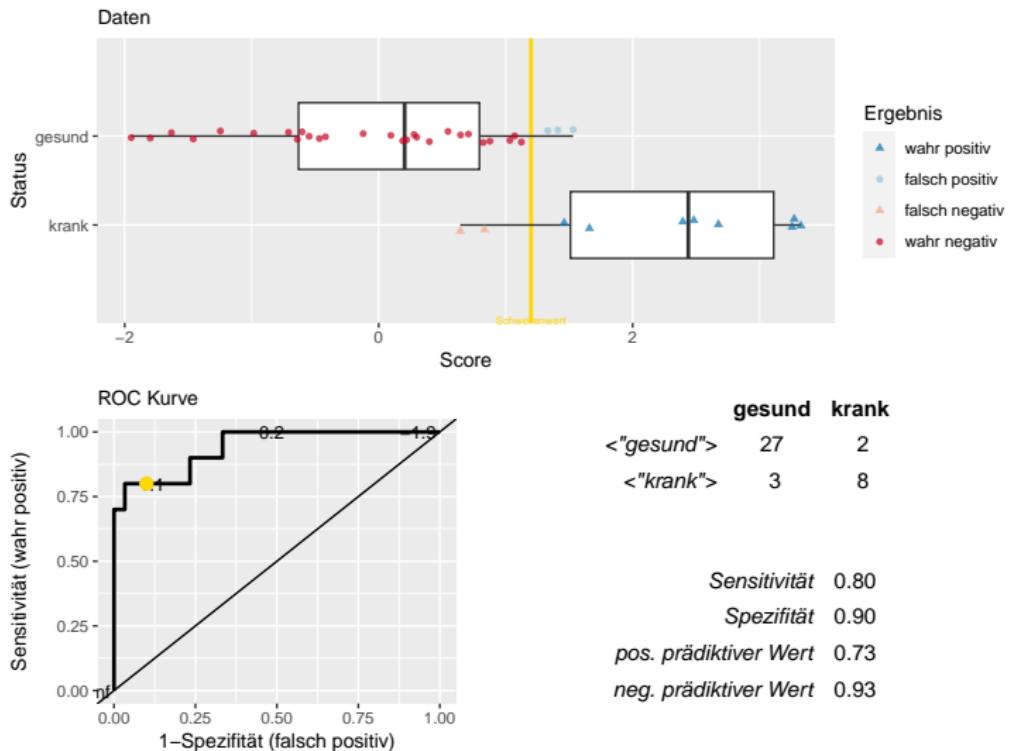
- ▶ verbindet die Punkte $(FPR(c), TPR(c))$, also (1-Spezifität, Sensitivität) bzw ("falsch positiv"-Rate, "wahr positiv"-Rate)
- ▶ für alle möglichen Schwellenwerte $c \in [x_{(1)}, x_{(n)}]$

Sie zeigt die Zuverlässigkeit der Vorhersagen für alle möglichen Schwellenwerte c an.

Es gilt:

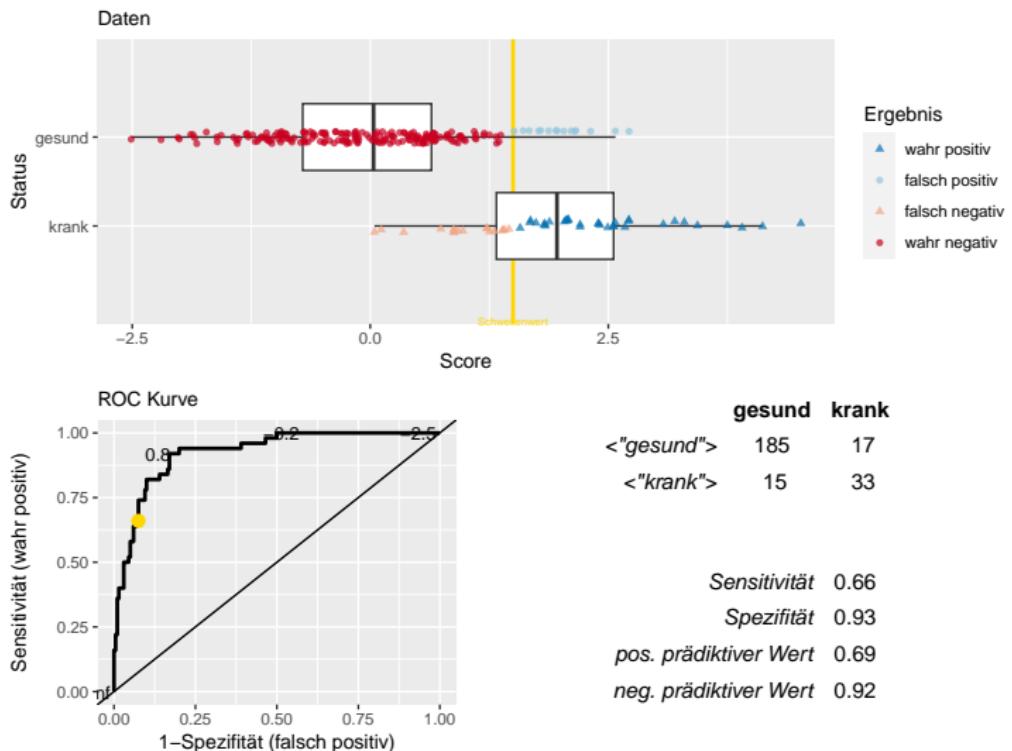
- ▶ Für $c < x_{(1)}$: Test immer "positiv", also
 $\hat{y}_i = 1 \forall i \implies (FPR(c), TPR(c)) = (1, 1)$
- ▶ Für $c > x_{(n)}$: Test immer "negativ", also
 $\hat{y}_i = 0 \forall i \implies (FPR(c), TPR(c)) = (0, 0)$

Beispiel 1: Mittelmäßig starker Zusammenhang



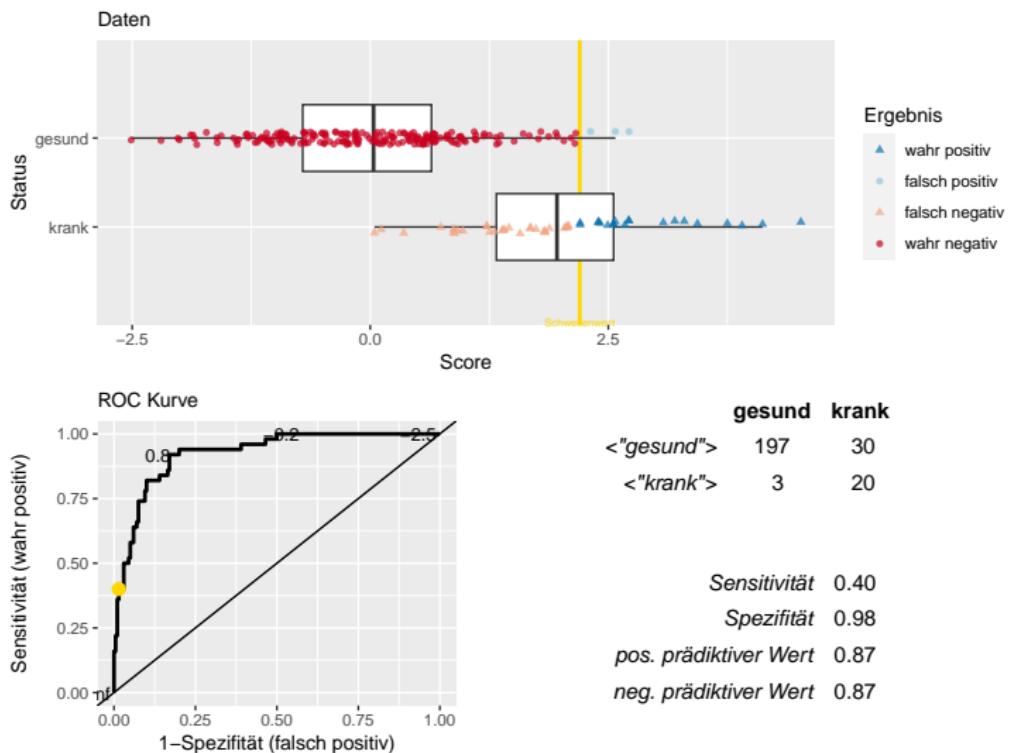
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 2: Mittelmäßig starker Zusammenhang



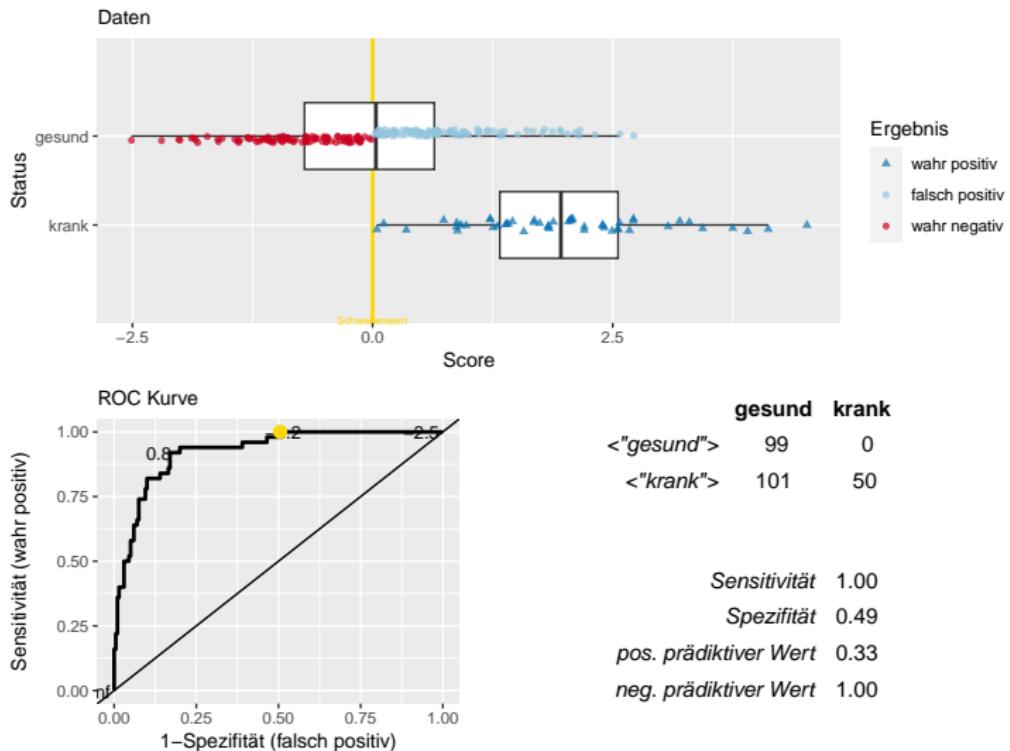
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 2: Mittelmäßig starker Zusammenhang



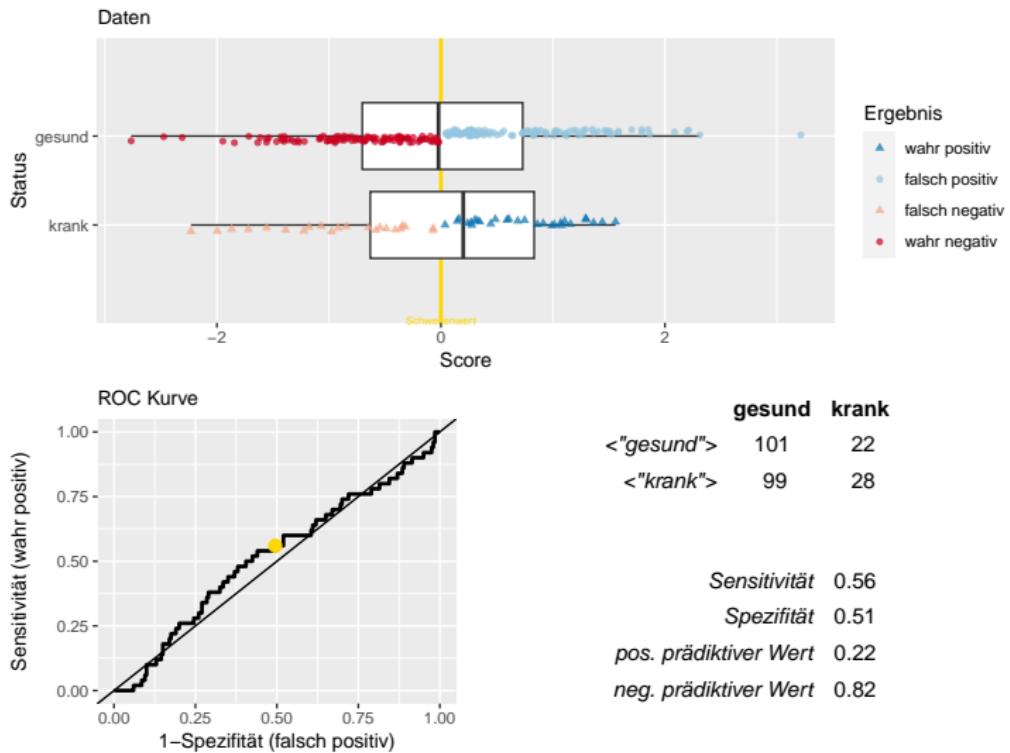
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 2: Mittelmäßig starker Zusammenhang



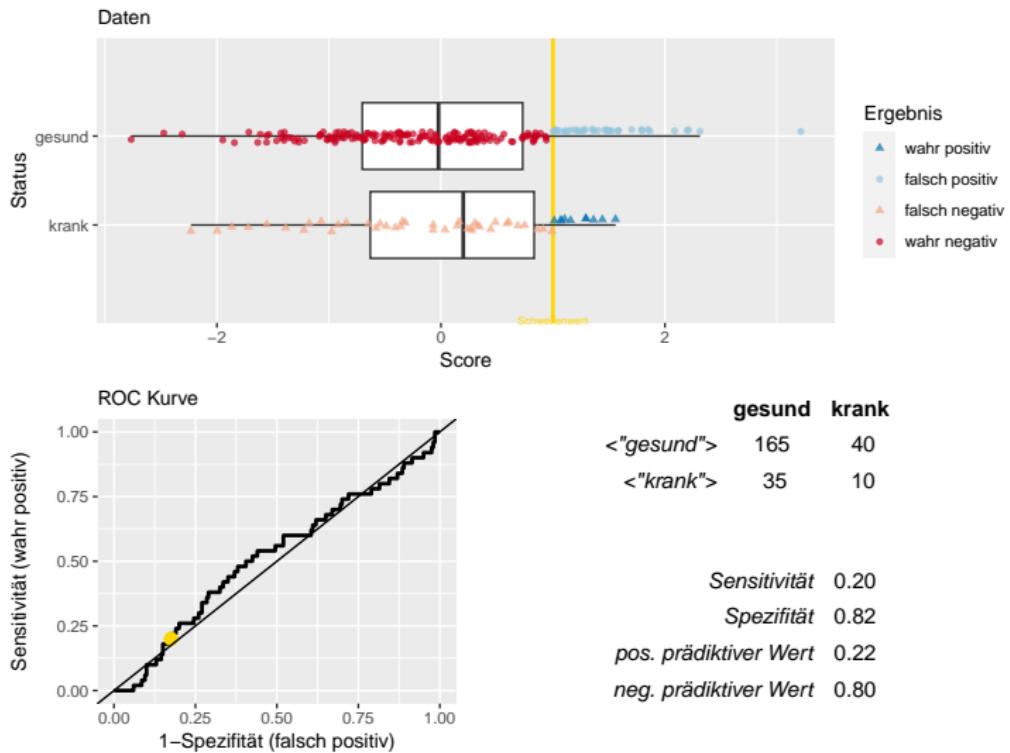
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 3: Kein Zusammenhang



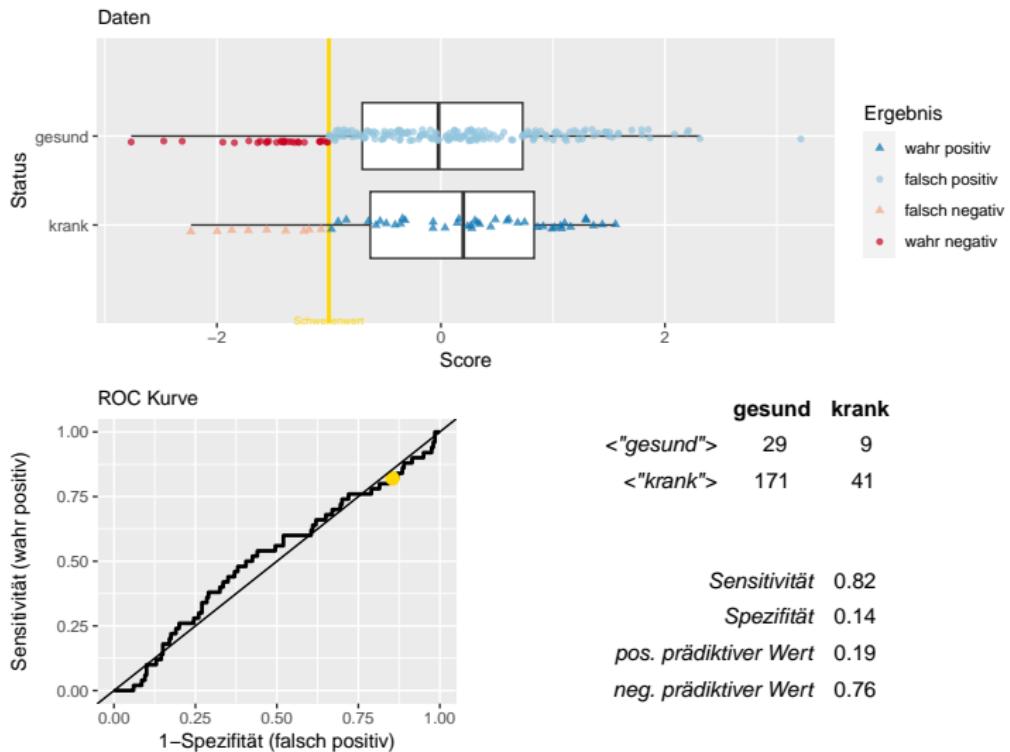
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 3: Kein Zusammenhang



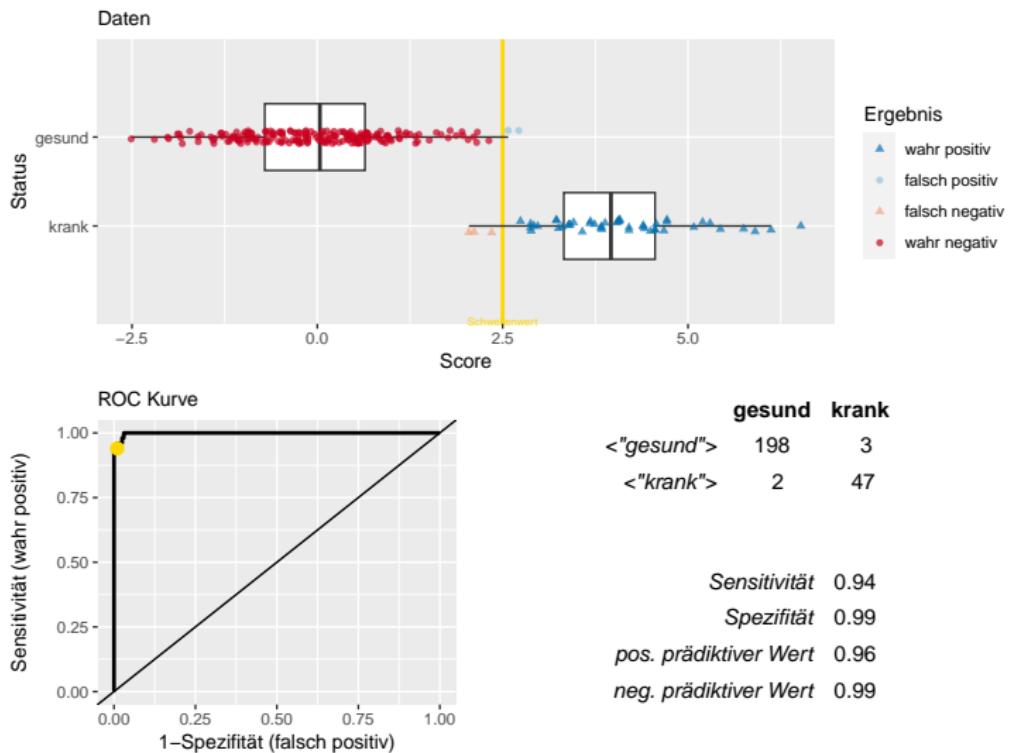
s.a. shinyapps.io/sensi-spezi-roc

Beispiel 3: Kein Zusammenhang



s.a. shinyapps.io/sensi-spezi-roc

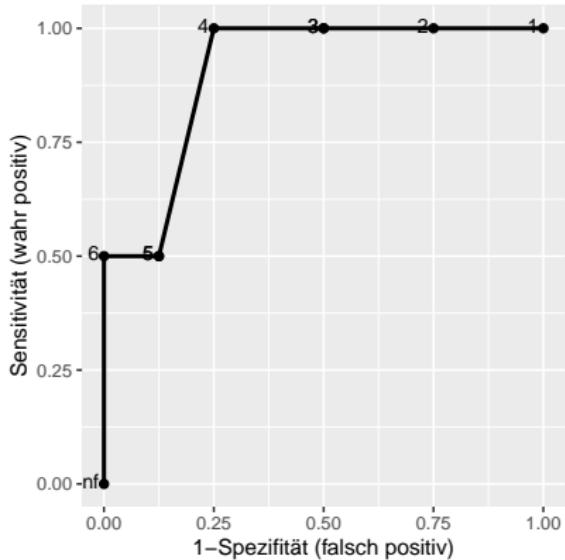
Beispiel 4: Sehr starker Zusammenhang



s.a. shinyapps.io/sensi-spezi-roc

Beispiel für ROC-Kurve mit Bindung

X	Y
1	1
2	1
3	2
4	2
5	3
6	3
7	4
8	4
9	5
10	6



Maß zur Bewertung der Kurve: AUC

Das AUC entspricht der Fläche unter der ROC-Kurve.

$$AUC := \frac{N_C + N_E/2}{N}$$

- ▶ Anzahl konkordante Paare
 $N_C = |\{(i, j) : (x_i > x_j \wedge y_i > y_j) \vee (x_i < x_j \wedge y_i < y_j)\}|$
- ▶ Anzahl Paare mit Bindungen in X
 $N_E = |\{(i, j) : x_i = x_j \wedge y_i \neq y_j\}|$
- ▶ Anzahl aller Paare mit unterschiedlichem Y
 $N = |\{(i, j) : y_i \neq y_j\}| = h_Y(0)h_Y(1)$

Interpretation AUC

- ▶ perfekte Trennbarkeit der Y-Gruppen durch den Scorewert X würde einen AUC-Wert von 1 ergeben
- ▶ Unabhängigkeit zwischen X und Y einen AUC-Wert von ca. 0.5.

⇒ AUC ist ein Maß dafür wie gut man Y auf Basis einer einfachen Schwellenwertregel für X vorhersagen kann.

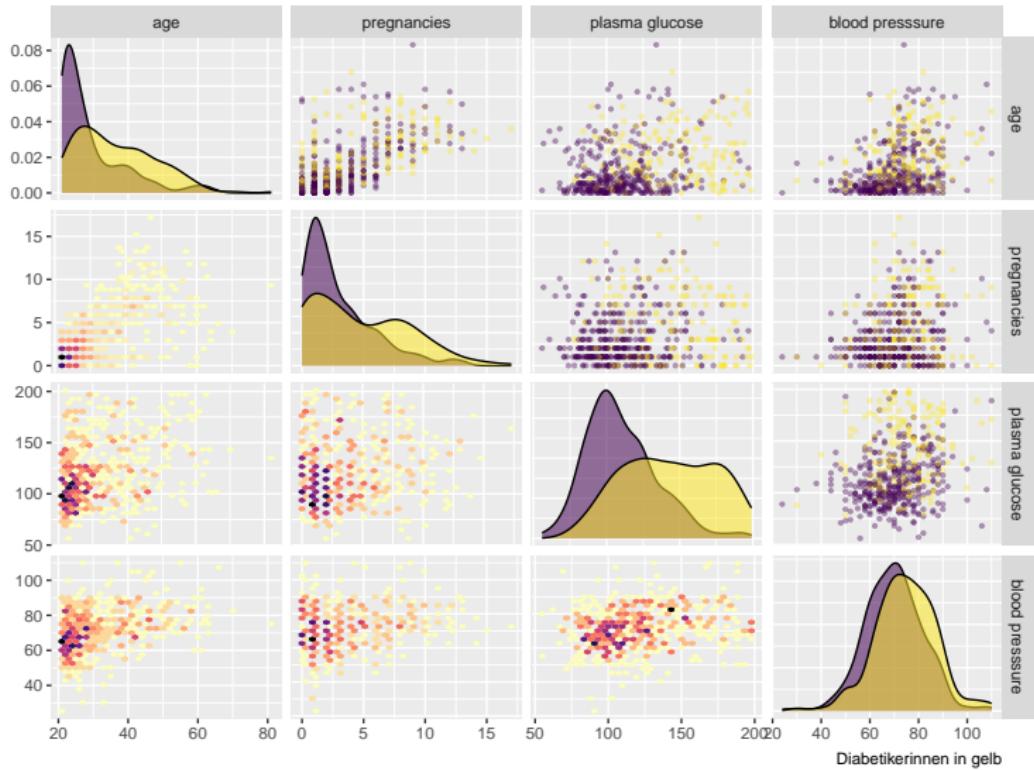
AUC entspricht (in etwa) der **relativen Häufigkeit mit der, in einem beliebigen Paar von Untersuchungseinheiten mit unterschiedlichem Y, die Beobachtung mit $Y = 1$ den höheren X-Wert hat.**

Beispiel: Pima Indian Diabetes-Daten

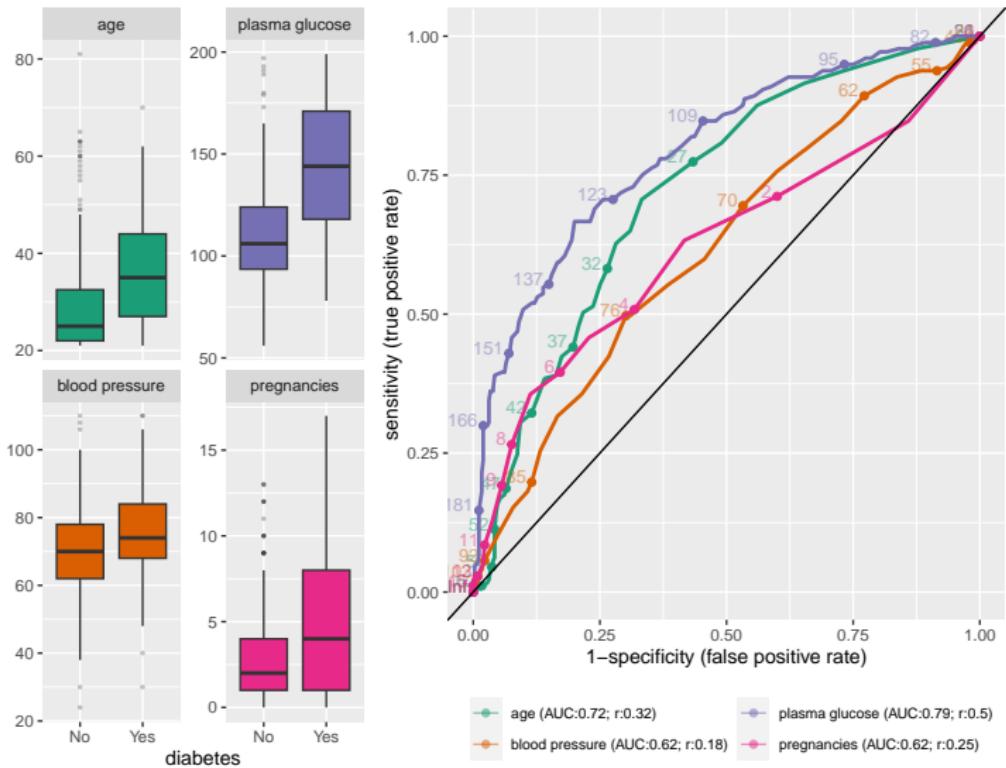
- ▶ 532 Frauen über 21
- ▶ vom Volk der Pima bei Phoenix, AZ
- ▶ sehr hohe Prävalenz von *Diabetes mellitus*: 33%
- ▶ viele zusätzlich erhobene Stoffwechselparameter (Blutplasmaglukose, Blutdruck, ...), Alter, Anzahl Schwangerschaften, etc
- ▶ was sind Risikofaktoren bzw was ist prognostisch verwendbar für Diabetes?

Datenquelle: US National Institute of Diabetes and Digestive and Kidney Diseases

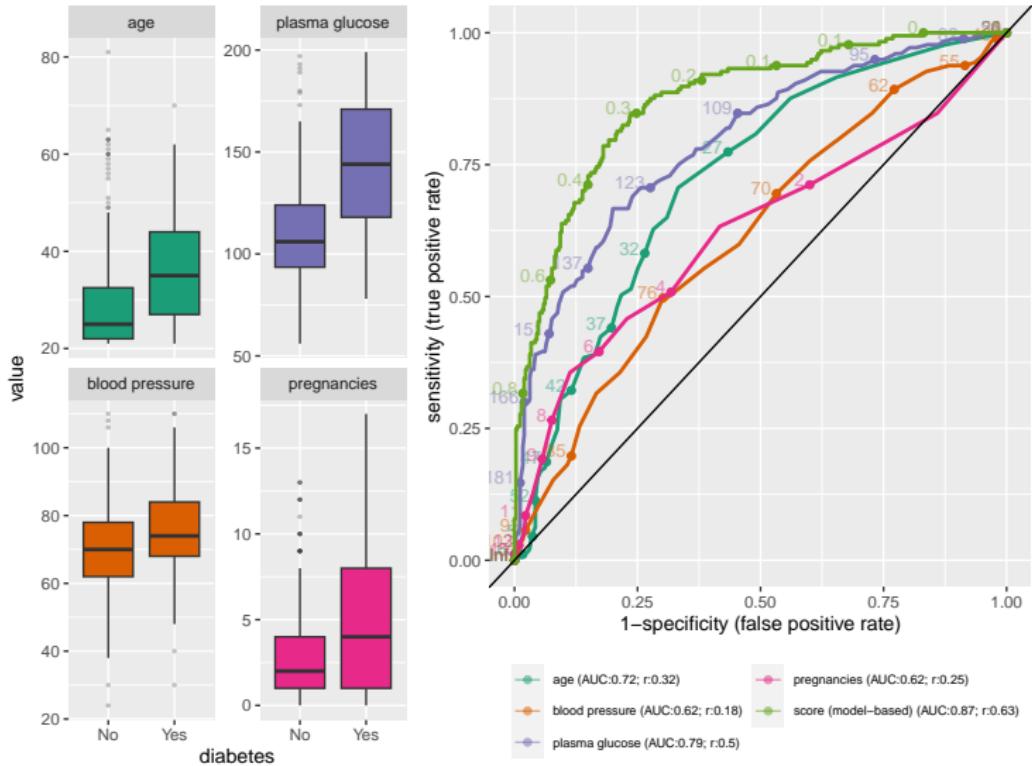
Beispiel: Pima Indian Diabetes-Daten



Beispiel: Pima Indian Diabetes-Daten



Beispiel: Pima Indian Diabetes-Daten



Kritik AUC

AUC basierend auf Sensitivität $f(\hat{Y} = 1 | Y = 1)$: "Welcher Anteil der Kranken wurde korrekt diagnostiziert?";

Spezifität $f(\hat{Y} = 0 | Y = 0)$: "Welcher Anteil der Gesunden wurde korrekt diagnostiziert?"

AUC behandelt Spezifität & Sensitivität gleichwertig –
oft aber dramatisch (!!) unterschiedlich wichtige Konsequenzen für
"falschen Alarm ausgelöst" (falsch positiv)
oder "keine Therapie eingeleitet da Krankheit nicht entdeckt", "Bauteil nicht gewartet/ersetzt da Defekt nicht gefunden", etc. (falsch negativ)

⇒ auch sehr wichtig:

- ▶ **positiv prädiktiver Wert (ppV)** des Diagnoseverfahrens: $f(Y = 1 | \hat{Y} = 1)$
Wie groß ist der Anteil echter Alarne an den ausgelösten Alarmen?
- ▶ **negativ prädiktiver Wert (npV)** des Diagnoseverfahrens: $f(Y = 0 | \hat{Y} = 0)$
Wie viele der als "gesund" Diagnostizierten sind tatsächlich gesund?

Positiv/negativ prädiktiver Wert: ppV & npV

Bsp: Nicht-invasive Pränataldiagnostik (NIPT) für Trisomie-21

- ▶ Prävalenz Trisomie-21 bei 22-jährigen Müttern:
ca. 8 von 10 000 Kindern (0.008%)
- ▶ NIPT: Sensitivität 99.2%, Spezifität 99.9%

Erwartete Häufigkeiten dementsprechend bei 1 000 000 getesteten 22-jährigen Schwangeren:

	Kind krank	Kind gesund	
NIPT: "Kind krank"	794	999	1793
NIPT: "Kind gesund"	6	998 201	998207
	800	999 200	1 000 000

$$npV := \frac{\text{"wahr negativ"}}{\text{"wahr negativ"} + \text{"falsch negativ"}} = 998201 / 998207 = 0.99994$$

$$ppV := \frac{\text{"wahr positiv"}}{\text{"wahr positiv"} + \text{"falsch positiv"}} = 794 / 1793 \approx 0.443$$

Also: Negative NIPT Diagnosen sind nahezu sicher korrekt, aber **nur 44% der positiven NIPT Diagnosen sind zutreffend.**

Zahlen aus Gießelmann, Kathrin: "Nichtinvasive Pränataltests: Risiko für Fehlinterpretationen." Dtsch Arztebl 2020; 117(7): A-320

But wait, there's more.....

Babylonische terminologische Zustände:

		True condition			
Total population	Condition positive	Condition negative	Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$	
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$	F ₁ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$		

Quelle: Wikipedia

Korrelation & Kausalität

Kausale & assoziative Strukturen

(Schein)Assoziation über Drittvariablen

(Schein)Assoziation durch Aggregation

(Schein)Assoziation durch Stichprobenauswahl

Korrelation & Kausalität

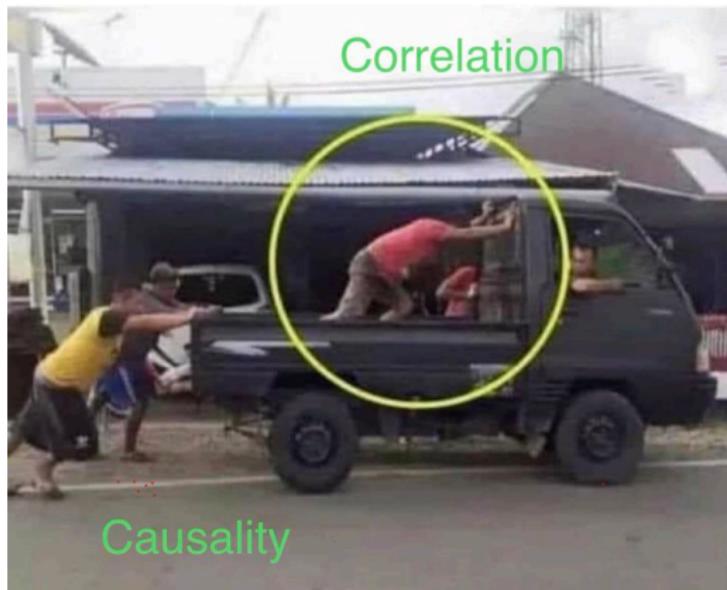
Kausale & assoziative Strukturen

(Schein)Assoziation über Drittvariablen

(Schein)Assoziation durch Aggregation

(Schein)Assoziation durch Stichprobenauswahl

Kausale & assoziative Strukturen

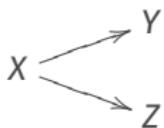


Kausale & assoziative Strukturen

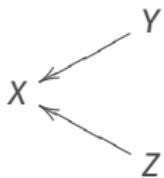
$A \longrightarrow B$: "A ist Ursache von B"

$A \sim\sim B$: "A und B sind assoziiert"

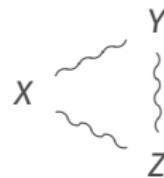
"Fork"



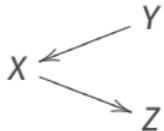
"Collider"



führen alle zu



"Pipe"



Kausale & assoziative Strukturen

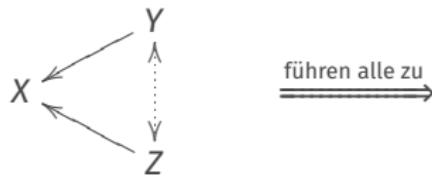
$A \longrightarrow B$: "A ist Ursache von B"

$A \sim\sim B$: "A und B sind assoziiert (nicht unabhängig)"

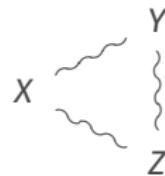
Still a "Fork"



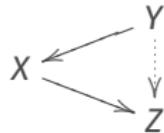
Still a "Collider"



führen alle zu



Still a "Pipe"



Kausale & assoziative Strukturen

- ▶ sehr unterschiedliche kausale Strukturen ergeben identische assoziative Struktur
- ▶ große Vorsicht bei (kausaler) Interpretation beobachteter Korrelationen / Assoziationen geboten
- ▶ "Kausalität" hier stochastisch/probabilistisch gemeint, (meistens) nicht deterministisch.

Korrelation & Kausalität

Kausale & assoziative Strukturen

(Schein)Assoziation über Drittvariablen

(Schein)Assoziation durch Aggregation

(Schein)Assoziation durch Stichprobenauswahl

Confounding

- ▶ Gemeinsame Ursache X für interessierende Variablen Y, Z
(Kausale Struktur: *Fork*)
- ▶ erzeugt oft marginale Abhängigkeiten zwischen nicht kausal verbundenen Y und Z :
“Scheinkorrelation”, *spurious correlation*
- ▶ kann auch evtl. vorhandene (bedingte) Assoziationen zwischen Y und Z gegeben X abschwächen oder sogar umkehren:
Simpson's "Paradox"

Confounding: Schein-Assoziationen

Drittvariablen erzeugen oft Assoziationen ohne kausale Entsprechung:
Beispiele:

Storchenpopulation

}

Geburtenrate

Ertrunkene/Monat

}

Eiscremeabsatz/Monat

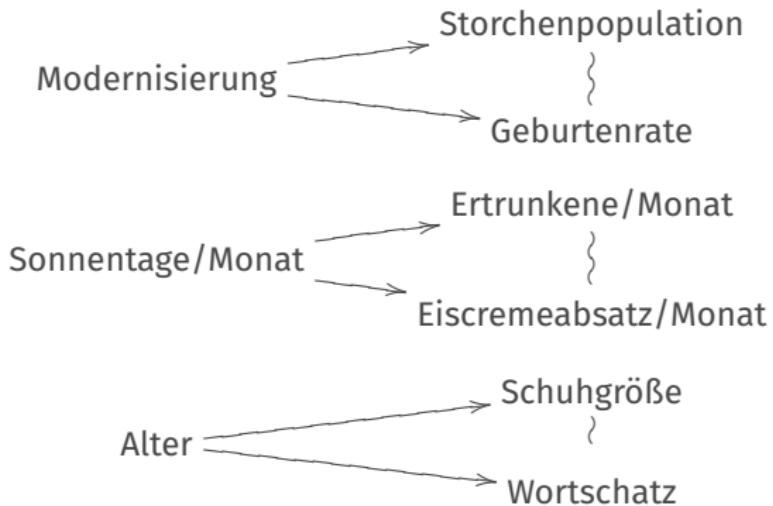
Schuhgröße

}

Wortschatz

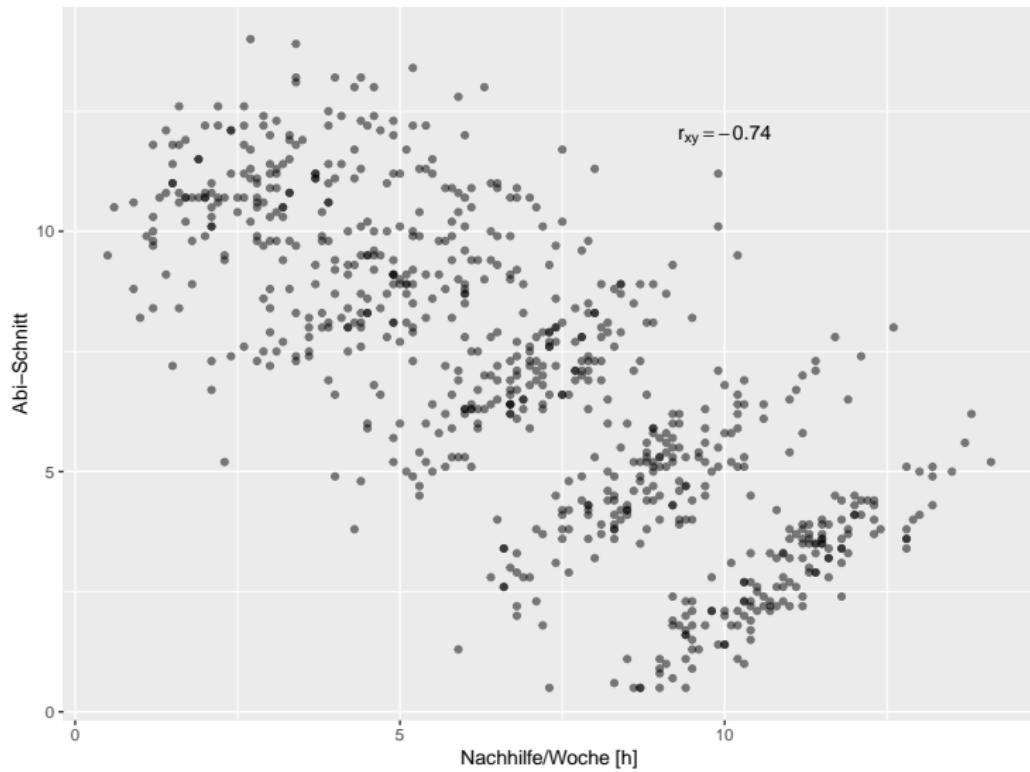
Confounding: Schein-Assoziationen

Drittvariablen erzeugen oft Assoziationen ohne kausale Entsprechung:
Beispiele:



Confounding: Noten & Nachhilfe

Zusammenhang Abschlußnote - Nachhilfestunden/Woche (fiktiv):



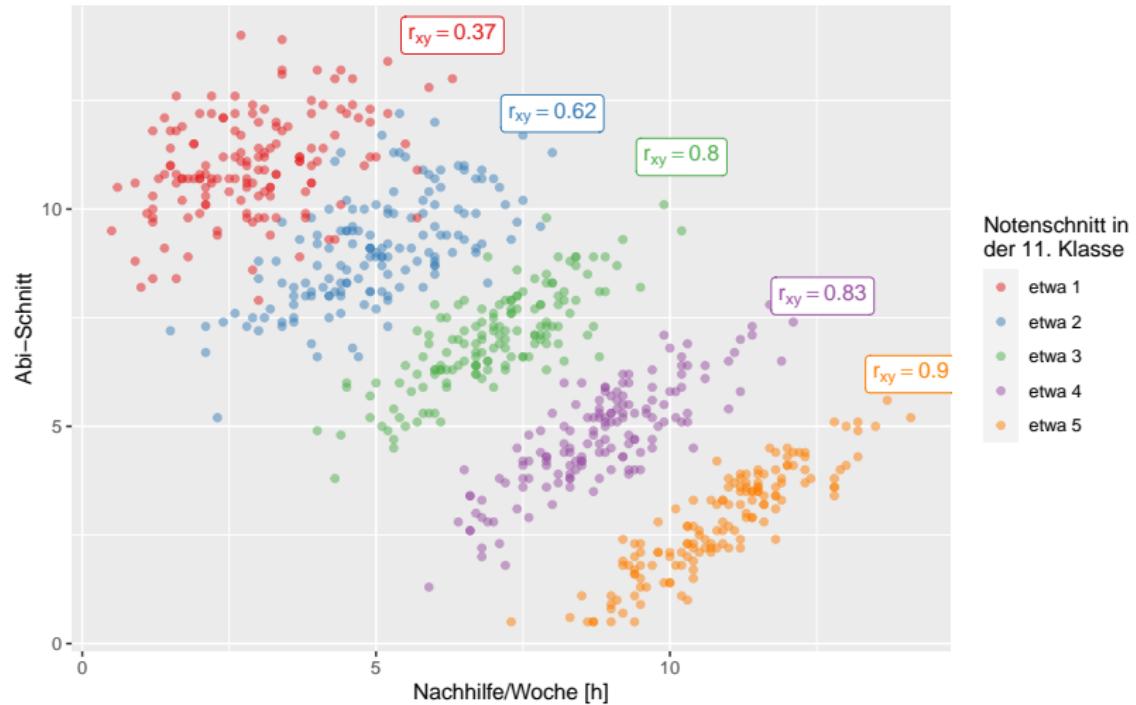
Confounding: Noten & Nachhilfe

Zusammenhang Abschlußnote - Nachhilfestunden/Woche (fiktiv):



Confounding: Noten & Nachhilfe

Zusammenhang Abschlußnote - Nachhilfestunden/Woche (fiktiv):



⇒ kausal interpretierbarer Fall von *Simpson's Paradox*

Korrelation & Kausalität

Kausale & assoziative Strukturen

(Schein)Assoziation über Drittvariablen

(Schein)Assoziation durch Aggregation

(Schein)Assoziation durch Stichprobenauswahl

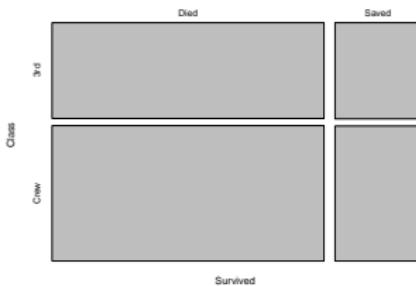
Simpson's Paradox

- ▶ Kein echtes Paradox
- ▶ Beschreibt häufig auftretendes Phänomen:
Durch Bedingen auf Drittvariable(n) können Assoziationen
 - ▶ entstehen (s. Titanic),
 - ▶ verschwinden (s. Berkeley),
 - ▶ oder ihre Richtung ändern. (s. Nachhilfe)
(Autsch.)
- ▶ Synonyme: *omitted variable bias*

Simpson's Paradox: Titanic

Überleben von erwachsenen Titanicpassagieren 3. Klasse und Crew:

	Died	Saved	Rate
3rd	476	151	0.24
Crew	673	212	0.24



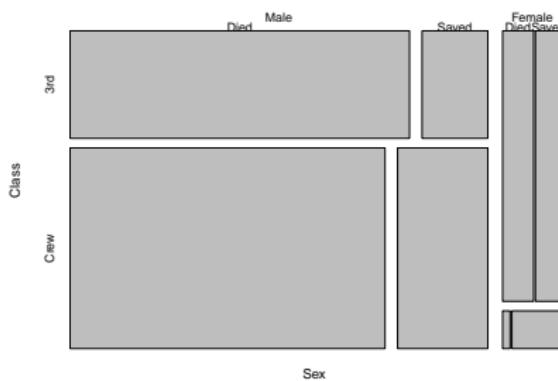
⇒ gleiche Überlebensraten für Passagiere 3. Klasse und Crew (?)

Simpson's Paradox: Titanic

Bedingt auf Geschlecht:

Males:	Died	Saved	Rate
3rd	387	75	0.16
Crew	670	192	0.22

Females:	Died	Saved	Rate
3rd	89	76	0.46
Crew	3	20	0.87



⇒ (deutlich) schlechtere Überlebensraten für Passagiere 3. Klasse, für beide Geschlechter.

Simpson's Paradox: Berkeley Admissions

Aufnahmekquoten der Uni Berkeley im Jahr 1973:

	Admitted	Rejected	Admission Rate
Male	1198	1493	0.45
Female	557	1278	0.30

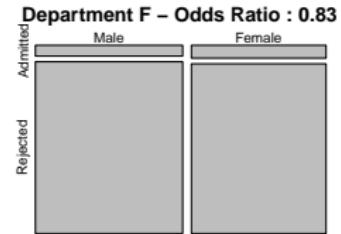
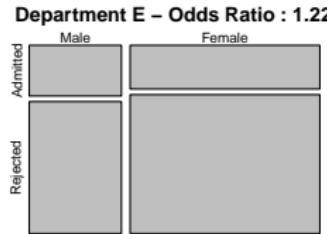
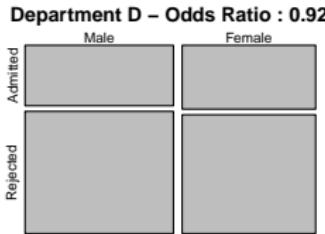
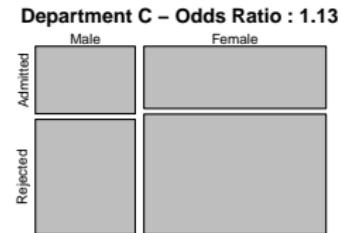
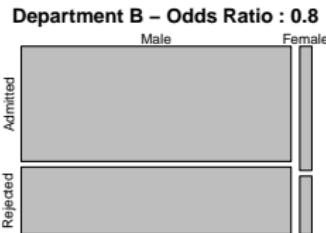
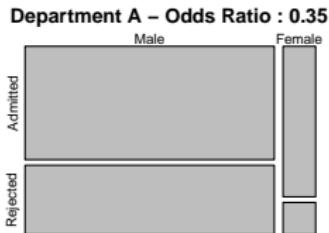


⇒ Benachteiligung von Bewerberinnen ...?

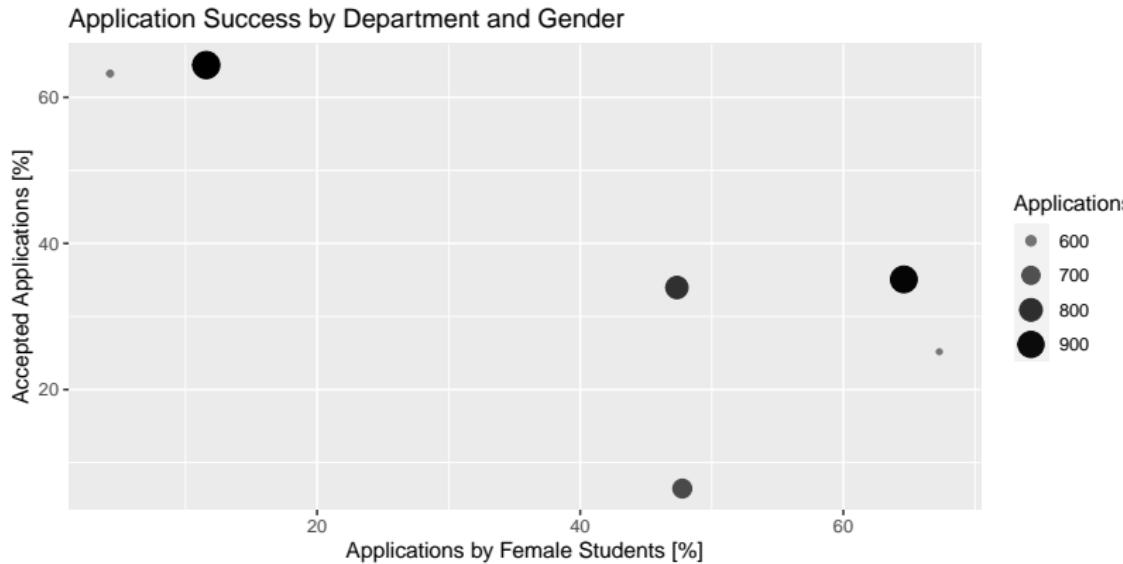
Odds Ratio: $\gamma(\text{Admitted}, \text{Rejected} | \text{Male}, \text{Female}) = 1.83!$

Simpson's Paradox: Berkeley Admissions

Bedingt auf Departments (A-F; absteigend nach Aufnahmequote):



Simpson's Paradox: Berkeley Admissions



Simpson's Paradox: Berkeley Admissions

Kausalität hier eher:



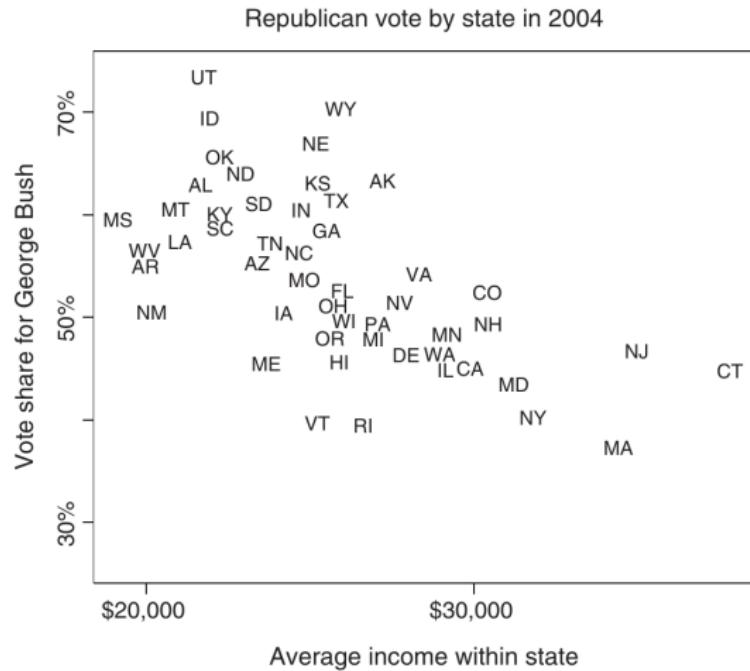
- hier kein “Confounding” mit *Fork*-Struktur im klassischen Sinn
- Stattdessen *Pipe*-Struktur: Durch Bedingen auf *Mediator* “Fach” wird Assoziation zwischen “Geschlecht” und “Aufnahme” komplexer, hier insgesamt schwächer.

Ökologischer Fehlschluss

Allgemeinere Art von “omitted variable bias”:
Unzulässige / falsche Schlüsse von aggregierten Daten auf Individualdaten.

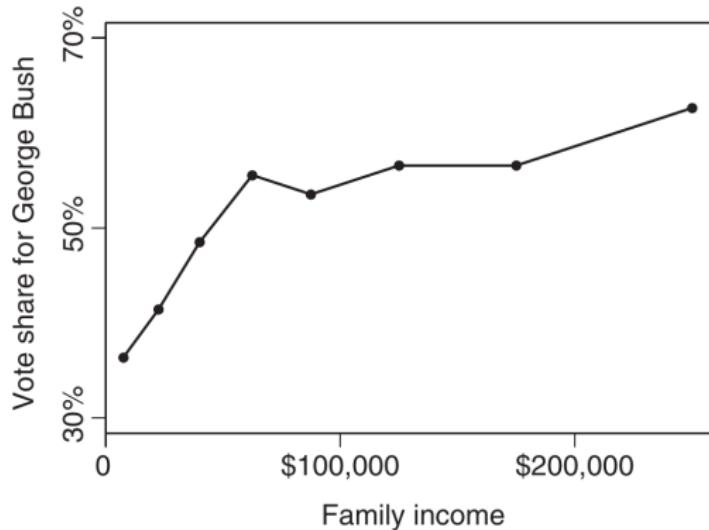
Aus einem bestehenden Zusammenhang auf Aggregatebene folgt nicht zwangsläufig ein entsprechender Zusammenhang auf Individualebene.

Einkommen & Politische Einstellung



Einkommen & Politische Einstellung

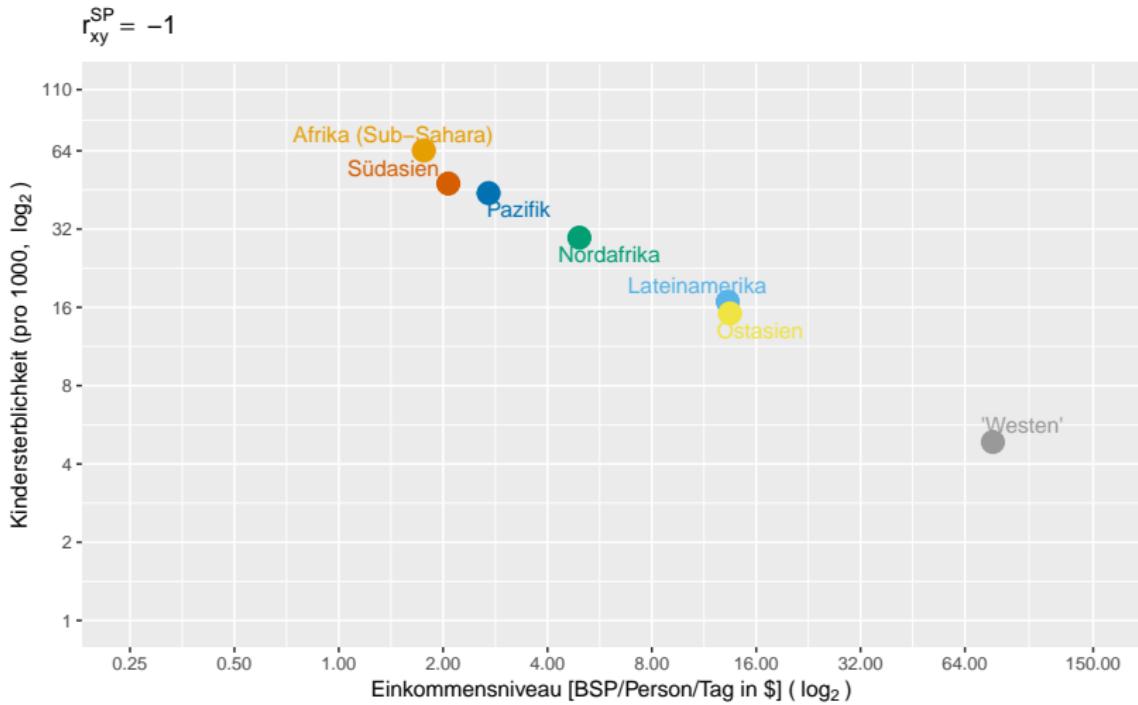
Bush vote in 2004 by income



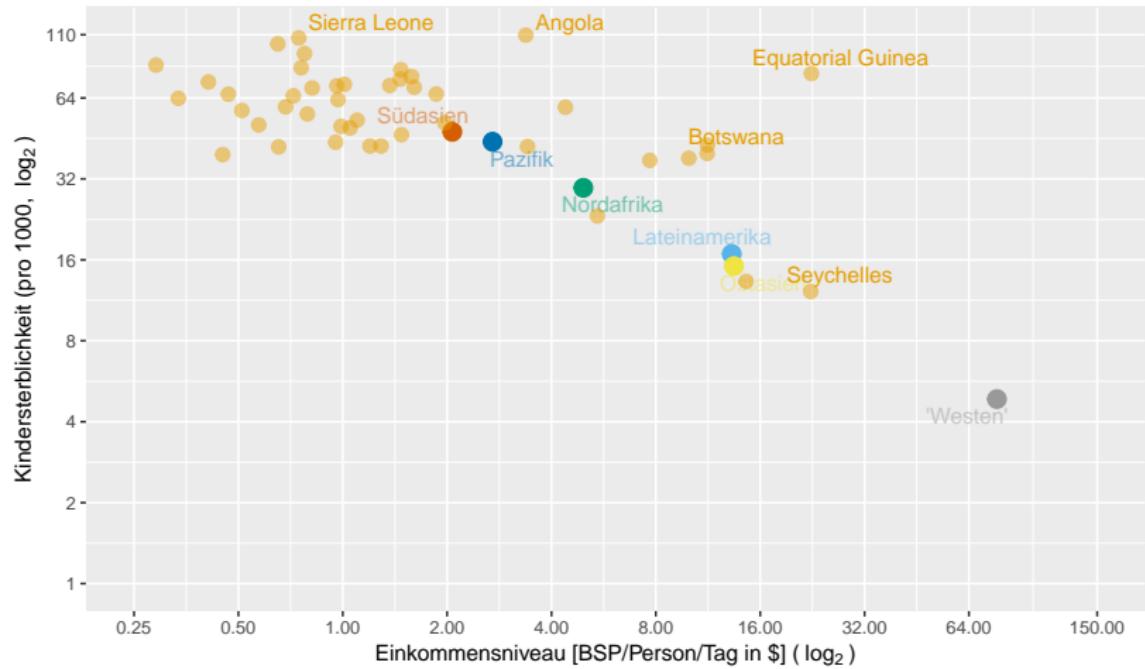
⇒ Aus “Republikaner gewinnen in ärmeren Bundesstaaten” folgt **nicht** “Höhere Zustimmung zu Republikanern bei Menschen mit niedrigerem Einkommen”!

Quelle: Gelman, A. et al. (2007) “Rich State, Poor State, Red State, Blue State: What’s the Matter with Connecticut?” *Quarterly Journal of Political Science* 2: 345-367.

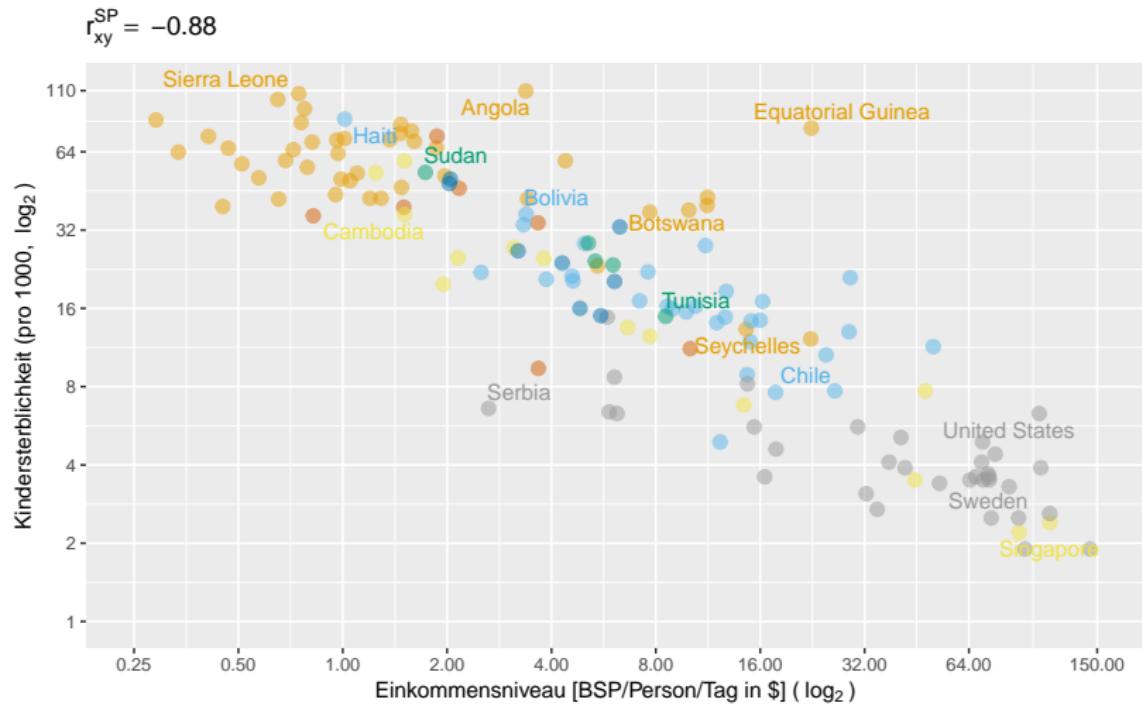
Einkommensniveau & Kindersterblichkeit



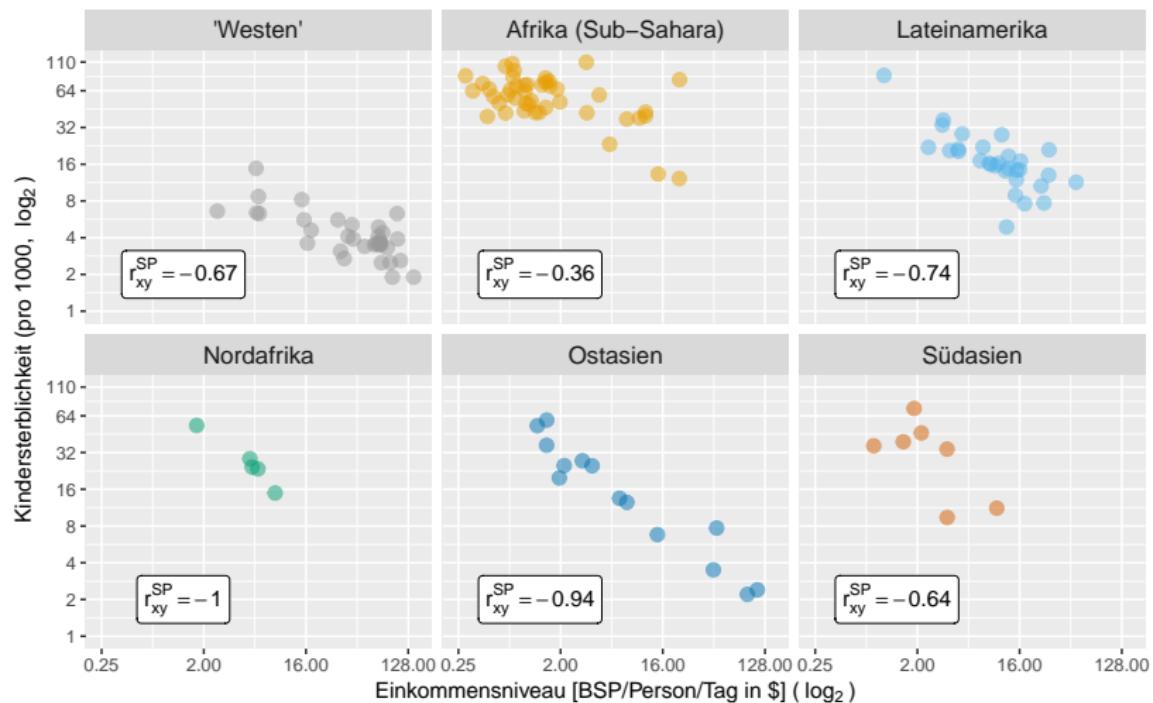
Einkommensniveau & Kindersterblichkeit



Einkommensniveau & Kindersterblichkeit



Einkommensniveau & Kindersterblichkeit



Quelle: WHO/OECD, Stand 2010 (dslabs::gapminder)

Korrelation & Kausalität

Kausale & assoziative Strukturen

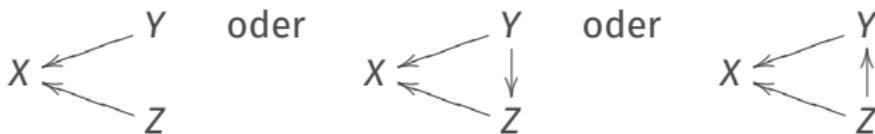
(Schein)Assoziation über Drittvariablen

(Schein)Assoziation durch Aggregation

(Schein)Assoziation durch Stichprobenauswahl

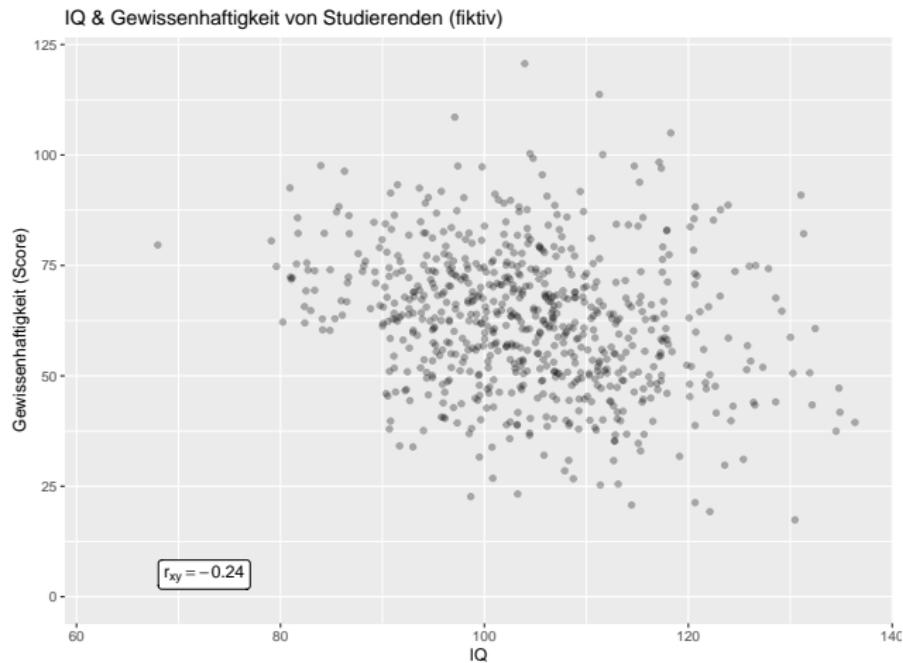
Conditioning on the Collider

- X ist “collider” wenn Y und Z beides Ursachen von X sind:



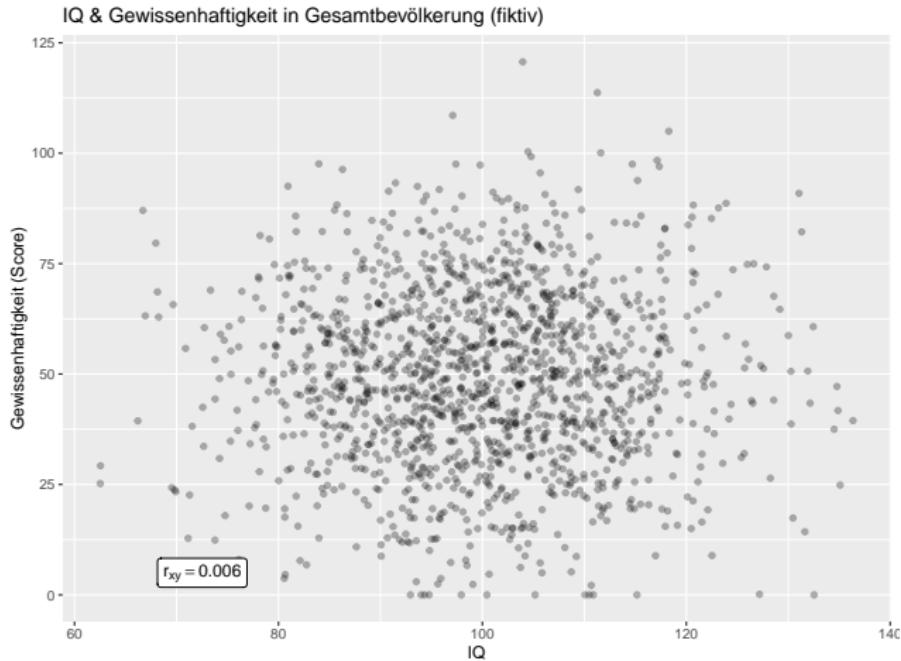
- Durch Bedingen auf X können *spurious correlations* zwischen Y und Z entstehen
- Durch Bedingen auf X können (marginale) Assoziationen zwischen Y und Z ihre Richtung ändern oder verschwinden

Conditioning on the Collider: Beispiel

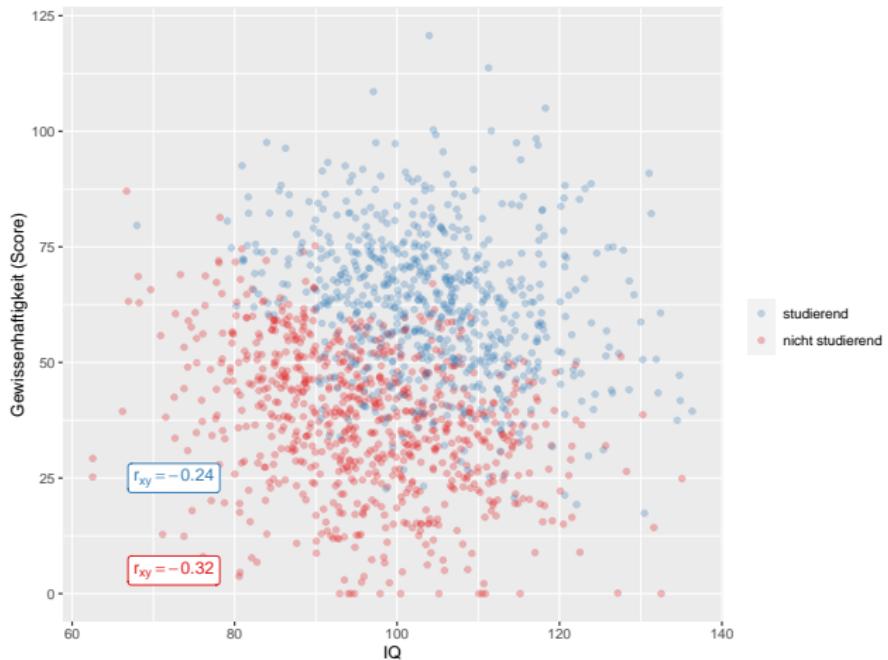


“Schlampige Genies” und “Stumpfe Arbeitsbienen”...?

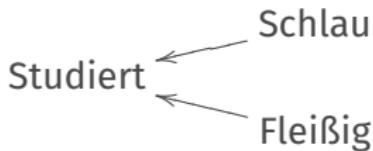
Conditioning on the Collider: Beispiel



Conditioning on the Collider: Beispiel



Conditioning on the Collider: Beispiel



Durch die Bedingung auf "Studiert" entsteht eine Scheinkorrelation zwischen Schläue und Fleiß:

nur diejenigen, die eher schlau oder eher fleißig oder sogar beides sind, können studieren,
dadurch entsteht eine negative Korrelation in der *beobachteten* Teilpopulation der Studierenden.

Anders gesagt: Durch den Auswahlprozess der Stichprobe sind "besonders schlau, nicht so fleißig" und "besonders fleißig, nicht so schlau" in der Stichprobe relativ häufiger als in der Gesamtpopulation. Dadurch entsteht eine negative Korrelation.

Conditioning on the Collider: *endogeneous selection bias*

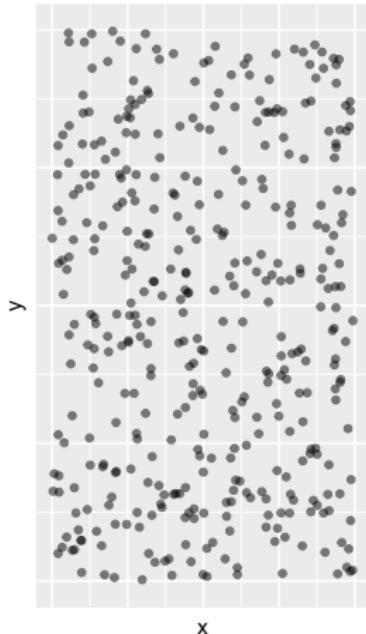
Hier also Spezialfall *endogeneous selection bias*:

Ob eine Untersuchungseinheit in der beobachteten Stichprobe ist oder nicht hängt von einer Variable ab, die von den untersuchten Variablen kausal abhängt.

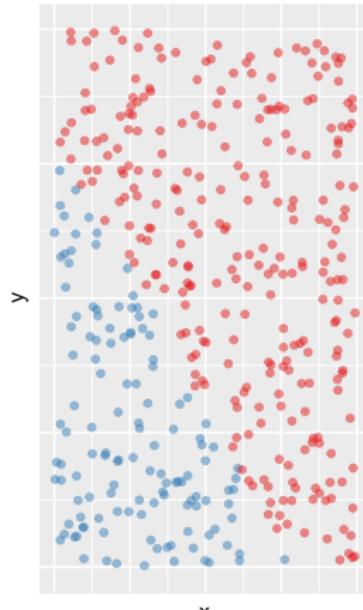
Auch bekannt als *selection-distortion effect*.

Conditioning on the Collider: *endogeneous selection bias*

Gesamtbevölkerung

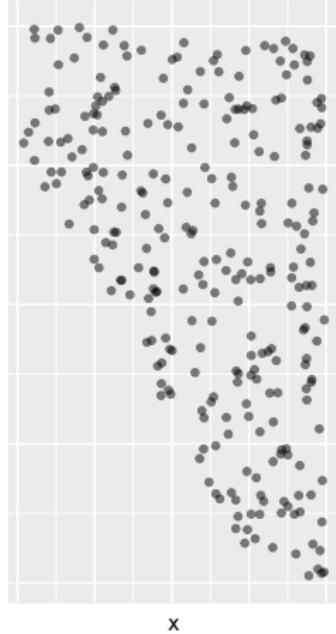


kein Zusammenhang!



beobachtet ● ja ● nein

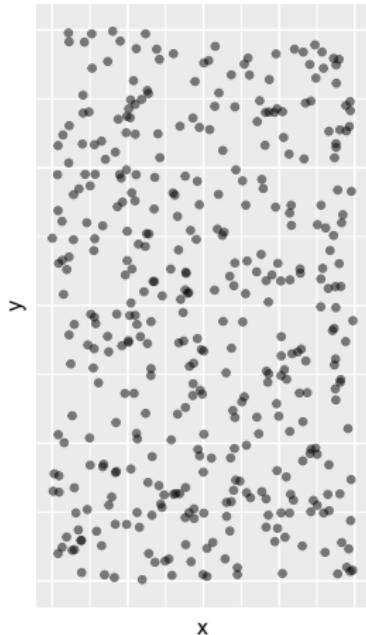
Beobachtet falls $x + y > \text{const}$



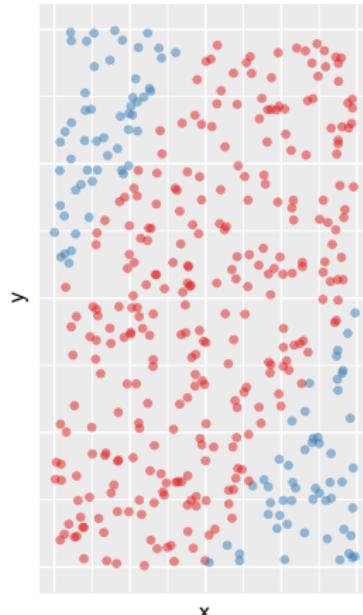
⇒ negative Korrelation

Conditioning on the Collider: *endogeneous selection bias*

Gesamtbevölkerung

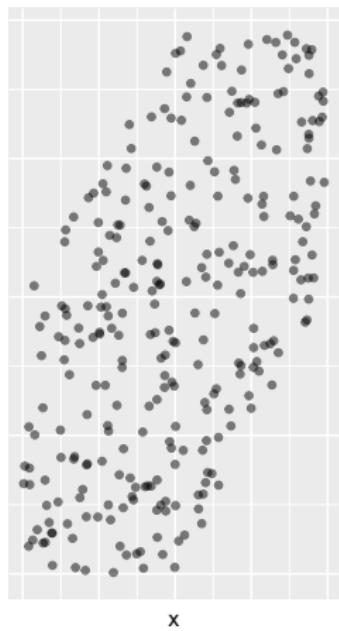


kein Zusammenhang!



beobachtet ● ja ● nein

Beobachtet falls $|x - y| < \text{const}$



⇒ positive Korrelation

Colophon

Material teils basierend auf früheren Vorlesungen von Helmut Küchenhoff, Torsten Hothorn und Leonhard Held.

Berechnungen mit R, gerendert mit pandoc via `{rmarkdown}` und X \exists T \mathbb{E} X.

R-Pakete:

- ▶ Grafiken mit `{ggplot2}`, `{ggrepel}`, `{grid}`, `{gridExtra}`, `{rayshader}`, `{patchwork}`, `{viridisLite}`, `{colorspace}`
- ▶ Datensätze aus `{dslabs}`, `{socviz}` & `{datasaurus}`
- ▶ Daten-Wrangling mit `{dplyr}` & `{tidyverse}`
- ▶ Interaktive Quizzes mit `{exams}`

Fonts: Fira Sans, Fira Math, XITS Math