

u^b

b
UNIVERSITÄT
BERN

Quantitative Methoden in der Geographie

Herbstsemester 2024

Jörg Franke

Viele Fragen können nur mit Hilfe von quantitativen Informationen und Statistik beantworten werden

Hat man einen finanziellen Vorteil, wenn man einen Abschluss an einer Hochschule macht?

Mit einem Hochschulabschluss verdient man:

- mehr oder
- durchschnittlich 10000 CHF mehr pro Jahr

Welche Aussage findet ihr hilfreicher?

u^b

b
UNIVERSITÄT
BERN

Stichwort «Data Literacy»

Was heisst z.B. ein Impfstoff hat eine Wirksamkeit von 95%?

Stichwort «Data Literacy»

Was heisst z.B. ein Impfstoff hat eine Wirksamkeit von 95%?

Wie viele Personen erkrankten mit Placebo, wie viele mit dem Impfstoff?

	Mit Scheinimpfstoff (Placebo)	Mit Impfstoff Comirnaty	Prozentuale Verringerung des Erkrankungsrisikos
Wie viele aller Teilnehmenden erkrankten an COVID-19?	93 von 10.000	5 von 10.000	Ca. 95 %

dies entspricht nur noch ~1 von 20 erwarteten Infektionen mit Impfung.

Stichwort «Data Literacy»

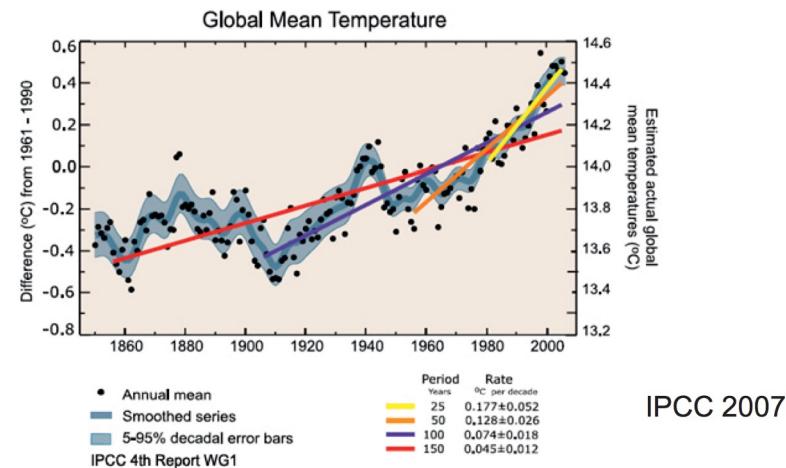
Was heisst z.B. ein Impfstoff hat eine Wirksamkeit von 95%?

Wie viele Personen erkrankten mit Placebo, wie viele mit dem Impfstoff?

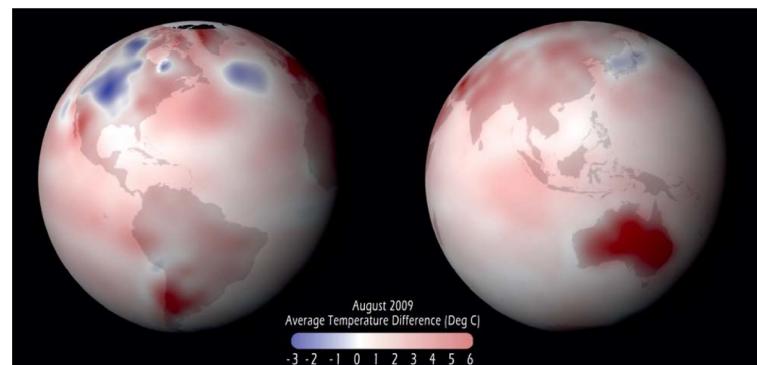
	Mit Scheinimpfstoff (Placebo)	Mit Impfstoff Comirnaty	Prozentuale Verringerung des Erkrankungsrisikos
Wie viele aller Teilnehmenden erkrankten an COVID-19?	93 von 10.000	5 von 10.000	Ca. 95 %

Was heisst eine 20% Regenwahrscheinlichkeit in der Wettervorhersage und wie wird diese berechnet?

Wissenschaft produziert quantitative Ergebnisse

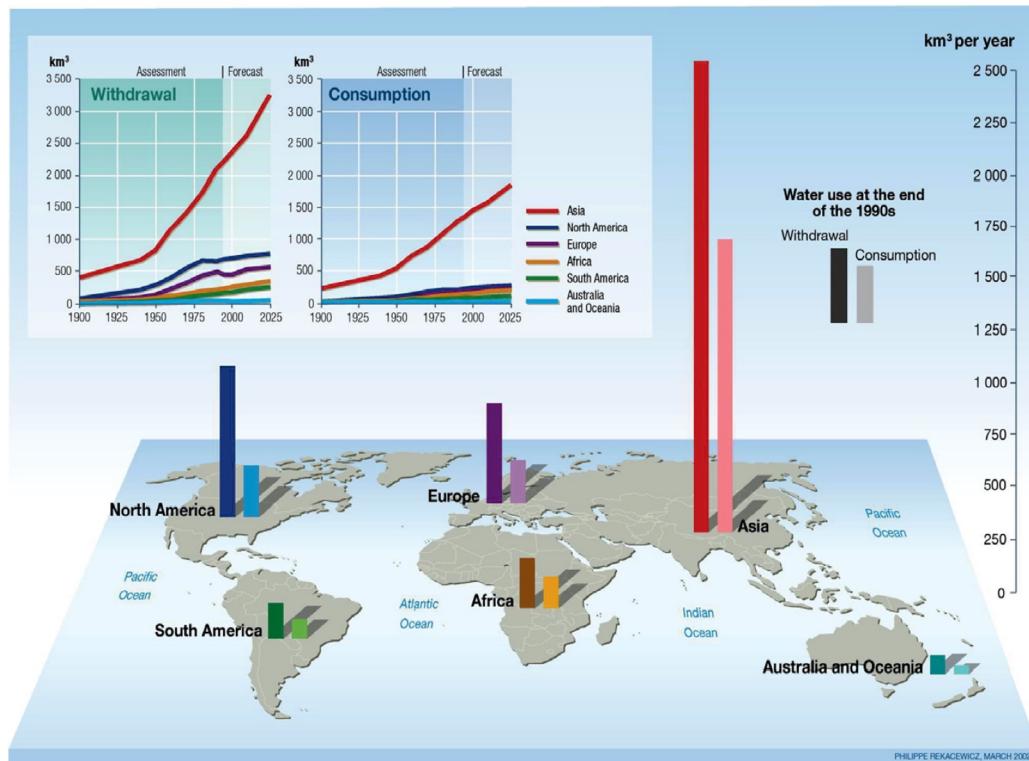


zeitliche Dimension



räumliche Dimension

Wissenskommunikation beruht darauf



Source: Igor A. Shiklomanov, State Hydrological Institute (SHI, St. Petersburg) and United Nations Educational, Scientific and Cultural Organisation (UNESCO, Paris), 1999; *World Resources 2000-2001, People and Ecosystems: The Fraying Web of Life*, World Resources Institute (WRI), Washington DC, 2000; Paul Harrison and Fred Pearce, *AAAS Atlas of Population 2001*, American Association for the Advancement of Science, University of California Press, Berkeley.

Was hat das mit Geographie zu tun?

Kompetenter Umgang mit Daten gehört zu physischer und Humangeographie, insbesondere:

- Solide Anwendung statistischer Methoden
- Interpretation der Resultate
- Darstellung von Daten (das ...graph in Geographie)

Qualitativ vs. Quantitativ

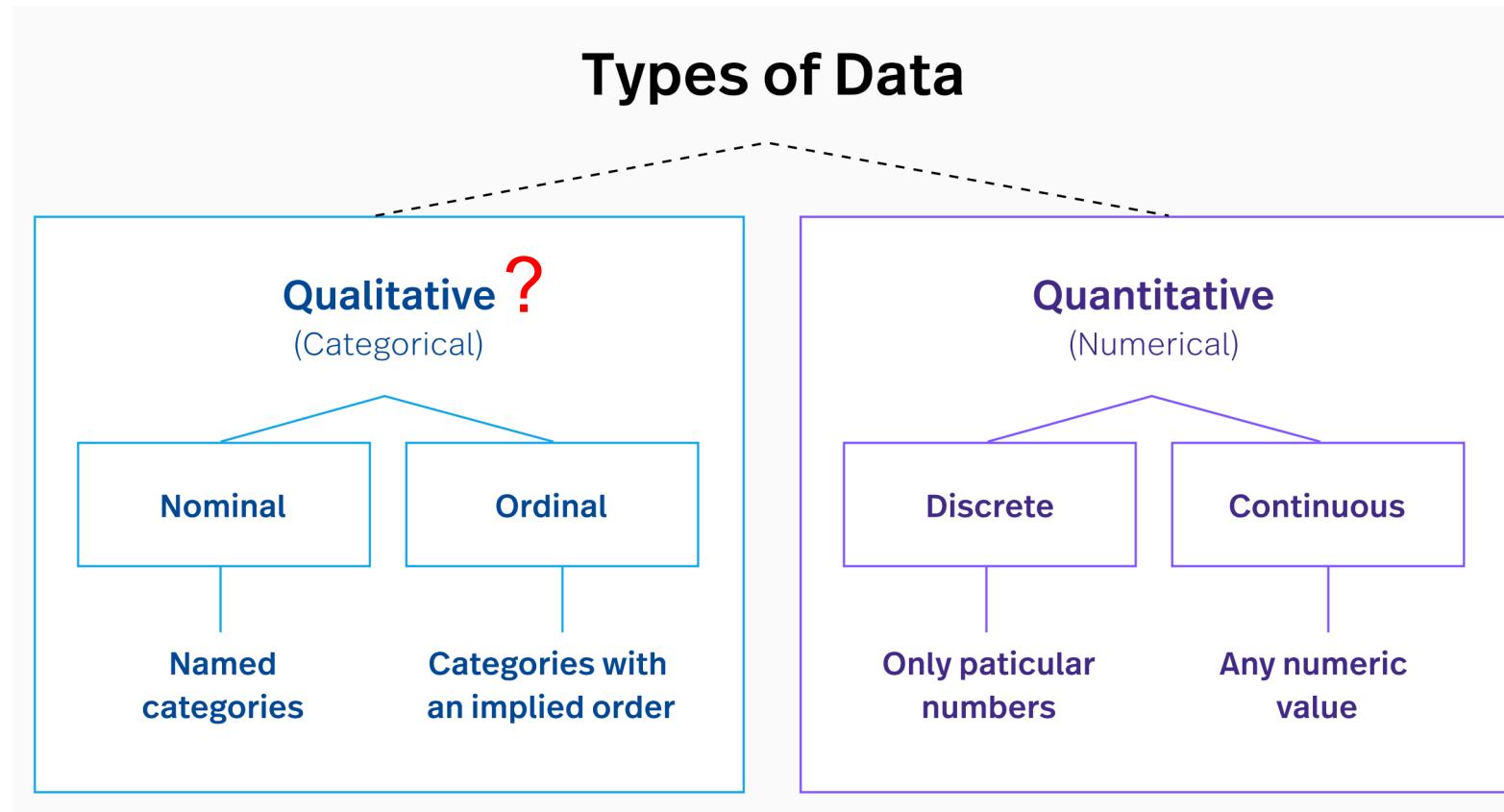
Qualitative Daten:

- Nur mit Worten und nicht mit Zahlen beschreibbar

Quantitative Daten:

- Zahlenwerte aus Messungen oder Zählungen
- Messskala existiert, so dass je nach Skalentyp Vergleiche möglich sind (z.B.: Alter, Jahresniederschlagsmenge; Pendlerdistanz, CO₂-Ausstoss von Fahrzeugen)

Qualitativ vs. Quantitativ



Daten oder besser Methoden?

“There is this **misconception about quantitative data** — that anything numbers-related is quantitative. But that’s not the case. When we talk about quantitative research, we are mostly talking about data which can be used for inference.

For example, say I ask a group of people how many of them drink coffee. Let’s say I found that 80 people out of 100 drink coffee. That’s a number — but it’s not quantitative data — it’s just a description.

Descriptive data without statistical significance, even if it has numbers, does not constitute quantitative data.”

Aditi Paul, PhD, Mixed-methods researcher

Qualitativ vs. Quantitativ

“Die Unterscheidung in qualitative und quantitative Untersuchungen ist **nicht so trennscharf** wie häufig angenommen. Sinnvoller wäre vermutlich eine Abgrenzung zwischen **hypotesenprüfenden und theoriebildenden Untersuchungen**. Ganz allgemein gilt, dass auch quantitative Verfahren – also alle Methoden, die mit nötigen Mindestmengen, mit Messen, Zählen und Berechnen zu tun haben, auf qualitativen Füßen stehen und umgekehrt. Wer quantitativ arbeitet, bedient sich vieler Annahmen oder theoretischer Setzungen, die (zumeist) nicht mehr infrage gestellt, sondern „nur“ genutzt werden. Was etwa ein Mittelwert oder eine Korrelation aussagen, muss vorab theoretisch bestimmt worden sein. ...”

Qualitativ vs. Quantitativ

Qualitative Forschungsansätze

Vorrangiges Ziel:

- Phänomene rekonstruieren
- Hypothesen und Theorien generieren

Voraussetzungen:

- offener, explorativer Zugriff auf das Phänomen
- Wissen über qualitative Verfahren und Methoden zur Datenerhebung, -auswertung und -interpretation

Quantitative Forschungsansätze

Vorrangiges Ziel:

- Phänomene messbar machen und statistisch auswerten
- Überprüfung von Hypothesen und Theorien

Voraussetzungen:

- Vorliegen von Hypothesen und Theorien, die überprüft werden können
- Wissen über statistische Verfahren und Methoden zur Datenerhebung, -auswertung und -interpretation

Qualitativ vs. Quantitativ

Qualitative Forschungsansätze

Typische Fragestellungen:

- Was heißt es, lese-/schreib-/rechenkompetent zu sein?
- Gibt es weitere gemeinsame Merkmale von Personen, die x haben, und wenn ja, welche?

Typische Verfahren der Datenerhebung:

- narratives Interview
- Gruppendiskussion
- Beobachtung

Quantitative Forschungsansätze

Typische Fragestellungen:

- Sind die SchülerInnen im Sinne von PISA lese-/schreib-/rechenkompetent?
- Trifft auf Personen, die x haben, auch y zu?

Typische Verfahren der Datenerhebung:

- standardisierter Fragebogen
- Experiment
- Messungen

Qualitativ vs. Quantitativ

Kriterium	Qualitative Forschung	Quantitative Forschung
Forschungsperspektive	Sicht der Betroffenen steht im Mittelpunkt des Interesses	Sicht aus der Außenperspektive des Forschers
Forschungskontext	„Weiche“, realitätsnahe Daten	„Harte“, replizierbare Daten
Forschungsprozess	Dynamisch	Statisch
Theoriebezug	Entdeckung und Entwicklung von Hypothesen und Theorien aus dem Material	Bestätigung von vorab festgelegten Hypothesen
Vorgehensweise	Induktiv, Sinnverstehen	Deduktiv, Messen
Erkenntnisinteresse	Erforschung von Lebenswelten und Interaktionen	Erklären kausaler Zusammenhänge Verallgemeinerbarkeit von Stichproben auf Populationen
Methode	z. B. Interview, Gruppen-diskussion, qualitative Inhaltsanalyse, Beobachtung	z. B. Versuch, Experiment, Beobachtung

Lernziele

- Die korrekten statistische Kennzahlen zur Beurteilung von Stichprobendaten auswählen und berechnen
- Statistische Tests verstehen, passenden Tests auswählen und durchführen können.
- Korrelations- und Regressionsanalysen selbstständig durchführen, unter Berücksichtigung von deren Anwendungsbedingungen.
- Alle behandelten statistischen Analysen kritisch zu beurteilen
- Grundlagen weiterführender statischer Verfahren und deren Anwendungsbereich verstehen.
- Fähigkeit diese Methoden mit Statistiksoftware R in der Praxis anzuwenden

Veranstaltung

Theorie:

- Vorlesung (erster Teil der Veranstaltung)
- Folien (Ilias)
- Bücher
- Unterlagen aus «Statistik für Naturwissenschaften»

Übungen:

- Zweiter Teil der Veranstaltung
- Praktische Anwendung
- Enthält oft Zusatzinformationen

Leistungskontrolle

- Multiple Choice Prüfung (voraussichtlich IliasExam)

Anmeldung via KSL

- **Meldet euch für die Lehrveranstaltung (LV) an**, dann seid ihr erreichbar via Email für Informationen und habt automatisch Zugriff auf den Kurs.
- Ihr müsst euch **auch zur Leistungskontrolle (LK) angemelden**, damit ich die Noten eintragen kann. (dies ist vermutlich erst in ein paar Wochen möglich!)

Themen

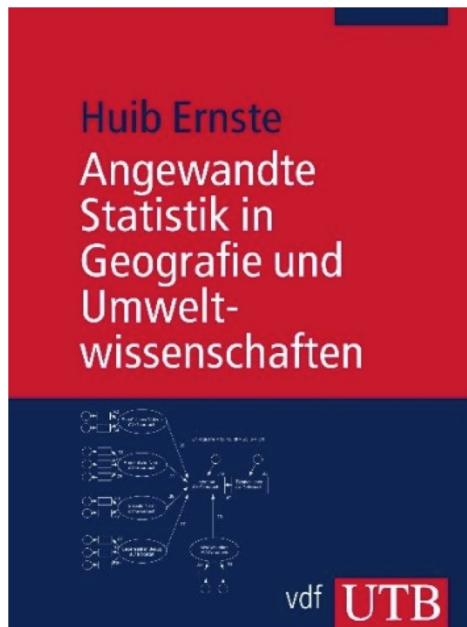
- | | |
|--------------------|---------------------------------------------------------------|
| 16. Sep. | Einführung |
| 23. Sep. – 7. Okt. | R - Programmierkurs |
| 14. Okt. | Deskriptive Statistik |
| 21. Okt. | Verteilungen, Wahrscheinlichkeiten, Satz von Bayes |
| 28. Okt. | Hypothesentests |
| 4. Nov. | Korrelationsanalyse |
| 11. Nov. | Einfache und multiple Regressionsanalyse, Trend |
| 18. Nov. | Modellvalierung |
| 25. Nov. | Clusteranalyse und Machine Learning Modelle |
| 5. Dez. | Hauptkomponentenanalyse, Extremwertanalyse, Zeitreihenanalyse |
| 12. Dez. | Fallen und Fehler in der Statistik |
| 19. Dez. | Fragestunde, Probeprüfung |

u^b

b
UNIVERSITÄT
BERN

Literatur

Ernste, H.: Angewandte Statistik in Geographie und Umweltwissenschaften. vdf ETH, Zürich.



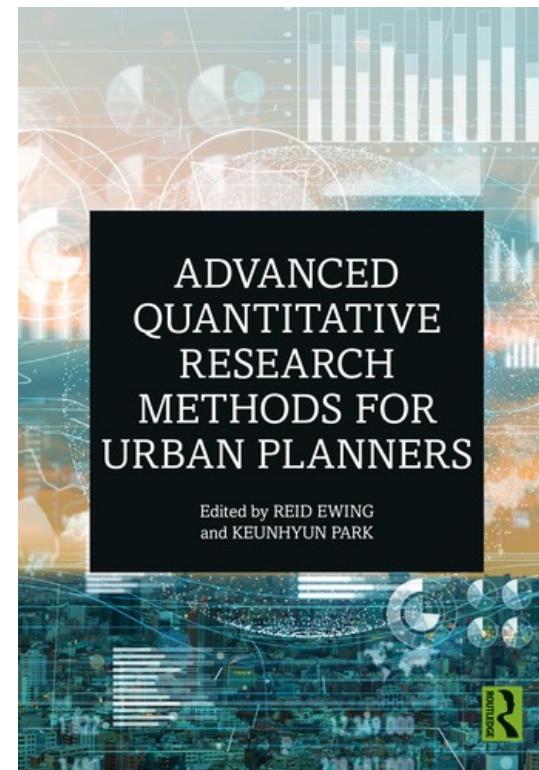
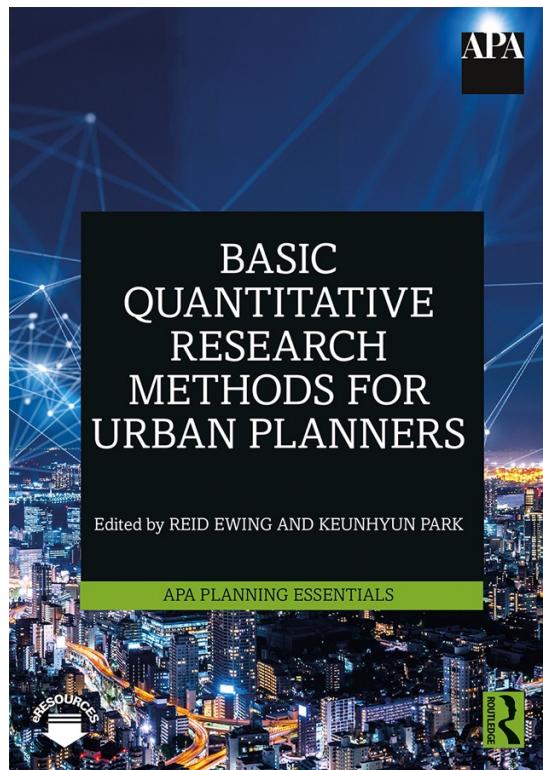
Bahrenberg / Giese / Nipper: Statistische Methoden in der Geographie - Band 1 und 2, Bornträger Verlag



u^b

b
UNIVERSITÄT
BERN

Fachspezifischere Literatur, z.B.



Literaturstudium

Gundbegriffe, Skalentypen, Maßzahlen	Bahrenberg I: Kap. 1 und 4
Verteilungen	Bahrenberg I: Kap. 4
Hypothesentests	Bahrenberg I: Kap. 5; Ernste Anh. B; Ewing I: Kap. 10
Korrelationsanalyse, Varianz-Kovarianz-Matrix	Ernste Kap. 2, Anh. A5.1-A5.2; Ewing I: Kap. 9
Einfache Regressionsanalyse, Trend	Bahrenberg I: Kap. 6; Ernste Kap. 3; Ewing I: Kap. 12
Multiple Regression	Bahrenberg I: Kap. 2; Ernste Kap. 4; Ewing I: Kap. 12
Hauptkomponentenanalyse	Bahrenberg II: Kap. 6; Ernste Kap. 9, 10; Ewing II: Kap. 5
Clusteranalyse	Bahrenberg II: Kap. 2; Ewing II: Kap. 6
Repetition Lineare Algebra	Ernste Anh. A1-A4

R installieren

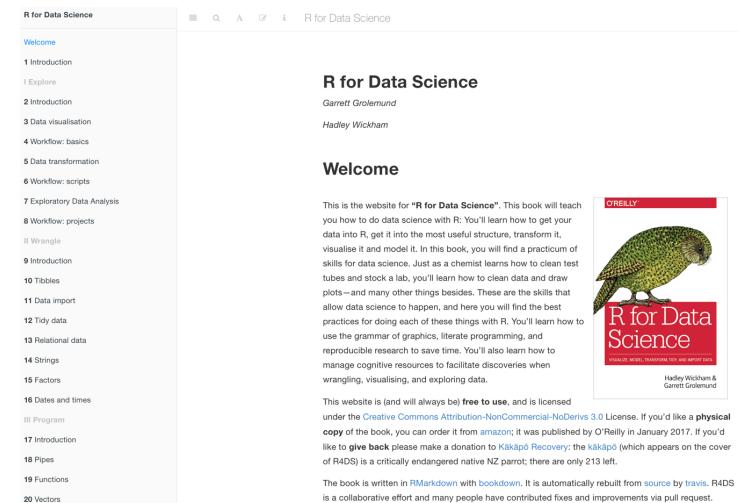
1. „R“ Basis Paket/Binaries installieren:
<http://www.r-project.org>

2. Graphische Oberfläche “**Rstudio Desktop Version**” installieren (läuft nicht ohne vorherige R Installation):
<http://www.rstudio.com>

Meldet Euch bei mir, falls Probleme bei der Installation auftreten

Optional kostenloses R Buch:
<https://r4ds.had.co.nz/index.html>

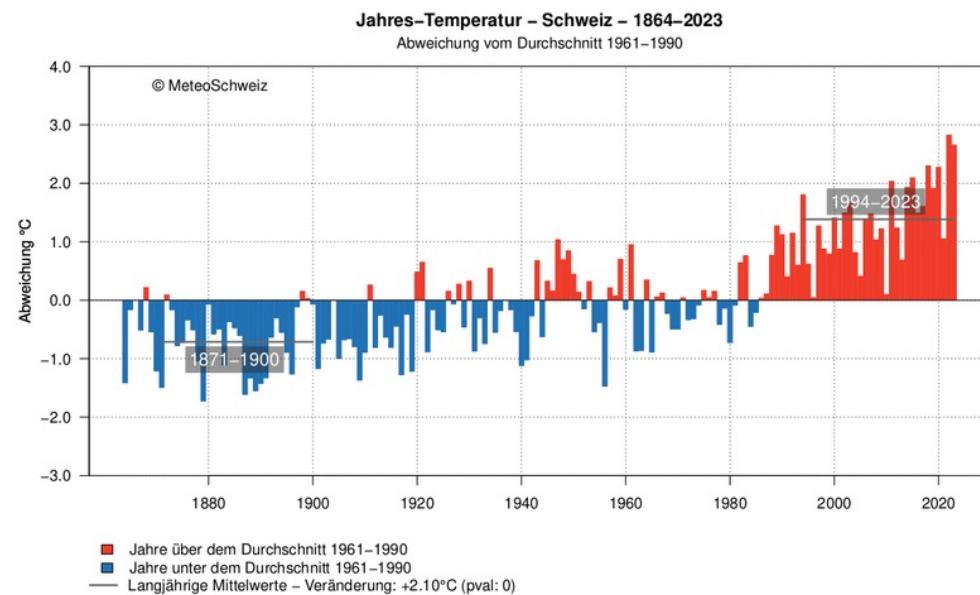
ACHTUNG: Alles weitere baut auf dem R-Kurs auf! UNBEDINGT gut verstehen!



The image shows a screenshot of the "R for Data Science" website on the left and the physical book cover on the right. The website has a sidebar with chapters like Welcome, Introduction, Explore, Data visualisation, Workflow: basics, Data transformation, Workflow: scripts, Exploratory Data Analysis, Projects, Wrangle, Tibbles, Data import, Tidy data, Relational data, Strings, Factors, Dates and times, Program, Introduction, Pipes, Functions, and Vectors. The book cover features a green parrot illustration and the title "R for Data Science" with the subtitle "Tidy Data, Transforming Text and Import Data". It also lists the authors, Hadley Wickham & Garrett Grolemund.

Beispielthema Wetter

- MeteoSchweiz Messstationen
- Temperatur- und Niederschlagsdaten



Messfeld-Station in Aigle mit 10 m hohem Windmasten, einer Messbrücke auf 2 m (Temperatur, Feuchtigkeit und Strahlung) und einem Pluviometer (Niederschlag), 1.5 m über Boden. © MeteoSchweiz

Beispielthema Mentale Gesundheit

Datenhebung:

- Stichprobe oder Grundgesamtheit?

The screenshot shows a news article from the SOTOMO CSS Gesundheitsstudie 2024 website. The article discusses the results of the study, noting a deterioration in mental health compared to previous years, followed by a slight improvement in 2024. The text is highlighted in yellow.

Wie die CSS Gesundheitsstudien der Vorjahre gezeigt haben, hat sich während und nach der Pandemie der physische und psychische Gesundheitszustand der Schweizer Bevölkerung mehr und mehr verschlechtert. 2024 zeigen sich nun erstmals zumindest punktuelle Aufhellungen. So fühlen sich wieder weniger Schweizerinnen und Schweizer ganz oder teilweise krank. Und der Anteil, denen es psychisch gut geht, hat nach markanten Verschlechterungen erstmals wieder zugenommen. Dennoch geht der Trend eher Richtung durchzogen als Richtung gut: Noch nie in dieser Befragungsreihe haben sich so wenige Befragte sehr gesund gefühlt und nie war der wahrgenommene Druck so gross, immer gesund und leistungsfähig zu sein. Der Druck der Leistungsgesellschaft und die Vermischung von Privat- und Berufsleben führen zunehmend zu Stress und Erschöpfung.

Beispielthema Mentale Gesundheit

6.1 DATENERHEBUNG UND STICHPROBE

Die Daten wurden zwischen dem 13. und 21. Juni 2024 erhoben. Die Grundgesamtheit der Befragung bildet die sprachlich integrierte Wohnbevölkerung der Deutschschweiz und der französisch- und italienischsprachigen Schweiz. Die Befragung erfolgte online über das Online-Panel von Sotomo. Nach Bereinigung und Kontrolle der Daten konnten die Angaben von 2456 Personen für die Auswertung verwendet werden.

- Repräsentativ?

The screenshot shows the header of the SOTOMO website for the CSS Health Study 2024. The header includes the SOTOMO logo, a search icon, and language links for FR. The main title "CSS Gesundheitsstudie 2024" is prominently displayed. Below the title, a text block discusses the state of health in Switzerland, mentioning a recent deterioration followed by a slight improvement, and noting increasing stress and exhaustion due to work and private life pressures.

Wie die CSS Gesundheitsstudien der Vorjahre gezeigt haben, hat sich während und nach der Pandemie der physische und psychische Gesundheitszustand der Schweizer Bevölkerung mehr und mehr verschlechtert. 2024 zeigen sich nun erstmals zumindest punktuelle Aufhellungen. So fühlen sich wieder weniger Schweizerinnen und Schweizer ganz oder teilweise krank. Und der Anteil, denen es psychisch gut geht, hat nach markanten Verschlechterungen erstmals wieder zugenommen. Dennoch geht der Trend eher Richtung durchzogen als Richtung gut: Noch nie in dieser Befragungsreihe haben sich so wenige Befragte sehr gesund gefühlt und nie war der wahrgenommene Druck so gross, immer gesund und leistungsfähig zu sein. Der Druck der Leistungsgesellschaft und die Vermischung von Privat- und Berufsleben führen zunehmend zu Stress und Erschöpfung.

Beispielthema Mentale Gesundheit

6.1 DATENERHEBUNG UND STICHPROBE

Die Daten wurden zwischen dem 13. und 21. Juni 2024 erhoben. Die Grundgesamtheit der Befragung bildet die sprachlich integrierte Wohnbevölkerung der Deutschschweiz und der französisch- und italienischsprachigen Schweiz. Die Befragung erfolgte online über das Online-Panel von Sotomo. Nach Bereinigung und Kontrolle der Daten konnten die Angaben von 2456 Personen für die Auswertung verwendet werden.

6.2 REPRÄSENTATIVE GEWICHTUNG

Da sich die Teilnehmenden der Umfrage selbst rekrutieren (opt-in), ist die Zusammensetzung der Stichprobe nicht repräsentativ für die Grundgesamtheit. Den Verzerrungen in der Stichprobe wird mittels statistischer Gewichtungsverfahren entgegengewirkt. Zu den Gewichtungskriterien gehören Geschlecht, Alter und Bildung, politische Orientierung und Eigentümerquoten. Die Randverteilungen dieser Merkmale wurden für die Sprachregionen der Schweiz jeweils separat berücksichtigt. Dieses Vorgehen gewährleistet eine hohe soziodemografische Repräsentativität der Stichprobe. Für die vorliegende Gesamtstichprobe beträgt das 95-Prozent-Konfidenzintervall (für 50 Prozent Anteil) +/-2 Prozentpunkte.

The screenshot shows the header of the SOTOMO website. On the left is a menu icon (three horizontal lines). In the center is the SOTOMO logo, which consists of the word "SOTOMO" in a bold, sans-serif font with a small circle integrated into the letter "o". To the right of the logo are two small icons: a magnifying glass for search and a flag for language selection. Below the header, the main title "CSS Gesundheitsstudie 2024" is displayed in large, bold, black font. A large block of text below the title discusses the findings of the study, mentioning a general decline in physical and mental health and a recent improvement. A yellow highlighted section quotes a participant's statement about stress and exhaustion.

Wie die CSS Gesundheitsstudien der Vorjahre gezeigt haben, hat sich während und nach der Pandemie der physische und psychische Gesundheitszustand der Schweizer Bevölkerung mehr und mehr verschlechtert. 2024 zeigen sich nun erstmals zumindest punktuelle Aufhellungen. So fühlen sich wieder weniger Schweizerinnen und Schweizer ganz oder teilweise krank. Und der Anteil, denen es psychisch gut geht, hat nach markanten Verschlechterungen erstmals wieder zugenommen. Dennoch geht der Trend eher Richtung durchzogen als Richtung gut: Noch nie in dieser Befragungsreihe haben sich so wenige Befragte sehr gesund gefühlt und nie war der wahrgenommene Druck so gross, immer gesund und leistungsfähig zu sein. Der Druck der Leistungsgesellschaft und die Vermischung von Privat- und Berufsleben führen zunehmend zu Stress und Erschöpfung.

Beispielthema Mentale Gesundheit

Studie Uni Bern (<https://sub.unibe.ch/admin/data/files/asset/file/2197/sub-umfrage-2023-grossbericht-final.pdf?lm=1699284939>)

Mentales Wohlbefinden WHO-5

Die Weltgesundheitsorganisation (WHO) definiert Mentale Gesundheit als "einen Zustand des Wohlbefindens, in dem jede*r Einzelne die eigenen Fähigkeiten erkennt, mit den normalen Belastungen des Lebens zurechtkommt, produktiv und fruchtbar arbeiten kann und in der Lage ist, einen Beitrag zur eigenen Gemeinschaft zu leisten.²⁹

Der WHO-5 ist ein kurzer Fragebogen, der aus fünf Fragen besteht, die das subjektive Wohlbefinden der Befragten erfassen. Dabei geben die Befragten bei jeder von fünf Aussagen an, wie sie sich in den letzten zwei Wochen gefühlt haben. Dabei wird eine sechsstufige Likert-Skala von 0 = "Zu keinem Zeitpunkt" bis 5 = "Die ganze Zeit" verwendet. Der Indexwert wird durch einfache Summierung der fünf Itemwerte gebildet, wobei sich ein Maximalwert von 25 ergibt. Dieser Wert wird mit 4 skaliert, um ein Ergebnis zwischen 0-100 zu erhalten. Höhere Werte zeigen ein besseres Wohlbefinden an.

(0 tiefstmögliche, 100 bestmögliche psychische Wohlbefinden)

Skala ?

Beispielthema Mentale Gesundheit

Studie Uni Bern (<https://sub.unibe.ch/admin/data/files/asset/file/2197/sub-umfrage-2023-grossbericht-final.pdf?lm=1699284939>)

WHO-5 In den letzten zwei Wochen...	Durchschnittswert, skaliert (0-100)
...war ich froh und guter Laune	59.04
...habe ich mich ruhig und entspannt gefühlt	45.48
... habe ich mich energisch und aktiv gefühlt	49.93
... habe ich mich beim Aufwachen frisch und ausgeruht gefühlt	39.83
...war mein Alltag voller Dinge, die mich interessieren	59.46
Total (n=1301)	50.75

Was bedeutet “n=...” ?

Beispielthema Mentale Gesundheit

Studie Uni Bern (<https://sub.unibe.ch/admin/data/files/asset/file/2197/sub-umfrage-2023-grossbericht-final.pdf?lm=1699284939>)

Im Wissen, dass 50 einen Schwellenwert darstellt, unter welchem Untersuchungen zu möglichen Symptomen einer Depression angebracht sind, weckt der durchschnittliche Score des WHO-5 Fragebogens unter allen Studierenden der Universität Bern mit 50.79 Aufmerksamkeit. Es ist schwierig Vergleiche anzustellen, da diese Fragen zum ersten Mal im Rahmen der SUB Umfrage gestellt werden und zurzeit keine vergleichbaren Ergebnisse unter Schweizer Studierenden vorhanden sind. Bei Betrachtung der Durchschnittsergebnisse unter anderen Personengruppen liegen die Werte der Studierenden an der Universität Bern unter anderem tiefer als die Durchschnitte für alle 25-34-jährigen in Deutschland, Österreich und den Niederlanden (65)³¹. Die Ergebnisse weisen auf ein bedenklich tiefes Wohlbefinden der Studierenden hin, was einen klaren Handlungsbedarf aufzeigt.

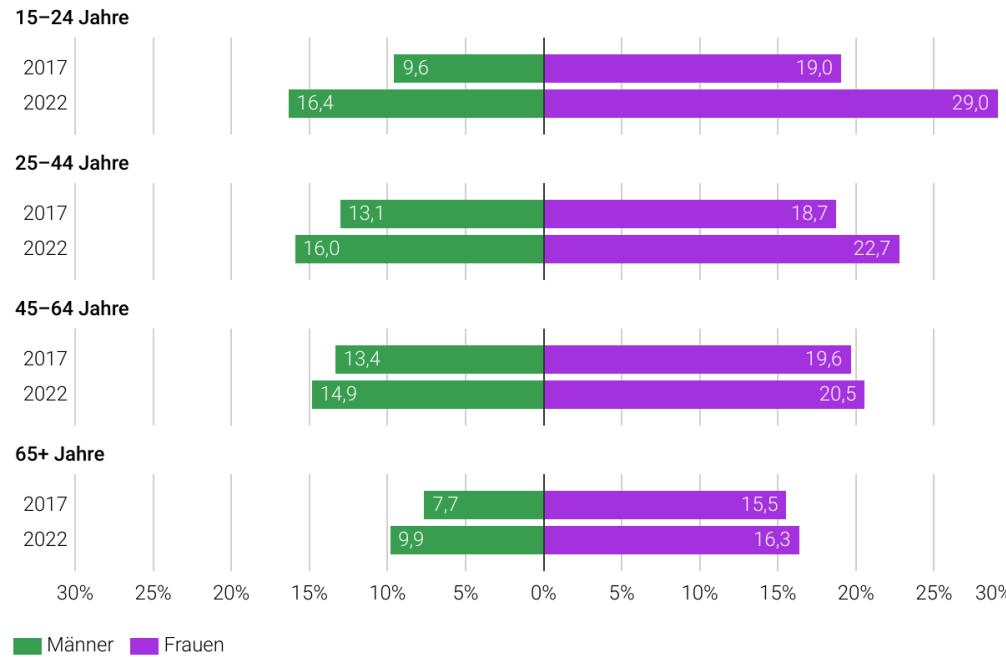
Beispielthemen für Datenanalysen

Bundesamt für Statistik

1. Was fällt euch auf?
2. Stellt Hypothesen für mögliche Ursachen auf!

Mittlere oder hohe psychische Belastung

Bevölkerung ab 15 Jahren in Privathaushalten



Quelle: BFS – Schweizerische Gesundheitsbefragung (SGB)

© BFS 2023

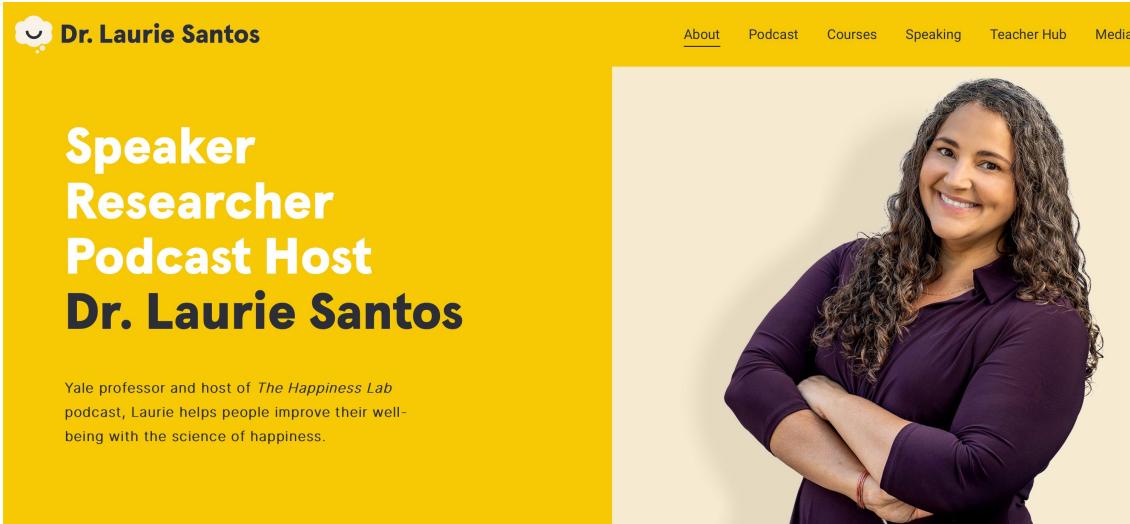
Beispielthemen für Datenanalysen

Sternstunde Philosophie (<https://www.srf.ch/play/tv/sternstunde-philosophie/video/thomas-fuchs---macht-uns-die-gesellschaft-krank?urn=urn:srf:video:381af386-15cb-4b65-8c9d-d56dfa0a3d6a>)

- Ängste (Zukunft wie Klimakrise, Arbeit, Kriege, ...)
- Machtlosigkeit, Handlungsunfähigkeit (individuell oder sogar der Politik)
- Lineare Beschleunigung in Gesellschaft, Leben, Arbeit
- Ständige Bewertung (neben Schule/Uni/Arbeit auch soziale Medien)

Beispielthemen für Datenanalysen

Wir lernen hier ein paar Studien kennen und machen hier ein Experimente und Analysen aus den Kursen: “Psychology and the Good Life” und “The Science of Well-Being“ von Yale Professorin Laurie Santos



The screenshot shows the homepage of Dr. Laurie Santos' website. At the top left is a yellow sidebar with a smiling emoji icon and the text "Dr. Laurie Santos". To the right is a navigation bar with links: About (underlined), Podcast, Courses, Speaking, Teacher Hub, and Media. The main content area features a large portrait of Dr. Santos, a woman with curly hair, wearing a purple long-sleeved shirt, standing with her arms crossed against a light background. To the left of the portrait, the text "Speaker", "Researcher", "Podcast Host", and "Dr. Laurie Santos" is displayed vertically. Below this, a smaller text block reads: "Yale professor and host of *The Happiness Lab* podcast, Laurie helps people improve their well-being with the science of happiness."

Aufgabe

Hört euch die erste Folge des Podcasts “The Happiness Lab” an, in eurer Podcast App oder hier: <https://www.pushkin.fm/podcasts/the-happiness-lab-with-dr-laurie-santos/you-can-change>

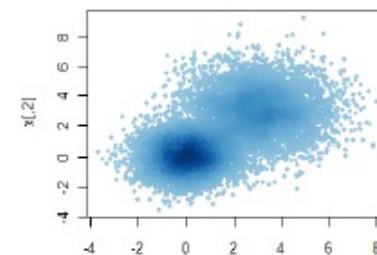
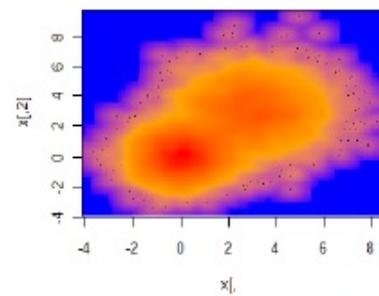
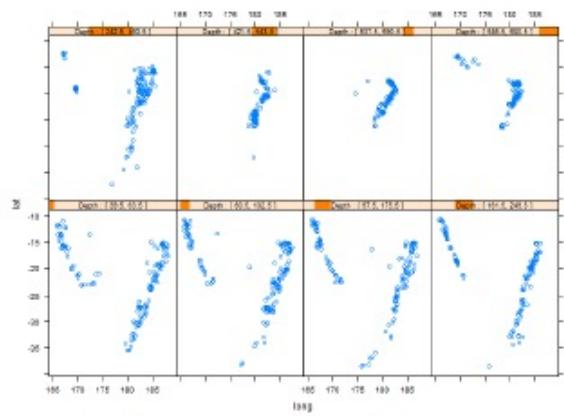
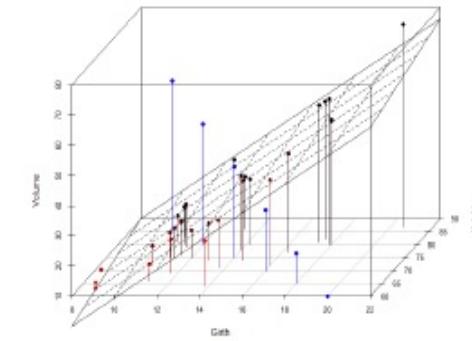
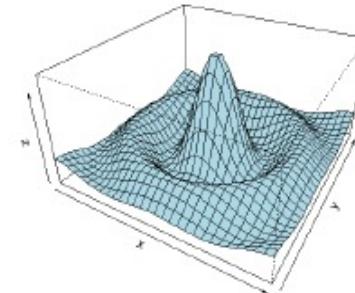
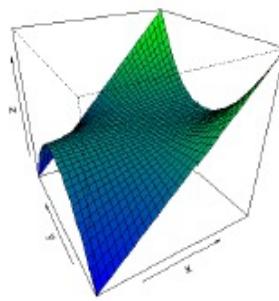
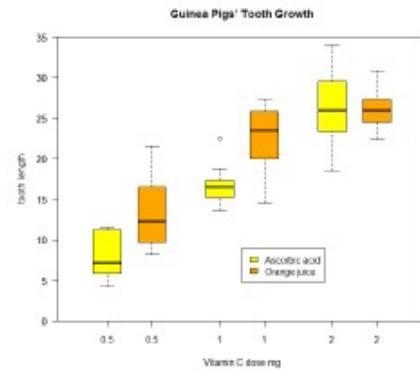
- Notiert euch alle Punkte, wo quantitative Forschungsmethoden erwähnt werden oder vermutlich genutzt wurden, um zu den wissenschaftlichen Erkenntnissen zu gelangen.
- Welche Methoden davon kennt ihr schon?
- Welche methodischen Schwierigkeiten werden erwähnt bzw. erkennt ihr im Podcast?
- Diskutiert mit euren Kommilitonen, was ihr neues gelernt habt, was für euch hilfreich war und ob diese sich die Ergebnisse aus den USA auf die Schweiz übertragen lassen?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

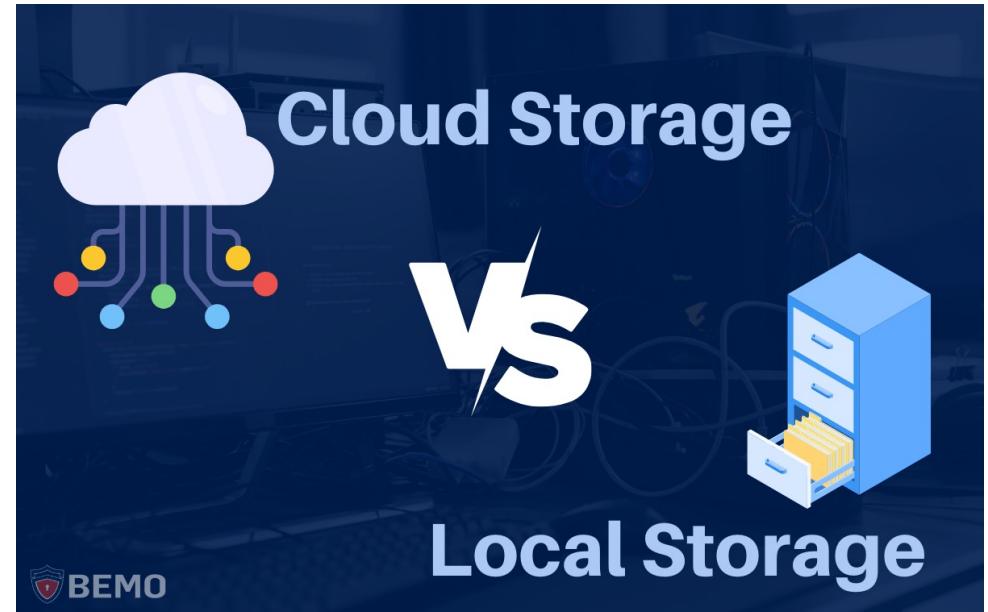
Computer Basics and R



Computer Basics

Dateisystem

- > Dateisystem und Mountpunkte in Finder und Konsole zeigen



Computer Basics

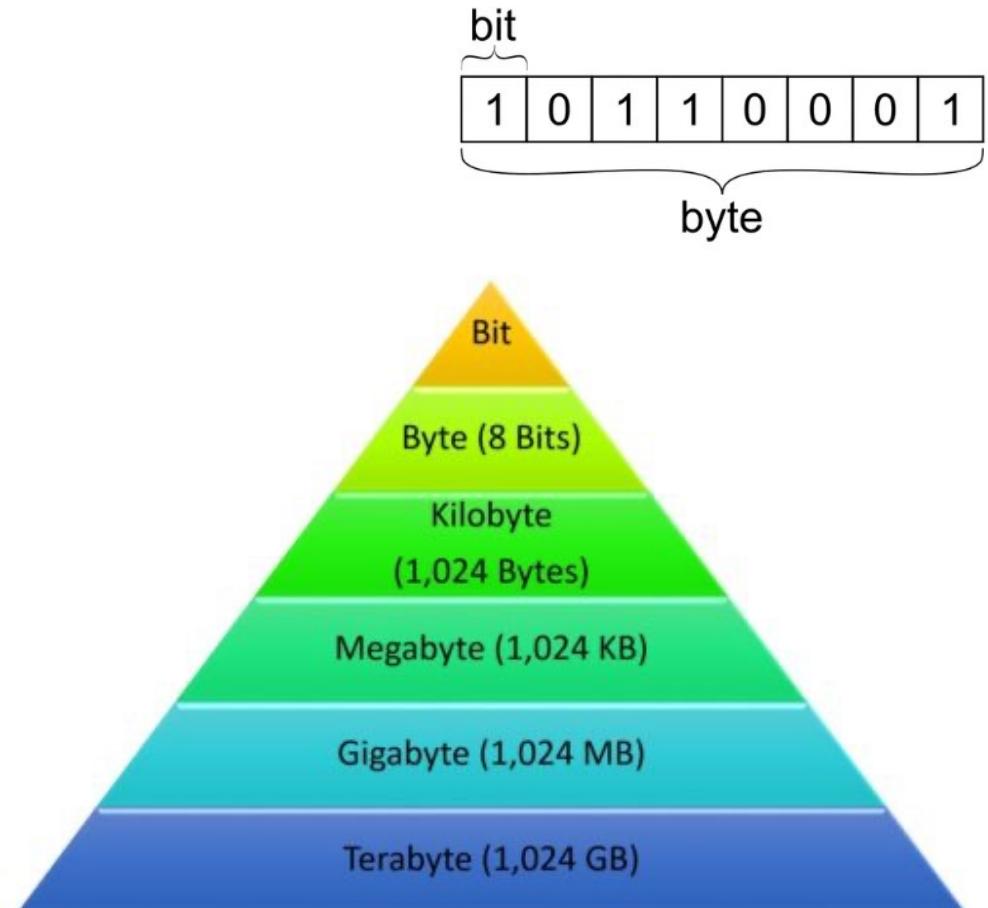
1 Bit: 1/0

1 Byte: kodiert einen Buchstaben

1 Megabyte ?

1 Gigabyte ?

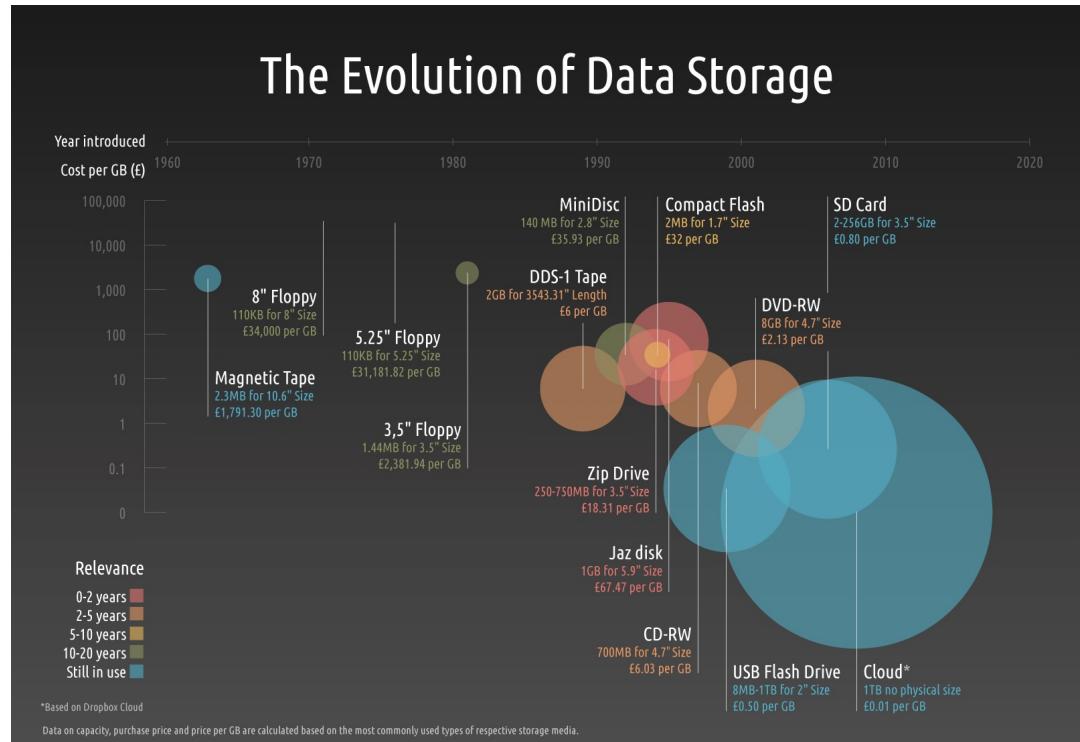
1 Terabyte ?



Computer Basics

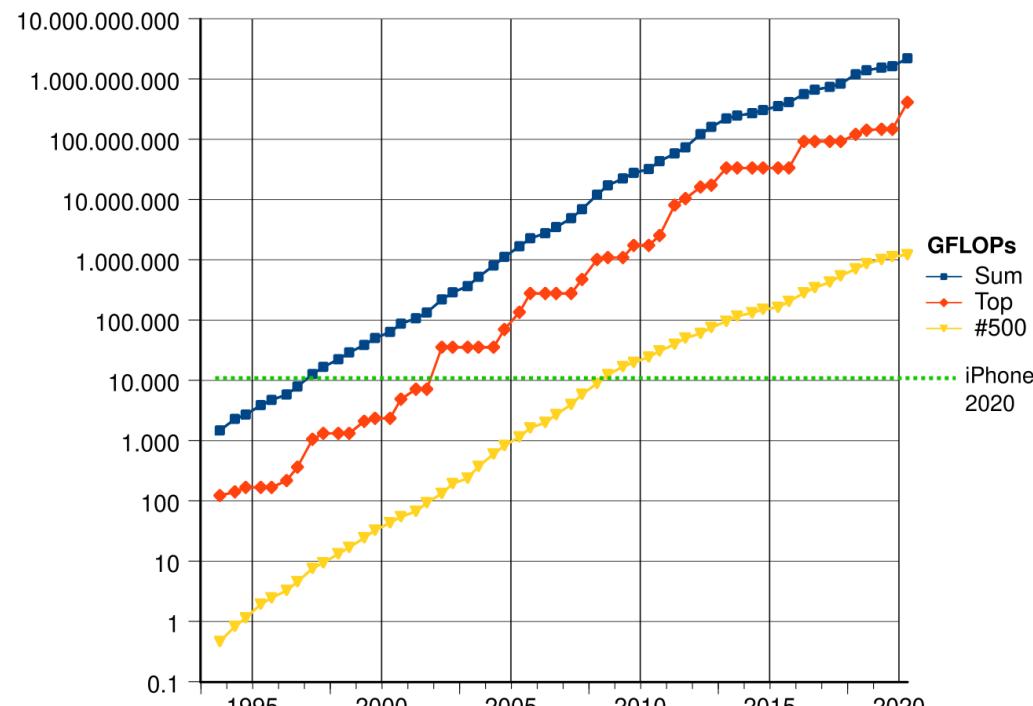
Datengrößen:

- > 100 Jahresmittel der Temperatur an einer Station wenige KB
- > 100 Jahre Tagesmittel einer Station ca. 1 MB
- > Zeitschritt eines Wettervorhersagemodells multivariat bei 1 km räumlicher Auflösung ca. 2 TB

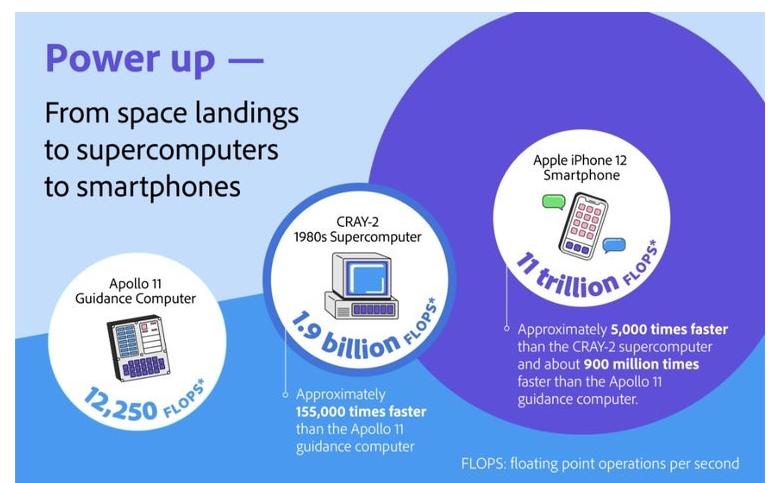


Computer Basics

Computerleistungen einschätzen (iPhone vs Supercomputer)



FLOPS = Floating point operations per second



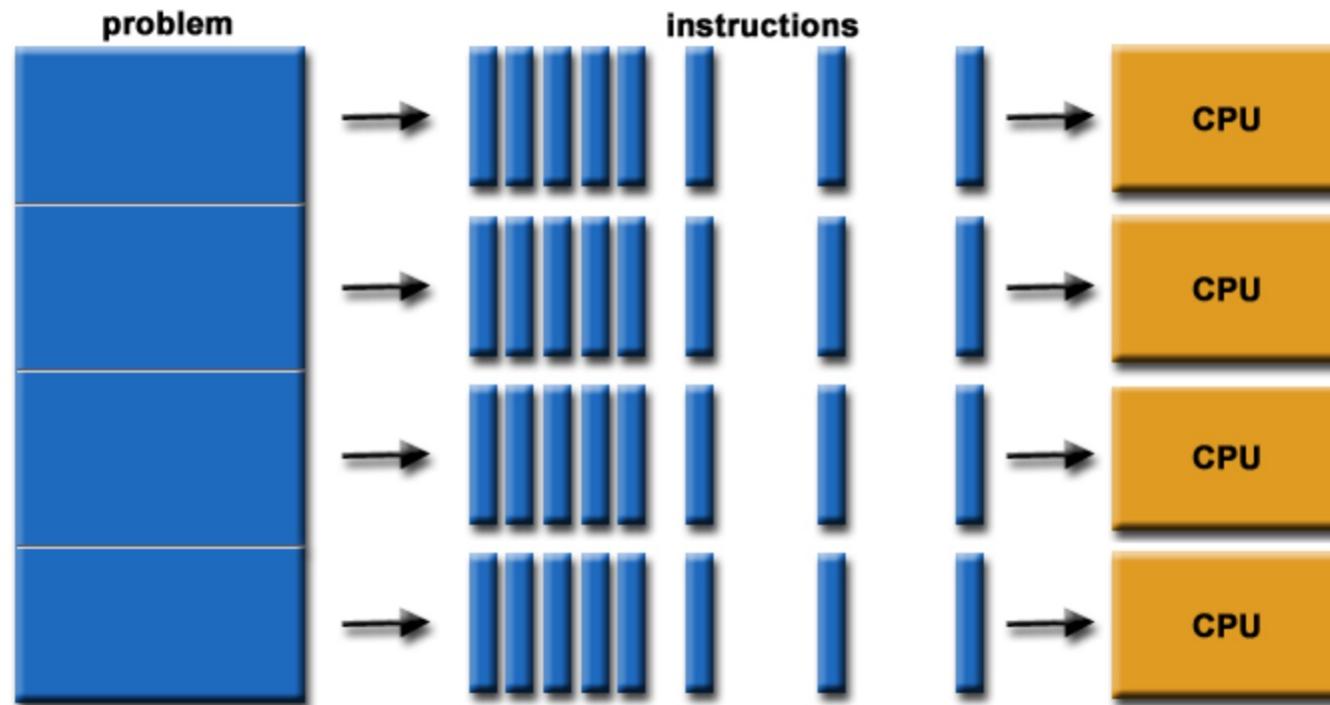
Cloud Computing

- › benutzt ihr die Online Office Version im Browser oder eine auf eurem Computer installierte und wieso?



Computer Basics

CPU Kerne / Parallelisierung



Flaschenhals: Rechenzeit oder Schreibzeit

Computer Basics

Speicherplatzoptimierung?

- > Zwischenschritte speichern?
- > Am Ende wieder löschen?
- > Dateiformate/Komprimierung

Codeoptimierung ?

- > Zeit zum Code schreiben / optimieren
- > Laufzeit des Codes
- > Anzahl der Codeläufe

Computer Basics

Organisation, Dateiformate und Dateinamen

- > KEINE Leerzeichen
- > KEINE Sonderzeichen wie Umlaute
- > Betriebssystemunabhängigkeit
- > ACHTUNG: Länderspezifische Besonderheiten wie Dezimalkomma vs Dezimalpunkt

1 FIGURE OUT LOCATION

Files can be stored in desktop, documents, downloads, or shared drive.



2 ORGANIZE BY FOLDERS

Organize by dates, departments, or events.



3 ORGANIZE BY SUBFOLDERS

Use date or project type.



4 FORMAT AND GROUP FILES

Format based off file type like excel, doc, pdf, jpeg, and png.



5 NAME FILES STRATEGICALLY

Think about how you would search for it in the future.

6 DOCUMENT THE PROCESS

Create a Standard Operating Procedure (SOP's) for consistency.



7 MAINTENANCE IS KING

Have a routine to update, move, and delete files regularly.

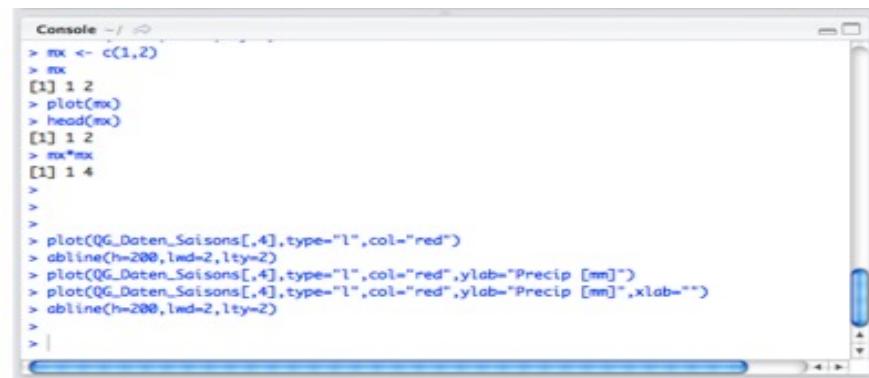


8 OTHER BEST PRACTICES

File immediately and do not store files on your desktop.

Was ist R?

- > Einfache Programmiersprache zur Datenanalyse, statistischen Auswertung und Visualisierung, d.h.
- > geeignet für jegliche Datenbearbeitung
- > inklusive Abbildungen jeder Art, auch Karten
- > "Einfach" da kein Compilieren notwendig ist
- > Freie Software
- > Kein offizielles Menü-System -> Kommando basiert + Text editor
- > 'packages' mit ständig von Benutzer ergänzten Zusatzfunktionen frei im Internet verfügbar
- > Gute Hilfe im Netz



The screenshot shows an R console window with the following text:

```
Console - / 
> mx <- c(1,2)
> mx
[1] 1 2
> plot(mx)
> head(mx)
[1] 1 2
> mx*mx
[1] 1 4
>
>
> plot(QG_Daten_Saisons[,4],type="l",col="red")
> abline(h=200,lwd=2,lty=2)
> plot(QG_Daten_Saisons[,4],type="l",col="red",ylab="Precip [mm]")
> plot(QG_Daten_Saisons[,4],type="l",col="red",ylab="Precip [mm]",xlab="")
>
>
```

Literatur: <https://r4ds.had.co.nz>

Welcome

☰ ⚅ A ⌂ ⓘ R for Data Science

1 Introduction

I Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

II Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

13 Relational data

14 Strings

15 Factors

16 Dates and times

III Program

17 Introduction

R for Data Science

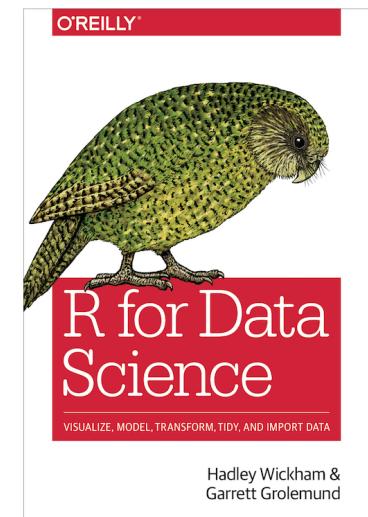
Garrett Grolemund

Hadley Wickham

Welcome

This is the website for “R for Data Science”. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You’ll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed under the [Creative Commons Attribution-NonCommercial-NoDerivs 3.0](#) License. If you’d like a **physical**



Editor, hier wird programmiert und der Code dann in der Konsole ausgeführt

R-Studio (graphische Benutzeroberfläche)

Konsole / Kommandozeile

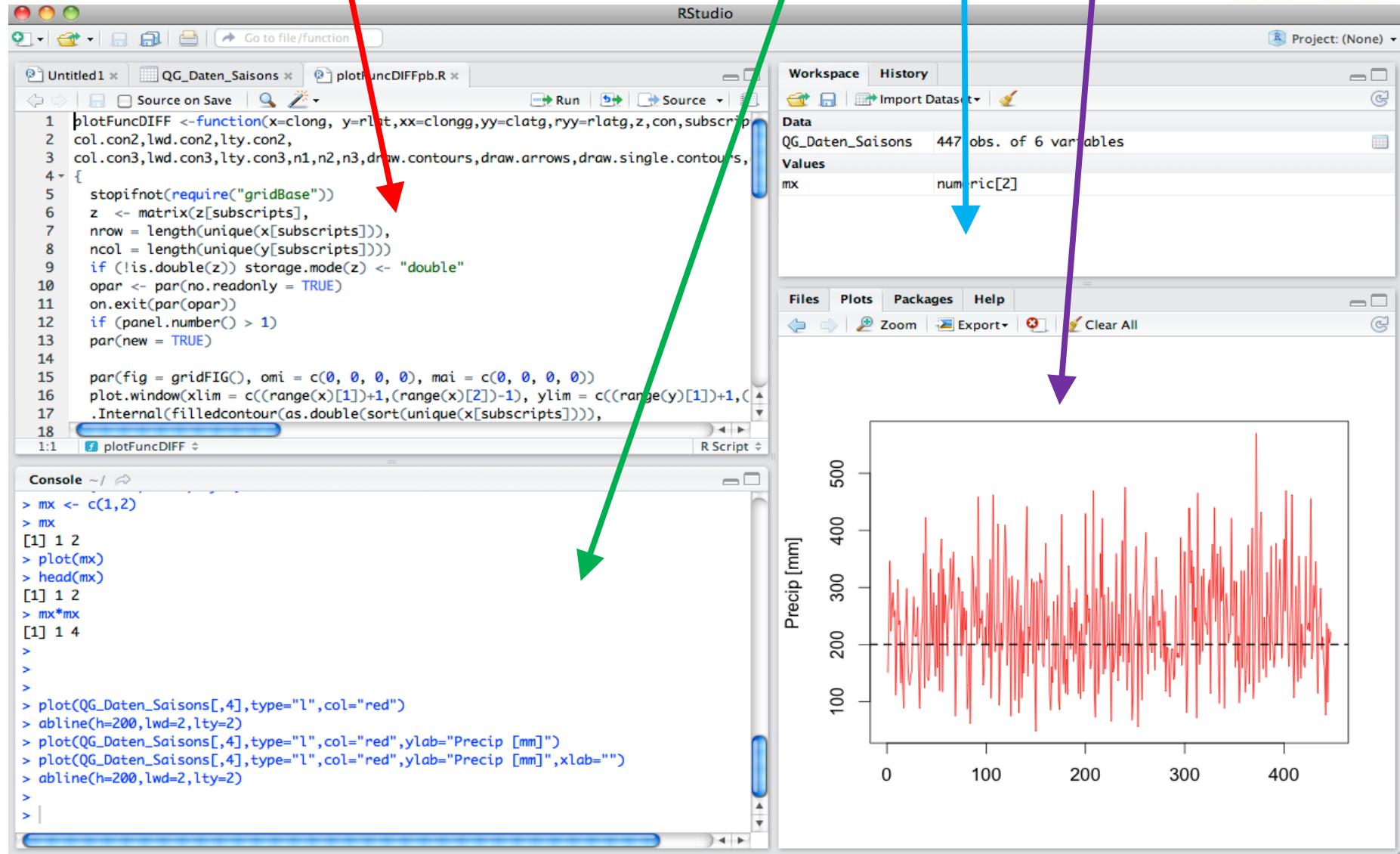
Plots/Hilfe

u^b

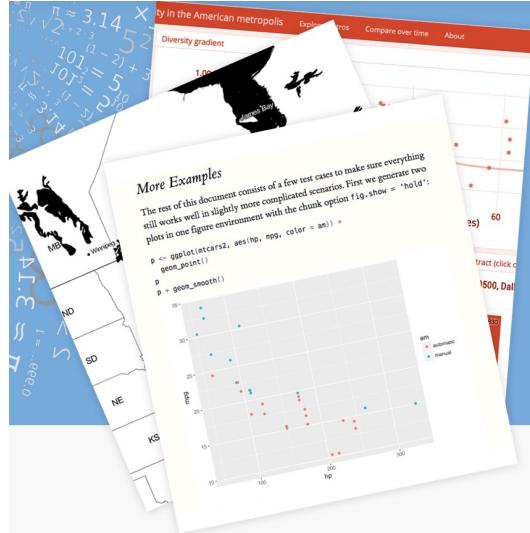
UNIVERSITÄT
BERN

OESCHGER CENTRE

RCH



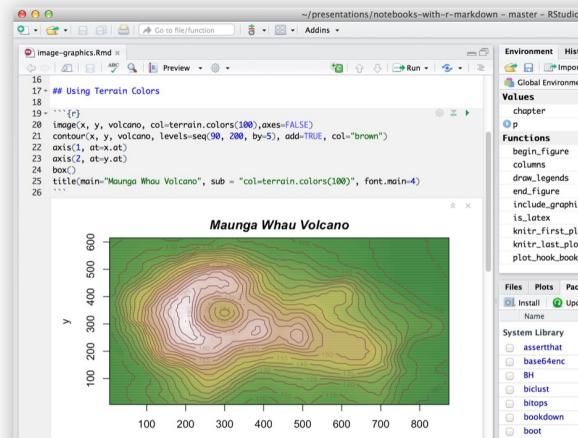
R Markdown or Notebooks



R Markdown documents are fully reproducible. Use a productive [notebook interface](#) to weave together narrative text and code to produce elegantly formatted output. Use [multiple languages](#) including R, Python, and SQL.

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown.
Turn your analyses into high quality documents, reports, presentations and dashboards.



R Markdown supports dozens of static and dynamic output formats including [HTML](#), [PDF](#), [MS Word](#), [Beamer](#), [HTML5 slides](#),

R Markdown

Cheat Sheet
learn more at rmarkdown.rstudio.com

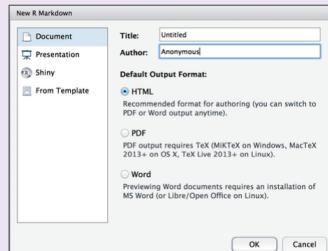
rmarkdown 0.2.50 Updated: 8/14



2. Open File

Start by saving a text file with the extension .Rmd, or open an RStudio Rmd template

- In the menu bar, click **File ▶ New File ▶ R Markdown...**
- A window will open. Select the class of output you would like to make with your .Rmd file
- Select the specific type of output to make with the radio buttons (you can change this later)
- Click OK



4. Choose Output

Write a YAML header that explains what type of document to build from your R Markdown file.

YAML

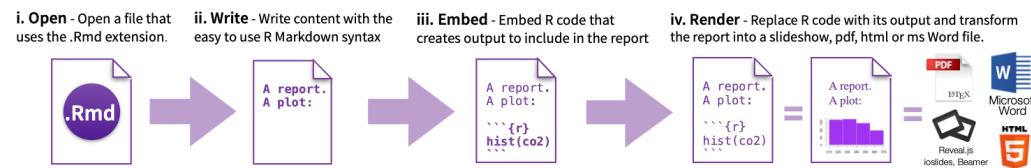
A YAML header is a set of key: value pairs at the start of your file. Begin and end the header with a line of three dashes (---)

```
title: "Untitled"
author: "Anonymous"
output: html_document
---
This is the start of my report. The above is metadata saved in a YAML header.
```

The RStudio template writes the YAML header for you

1. Workflow

R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more. To make a report:



3. Markdown

Next, write your report in plain text. Use markdown syntax to describe how to format text in the final report.

syntax

Plain text
End a line with two spaces to start a new paragraph.
italics and _italics_
bold and __bold__
superscript^2^
~~strikethrough~~
[link](www.rstudio.com)

```
# Header 1
## Header 2
### Header 3
#### Header 4
##### Header 5
###### Header 6

endash: --
emdash: ---
ellipsis: ...
inline equation: $A = \pi r^2$
image: 

horizontal rule (or slide break):
***
```

becomes

Plain text
End a line with two spaces to start a new paragraph.
italics and italics
bold and bold
superscript²
strikethrough
[link](#)

Header 1

Header 2

Header 3

Header 4

Header 5

Header 6

endash: –
emdash: —
ellipsis: ...
inline equation: $A = \pi r^2$

image:

horizontal rule (or slide break):

block quote

R-Struktur, Objektarten

Datentypen: numerisch, alphanumerisch (Text), logisch (TRUE/FALSE)

Objektarten

Vektoren: kein Mischen von Datentypen

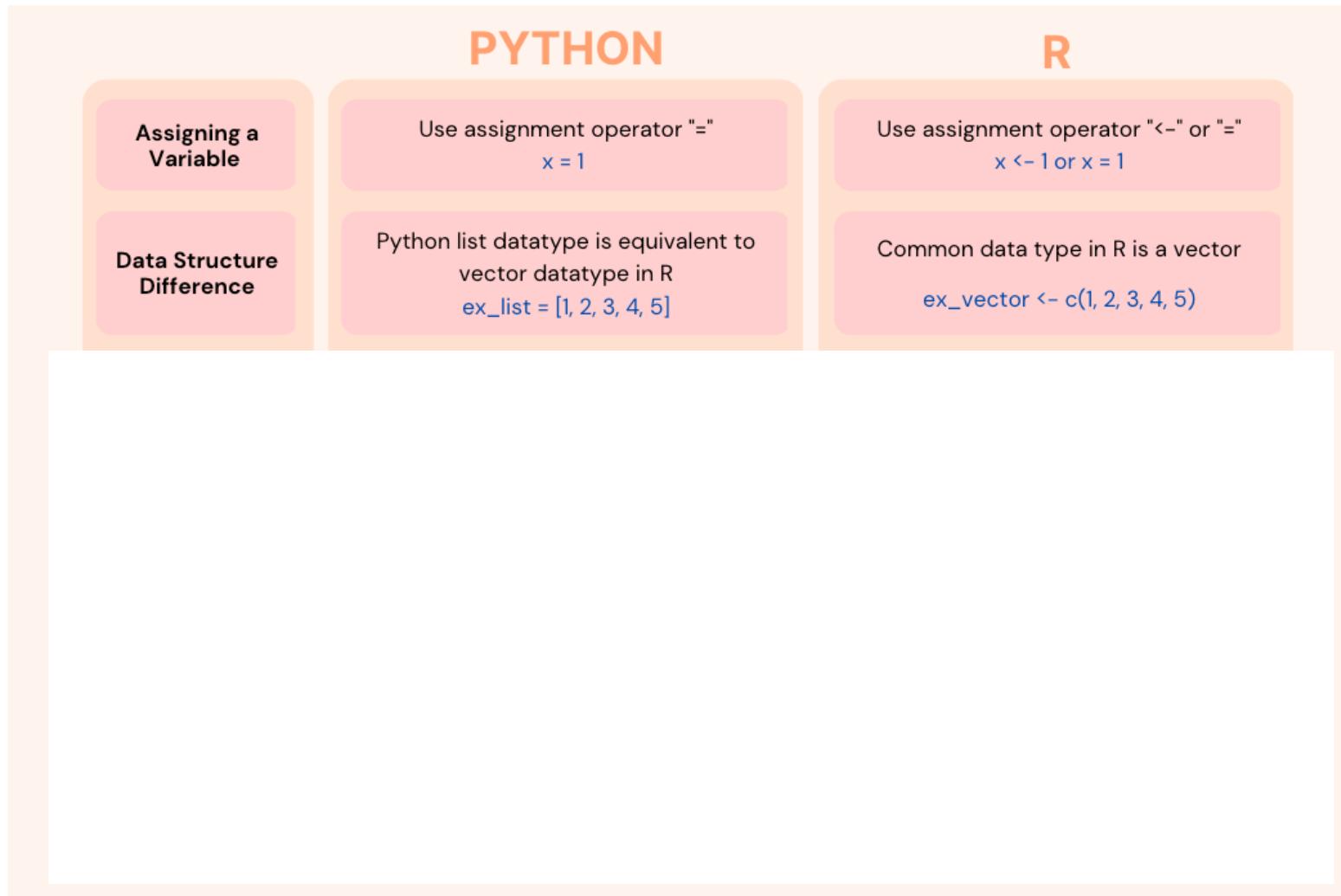
Data Frames ('Tabellen'): Mischen möglich

Matrizen: kein Mischen von Datentypen

Listen: mehrere Vektoren, Data Frames, Matrizen können in Listen zusammengefasst werden

Objekte erzeugt man in dem man einen Objekt-Namen vergibt und Werte mit <- oder = zuweist!

Python vs R



ACHTUNG: copy/paste funktioniert NICHT!!!



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Objektart: Vektor

Numerische Vektoren

```
> a <- c(4, 2, 7) # c Funktion steht für combine or concatenate  
# Parameter von Funktionen immer mit RUNDEN Klammern
```

```
> a
```

```
[1] 4 2 7
```

```
> b <- c(3.1, 5, -0.7)
```

```
> c <- c(a, b)
```

```
> c
```

```
[1] 4.0 2.0 7.0 3.1 5.0 -0.7
```

Einfache oder doppelte Anführungszeichen funktionieren!

Alphanumerische Vektoren

```
> station <- c('Bern', 'Biel', 'Olten')
```

Logische Vektoren

```
> e <- c(TRUE, FALSE, TRUE)
```

Objektart: Vektor

Logische Vektoren

```
> f <- 1:5  
[1] 1 2 3 4 5  
> f >= 3  
[1] FALSE FALSE TRUE TRUE TRUE
```

Vergleichsoperationen <, <=, >, >=, ==, !=

Operationen & (und), | (oder), ! (nicht)

```
> g <- (f>2) & (f<5)  
> g
```

```
[1] FALSE FALSE TRUE TRUE FALSE
```

Objektart: Vektor

```
> seq(0, 3, by=0.5) ... erzeugt Sequenz von 0 bis 3 im 0.5 Abstand  
[1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0
```

```
> rep(5, 3) ... wiederholt 3x den Wert 5  
[1] 5 5 5
```

```
> rep(c(0, 3), length=9) ) ... wiederholt 0 u. 3 bis Vektor 9 Elemente hat  
[1] 0 3 0 3 0 3 0 3 0
```

```
> rep(c(0:3), length=9) ... wiederholt 0,1,2,3 bis Vektor 9 Elem. hat  
[1] 0 1 2 3 0 1 2 3 0 1
```

Objektart: Vektor

- > Wichtige grundlegende Funktionen

Funktion, Beispiel	Bedeutung
<i>length(a)</i>	Länge, Anzahl Elemente
<i>sum(a)</i>	Summe aller Elemente
<i>mean(a)</i>	arithmetisches Mittel der Elemente
<i>var(a)</i>	empirische Varianz
<i>sd(a)</i>	emp. Standardabweichung
<i>range(a)</i>	Wertebereich
<i>min(a), max(a)</i>	Minimum, Maximum Wert des Vektors

Python vs R

PYTHON	R
<p>Chaining operations</p> <p>Method chains with the "." operator <code>df.head(n)</code> or <code>df.describe()</code></p>	<p>Can chain operations using following symbol: "%>%" <code>df %>% head(n)</code> or <code>df %>% summary()</code></p>

Objektart: Data Frame

Vektoren unterschiedlicher Datentypen als Tabelle (=Data Frame) zusammenfassen

> *Dat <- data.frame(station, a, b)* ... Data Frame mit Namen 'Dat' erzeugen

	<i>station</i>	<i>a</i>	<i>b</i>	... Spaltennamen (z.B. <i>a</i> ist Bewölkung, <i>b</i> ist Temperatur zu einem Zeitpunkt)
1	<i>Bern</i>	4	3.1	
2	<i>Biel</i>	2	5.0	
3	<i>Olten</i>	7	-0.7	

Elemente aus Objekten auswählen

Vektor [ELEMENTNUMMER]

> *a[3]; a[1:2]* ... 3. Element; ... 1. bis 2. Element (also 1. und 2.)

Data Frame/Matrix [ZEILE , SPALTE]

> *Dat[1,1]* ... 1. Zeile, 1. Spalte

> *Dat[c(1,2),1]* ... 1. und 2. Zeile der 1. Spalte

> *Dat[1:2,3]* ... 1. bis 2. Zeile der 3. Spalte

> *Dat[,3]* ... alle Zeilen der 3. Spalte (*leer = alle*)

> *Dat[-3,]* ... alle Zeilen bis auf die 3. Zeile, alle Spalten

oder mit Spalten/Zeilennamen

> *Dat[1:2,'b']*

[1] 3.1 5.0

Python vs R

PYTHON

R

Indexing and slicing

Indexing starts from 0, inclusive of the start index and excludes the end index.
`ex_list[0]` and `ex_list[0:2]`
`output: 1` `output: [1, 2]`

Indexing starts from 1, Indexing is inclusive of both start and end index
`ex_vector[1]` and `ex_vector[1:2]`
`output: 1` `output: 1, 2`

Hilfe in R

mit **?Funktionsname** oder **help(Funktionsname)**

```
> ?seq  
> ??histogram  
> help(sum)  
> example(histogram)  
> ...
```

Hilfe im Netz: **GOOGLE, CHATBOTS!**

R als Taschenrechner

> *2 + 5*

[1] 7

> *(2:5)^2* ... auf Vektoren Operationen elementweise angewandt;

[1] 4 9 16 25 ^ oder ** =Potenzfunktion

> *(a+1) * b* ... Klammern wie üblich

[1] 15.5 15.0 -5.6 11.7

> *abs(b)* ... Absolutwert (Betrag)

[1] 3.1 5.0 0.7 1.3

> *exp(a)* ... Exponentialfunktion (e^a)

[1] 54.598150 7.389056 1096.633158 2980.957987

> *log(a)* ... Natürlicher Logarithmus ($\exp(1)^{\log(a)}=a$)

[1] 1.3862944 0.6931472 1.9459101 2.0794415

> *sqrt(a)* ... Wurzel (Square root)

Neues R-Skript oder R Notebook erstellen und umbedingt speichern!!!

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE

RCH

The screenshot shows the RStudio interface. The top bar displays the title "RStudio". The left pane contains a script editor with the code for "plotFuncDIFF" and a console window below it. The right pane includes a "Workspace" panel showing the dataset "QG_Daten_Saisons" and its variables, and a "Plots" panel displaying a line plot of precipitation over time.

Script Editor (Source)

```
1 plotFuncDIFF <- function(x=clong, y=rlat,xx=clongg,yy=clatg,z,con,subscript
2 col.con2,lwd.con2,lty.con2,
3 col.con3,lwd.con3,lty.con3,n1,n2,n3,draw.contours,draw.arrows,draw.single.contours,
4 {
5 stopifnot(require("gridBase"))
6 z <- matrix(z[subscripts]),
7 nrow = length(unique(x[subscripts])),
8 ncol = length(unique(y[subscripts])))
9 if (!is.double(z)) storage.mode(z) <- "double"
10 opar <- par(no.readonly = TRUE)
11 on.exit(par(opar))
12 if (panel.number() > 1)
13 par(new = TRUE)
14
15 par(fig = gridFIG(), omi = c(0, 0, 0, 0), mai = c(0, 0, 0, 0))
16 plot.window(xlim = c((range(x)[1])+1,(range(x)[2])-1), ylim = c((range(y)[1])+1,(range(y)[2])-1),
17 .Internal(filledcontour(as.double(sort(unique(x[subscripts])))),
18
1:1 f plotFuncDIFF
```

Console

```
> mx <- c(1,2)
> mx
[1] 1 2
> plot(mx)
> head(mx)
[1] 1 2
> mx*mx
[1] 1 4
>
>
>
> plot(QG_Daten_Saisons[,4],type="l",col="red")
> abline(h=200,lwd=2,lty=2)
> plot(QG_Daten_Saisons[,4],type="l",col="red",ylab="Precip [mm]")
> plot(QG_Daten_Saisons[,4],type="l",col="red",ylab="Precip [mm]",xlab="")
> abline(h=200,lwd=2,lty=2)
>
> |
```

Plots

A line plot titled "Precip [mm]" showing precipitation over time. The x-axis ranges from 0 to 450, and the y-axis ranges from 0 to 500. A horizontal dashed line is drawn at 200 mm. The plot shows a highly variable pattern with several sharp peaks exceeding 400 mm.

Verzeichnis finden

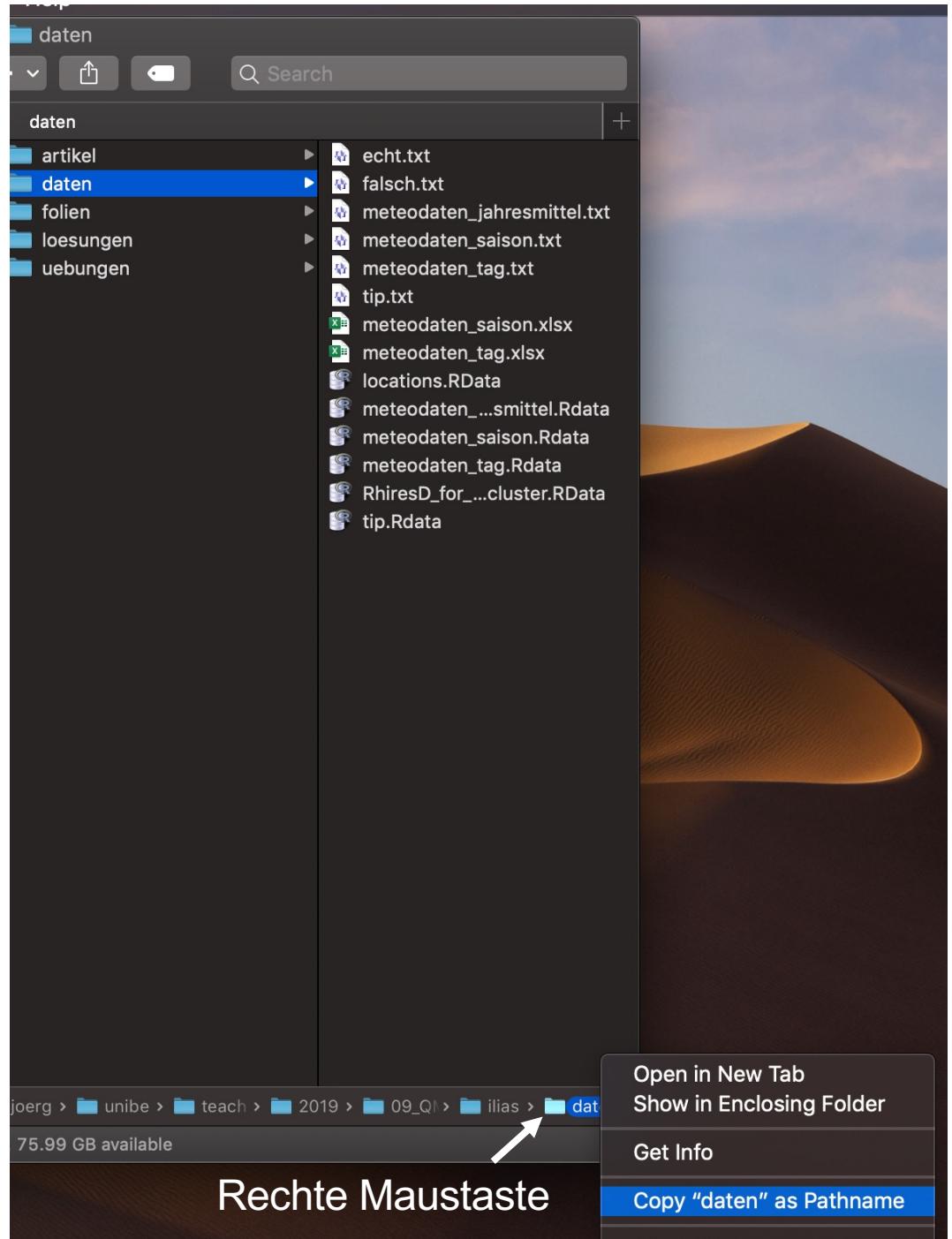
1. Im Datei "Explorer" (Windows) oder "Finder" (Mac) das Verzeichnis finden, wo ihr eure Dateien zu dieser Veranstaltung speichert
2. Beispieldaten von Ilias in dieses R Verzeichnis speichern/runterladen
3. Vollständigen Pfadnamen kopieren (siehe Abb. rechts)

Mac: '/Users/name/quant_meth...')

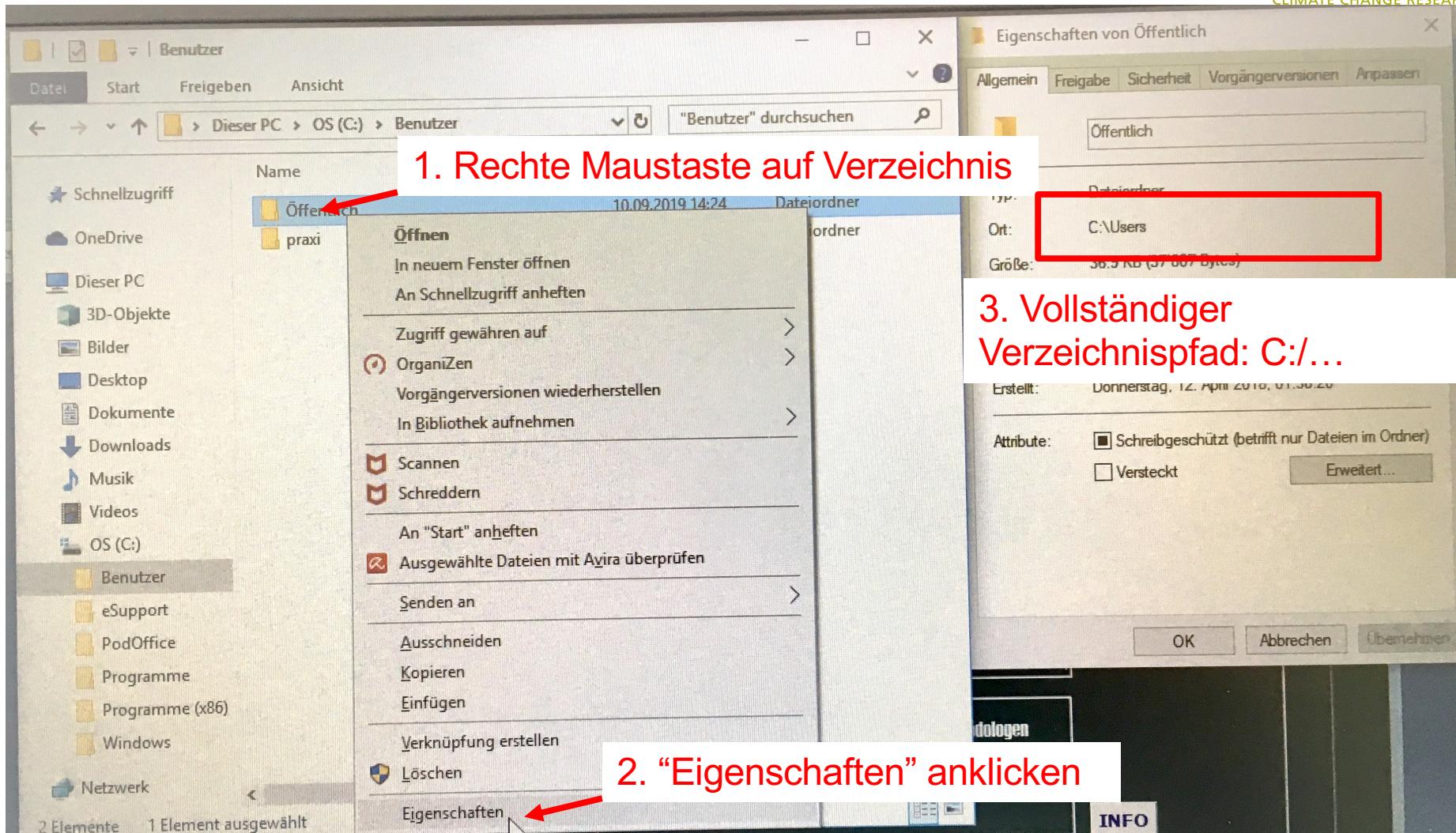
bzw.

Windows: 'C:/Users/...'

ACHTUNG bei Windows wird \ benutzt und muss für R in / verwandelt werden!



Arbeitsverzeichnis im Windows Explorer



Daten einlesen

Dateien im Verzeichnis anzeigen

```
> list.files('/Users/name/quant_meth...')
```

sollte alle Dateien und Unterverzeichnisse im angegeben Verzeichnis anzeigen

ASCII Daten einlesen:

Textdatei 'file.txt' bzw. 'file.csv' aus Excel exportiert

```
> saison <- read.table('/.../.../.../meteodaten_saison.csv',  
    # Pfad zur Datei angeben in ''  
    header=...,          # Erste Zeile Spaltennamen? TRUE oder FALSE  
    na.strings='..',      # wie sind Fehlwerte kodiert? z.B. '-99', NA  
    sep='... '           # optional: wie sind Werte separiert? wenn durch  
    # tabs getrennt nicht notwendig,  
    # sonst z.B. ',' für .csv (Comma separated values)
```

Wie sehen eingelesene Daten aus?

- > *str(saison)* ... zeigt die Struktur des Objekts
- > *head(saison)* ... zeigt die ersten paar Zeilen
- > *tail(saison)* ... zeigt die letzten paar Zeilen
- > *summary(saison)* ... einige statistische Grunddaten
- > *saison[1:10,]* ... zeigt die ersten 10 Zeilen an

Auswahl von Elementen/Spalten/Zeilen

- > *saison[1,1]*
- > *saison[23:24, 'SPALTENNAME']*
- > ...

Euer Skript startet mit:

```
> saison <- read.table('/euer_r_verzeichnis/meteodaten_saison.csv',  
sep=",", header=T) # 'C:/...' unter Windows  
  
> head(saison) # zeigt erste Zeilen der geladenen Daten an  
  
> str(saison) # zeigt an, ob Daten korrekt als numerisch gelesen wurden
```

Skript speichern nicht vergessen!

Auf korrekte GROSS- und klein-Schreibung achten!

Programmierempfehlungen

Kommentare, falls ihr nicht mit “Notebooks/Markdown” arbeitet

```
> # Zeitreihe der Sommertemperatur in Genf, geglättet mit 31-jährigem laufenden  
Mittelwert  
> plot(running.mean(saison[saison[,2]=='Sommer JJA',3],31),ty='l,col='red)
```

Selbsterklärende Objektnamen

```
JJA_temp_genf_31yr-smooth <- running.mean(saison[saison[,2]=='Sommer  
JJA',3],31)  
> plot(JJA_temp_genf_31yr-smooth ,ty='l,col='red)
```

Struktur mit Zeileneinschüben, Leerzeichen, etc.:

```
> for ( i in 3:5 ) {  
    print ( i + 1 )  
}
```

u^b

b
**UNIVERSITÄT
BERN**

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

ÜBUNGEN 1

R-Übungen 1

ACHTUNG: wenn + statt > am Zeilenanfang steht, ist eine Funktion nicht beendet, d.h. ',),]' oder } fehlen



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

1.1) Vektoren: Überlegt euch die erwarteten Lösungen vor dem Eintippen!

```
x <- c(5,2,1,4)  
xx <- c(1,10,15,18)  
y <- rep(1,5)  
z <- c(TRUE, FALSE, TRUE, TRUE)
```

- a) sum (x)
range(x)
length(x)
sum(x)
- b) c(x,y,13)
- c) x[4] * y[2]
xx[2:4] + x[1:3]
- d) xx <= 12
xx [xx <=12]

- e) plot(x,xx)
plot(x[z],xx[z])

1.2) Zahlenfolgen: Erzeugt mit den *rep* und *seq* Funktionen folgende Zahlenfolgen:

- a) 1 2 3 4 5 6 7 8 9
- b) 'm' 'w' 'm' 'w' 'm' 'w'
- c) 1 2 3 4 1 2 3 4 1 2 3 4
- d) 1 2 2 3 3 3 4 4 4 4

1.3) Lest die Datei
"meteodaten_saison.csv" in R ein:

```
> saison <- read.table('/pfad/.../  
meteodaten_saison.csv', sep=',',  
header=TRUE)
```

Überprüft, ob der Import korrekt verlief?

Welche Spalten sind chr, int, num, ...?

```
> head(), str(), summary(), tail(), class()...
```

Prüfung

- > Montag, 10.02.2025
- > 14:15 – 15:15
- > Hauptgebäude, Aula 210

- > Wer ein Auslandssemester macht, bitte persönlich melden

Daten auswählen aus data.frame bzw. matrix

Auswahl in kleinerem data.frame speichern

```
> sommer <- saison[saison[,2]=='Sommer(JJA)',]  
# Objekt saison, aber nur Zeilen, wo in 2. Spalte 'Sommer (JJA)' steht
```

Einzelne Variable/Spalte auswählen und in Vektor speichern

```
> sommer_temp_genf <- sommer[,3] oder  
> sommer_temp_genf <- sommer[, 'Genf_Mitteltemperatur']  
  
> jahre <- sommer[,1]
```

Rechnungen über ganze Spalten / Zeilen mit APPLY

> *apply(saison[,3:6], 2, mean)* ...2 = spaltenweise (1 wäre zeilenweise)

<i>Genf_Mitteltemperatur</i>	<i>Genf_Niederschlagssumme</i>	<i>GrStBernhard_Mitteltemperatur</i>	<i>GrStBernhard_Niederschlagssumme</i>
9.732662	234.364206	-1.182476	549.379418

Nur für Auswahl (z.B. Sommersaison; siehe vorherige Folie)

> *apply(saison[saison[,2]=='Sommer(JJA)',3:6], 2, mean)*

<i>Genf_Mitteltemperatur</i>	<i>Genf_Niederschlagssumme</i>	<i>GrStBernhard_Mitteltemperatur</i>	<i>GrStBernhard_Niederschlagssumme</i>
18.168750	254.375893	5.987202	432.576786

oder direkt mit dem neu erzeugen data.frame 'sommer':

> *apply(sommer[,3:6], 2, mean)*

oder einfach das Mittel des vorher erzeugten Vektors 'sommer_temp_genf':

> *mean(sommer_temp_genf)*

Rechnungen nach Klassen

Funktion **aggregate()**

Syntax: **aggregate(numerischer Vektor(en), list(Klassen), FUN= Funktion)**

> **s.agg <- aggregate(saison[,3:6], list(saison[,2]), FUN = mean)**

Group.1	Genf_Mitteltemperatur	Genf_Niederschlagssumme	GrStBernhard_Mitteltemperatur
1 Fruehling(MAM)	9.369940	209.3482	-3.325595238
2 Herbst(SON)	9.905706	264.9090	-0.006306306
3 Sommer(JJA)	18.168750	254.3759	5.987202381
4 Winter(DJF)	1.487798	209.0964	-7.374702381

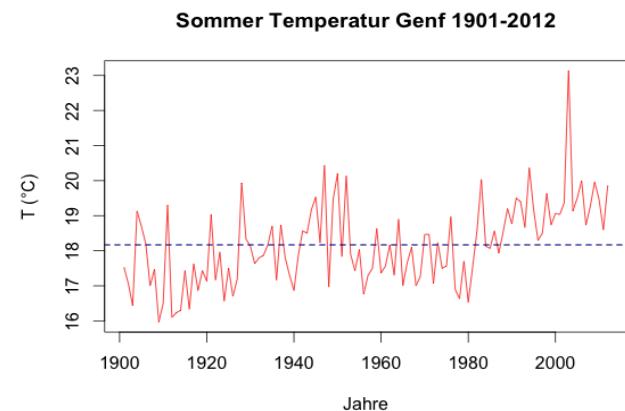
	GrStBernhard_Niederschlagssumme
1	596.0366
2	558.9613
3	432.5768
4	610.0286

Grafiken in R

„High level‘ Plot Funktionen – „zaubern aus allem eine Grafik‘, jedoch nicht immer nach Wunsch, je nach Daten/Objekt

plot(), hist(), boxplot(), pairs(), ...

```
> plot(saison)  
> plot(saison[, 'Genf_Mitteltemperatur'])  
> plot(sommer[,3])
```



„Low level‘ Plot Funktionen – kann man einer bestehenden Grafik nur hinzufügen

lines(), abline(), points(), ...
> abline(h=18)

Funktion ***plot(x,y)***

```
> ?plot
```

```
> plot(sommer[,1],sommer[,3])
```

```
> plot(sommer[,1],sommer[,4],
```

```
      type ='l',
```

....Darstellung: Punkte „p', Linie „l',

```
      xlab='Jahr',
```

....Beschriftung x-Achse,

```
      ylab='Niederschlag (mm)', ....Beschriftung y-Achse,
```

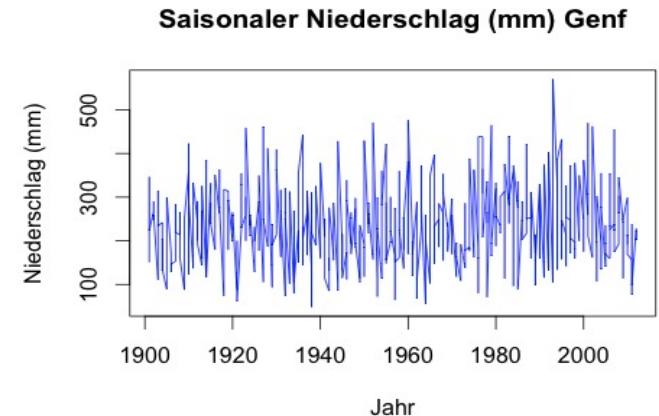
```
      main='Saisonaler Niederschlag (mm) Genf', ...Überschrift
```

```
      col='blue',
```

...Farbe z.B. „green', 'red', ...

```
      ylim=c(80,600), xlim=c(...,...),
```

```
      ...)
```



weitere Grafik-Parameter findet ihr mit: **?par**

Funktion ***boxplot(y~x)***

> *?boxplot*

> *boxplot(saison[,3] ~ saison[, 'Saison'])*

... *boxplot(daten~gruppeninfo)*

> *boxplot(saison[,3] ~ saison[,2],*

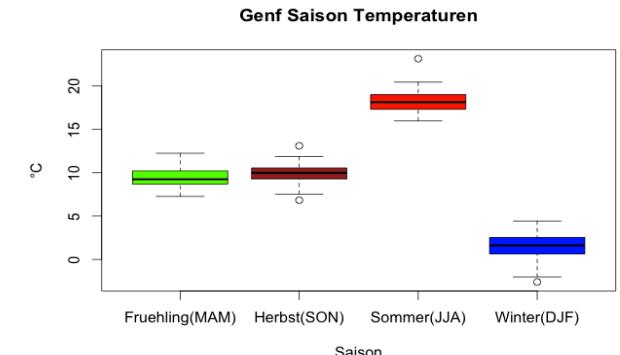
main='Genf Saison Temperaturen', ... Überschrift

ylab='° C', , ... Beschriftung y-Achse,

xlab='Saison', , ... Beschriftung x-Achse,

col=c('green', 'brown', 'red','blue') ... Farbe(n)

...)



Funktion *hist(x)*

```
> ?hist  
> hist(sommer[,3])  
  
> hist(sommer[,3],  
       main='Histogramm',  
       ylab='Häufigkeiten',  
       xlab='° C',      ,  
       col='green',  
       breaks=20,  
       ...)
```

... Überschrift
... Beschriftung y-Achse,
... Beschriftung x-Achse,
... Farbe(n)
... ungefähre Anzahl Klassen

Funktion **barplot(x)**

```
> ?barplot  
> s.agg <- aggregate(saison[,3:6], list(saison[,2]), FUN = mean)  
                                ... von Folie mit aggregate() Funktion  
> barplot(s.agg[,2])  
  
> barplot(s.agg[,3],  
          names.arg=s.agg[,1],  
          main='Barplot',  
          ylab='° C',  
          xlab='Saison',    ,  
          col=c('green', 'brown', 'red','blue'), ... Farbe(n)  
          ...)
```

... Beschriftung der bars

... Überschrift

... Beschriftung y-Achse,

... Beschriftung x-Achse,

... Farbe(n)

Hinzufügen von „low-level“ Plot Funktionen

zuerst High-Level Funktion z.B. **plot(x,y)**, Grafik öffnet sich

> **plot(sommer[,1], sommer[,3], type='l', ...)**

Hinzufügen von „low-level“ Funktionen bei offener Grafik

> **abline(h=20)** ... horizontale Linie bei 20 (auch **vertikal**)
> **abline(h=mean(sommer[,3]), col='blue')** ... horiz. Linie bei Mittelwert
> **points(sommer[,1], sommer[,3], pch=2)** ... Punkte hinzufügen
 ... Punkt-typ
> **lines(x,y,...)**
> **legend(...)** ... Legende hinzufügen

Grafikfunktion **par(...)** vor plots

```
> ?par          ... Grafikparameter für alle Grafiken festlegen  
> par(mfrow=c(2,1))    ... plot mit 2 Grafiken untereinander (2 Zeilen)  
> barplot(s.agg[,2] names.arg=s.agg[,1])  
> plot(sommer[,1], sommer[,3], type='l', xlab='Jahr', col=„red“)
```

Legende hinzufügen, nach plots (low-level Funktion), **?legend**

```
> legend('topleft',  
         legend='Sommer',  
         col='red',  
         lty=1,      ... wenn Linie->Linientyp (1=durchgezogen)  
                   wenn Symbol: pch (pointcharakter)=(z.B.) 1  
         cex=0.8, ...)        ... Grösse der Legende (relative zu 1)
```

u^b

b
**UNIVERSITÄT
BERN**

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

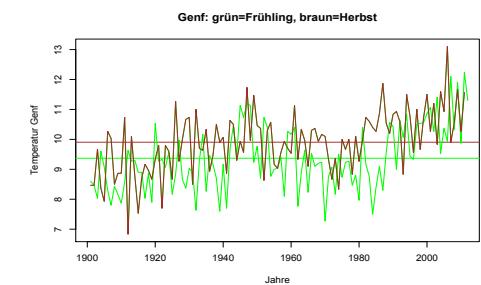
ÜBUNGEN 2

R-Übungen 2

ACHTUNG: immer Teil des Codes einzeln ausführen, wenn etwas nicht funktioniert! Hier z.B.:
`saison[,2]=='Fruehling(MAM)'`

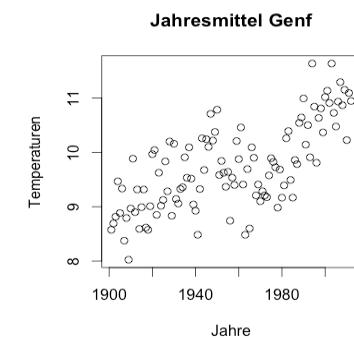
2.1) Grafik erstellen

- > Extrahiert aus den saisonalen Daten erstens nur die Frühlingsdaten, zweitens nur die Sommerdaten und drittens nur die Herbstdaten z.B.
`fruehling <- saison[saison[,2]=='Fruehling(MAM)',]`
- > Erstellt einen Plot, mit den Jahren auf der x-Achse und der Temperatur in Genf auf der y-Achse. Stellt dabei die **Frühlings-, Sommer- und Herbsttemperaturen** als Linien mit unterschiedlichen Farben im gleichen Plot dar:
 - > zuerst `plot(x,y,col=' ', xlab=' ',...)`
 - > dann mit `lines(x,y,col=..)` weitere Saisons
 - > Vergebt eine Überschrift und beschriftet auch beide Achsen
 - > Fügt Linien der beiden Mittelwerte hinzu mit `abline(h=...)`

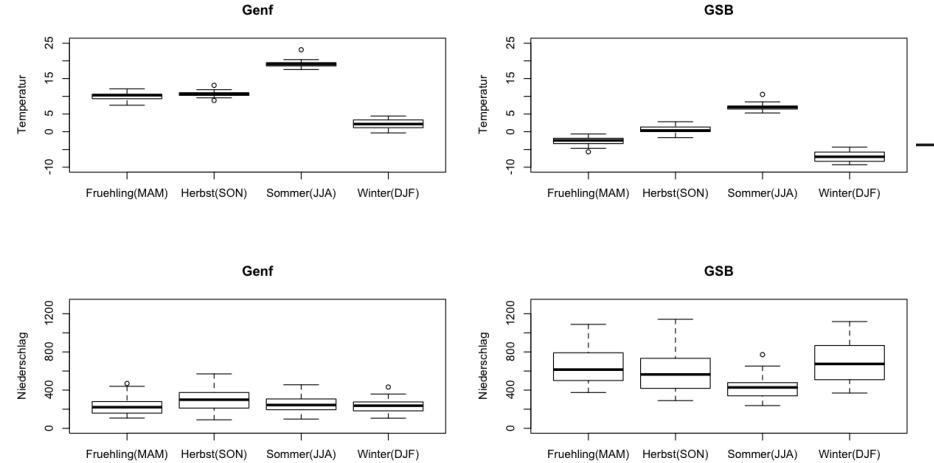


2.2)

- > Erstellt mittels `aggregate()` die **Jahresmittelwerte** der Temperatur für Genf und stellt diese in einem Scatterplot mit Punkten `plot(x,y)` dar.
- > Beschriftet die Achsen und vergeb einen Titel.



R-Übungen 2



2.3)

- > Wählt den **Zeitraum 1981-2010**, z.B. `zeit <- saison[,1]>=1981 & ...`
- > Stellt die Temperatur- und Niederschlagsverteilungen der Saisons in Genf und Gr. S. Bernhard für diesen Zeitraum in vier `boxplot()` dar.
- > Das Grafikausgabefenster kann mit `par(mfrow=c(2,2))` in 2 Zeilen und 2 Spalten geteilt werden.
- > Beschriftet wieder die Achsen und vergeb Titel. Achtet darauf, für beide Stationen gleiche y-Achsen zu wählen, so dass die Plots gut visuell vergleichbar sind, z.B. bei Niederschlag je `ylim=c(0,1300)`
- > Das Grafikausgabefenster mit `par(mfrow=c(1,1))` wieder auf 1 Zeilen und 1 Spalten zurücksetzen.

u^b

^b
**UNIVERSITÄT
BERN**

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Graphik speichern

als eps, jpg, pdf,... '**plot.pdf**'

in R-Studio:

'Export' ... 'Save Plot as PDF' oder 'Image' ... (bei 'Image' gewünschtes Format wählen)... 'Directory' angeben (=Pfad) oder manuell

- > **pdf('name.pdf', width=9, height=4.5, ...)** ... oder
- > **png(...)** ... zum Grafikdatei anlegen/öffnen, je nach Formatwunsch
- > **plot(...)** ... *Grafik erstellen*
- > **dev.off()** ... *Grafik abschliessen*

Daten speichern

als ASCII Textdatei 'file.txt'

```
> write.table(saison, ... welches Objekt soll gespeichert werden  
  file='.../.../.../file.txt') ... Pfad und gewünschten Dateinamen angeben
```

Es gibt spezielle Funktionen zum Lesen und Schreiben vieler weiterer Datenformate. Diese findet ihr bei Bedarf leicht im Internet.

Umgang mit Fehlwerten in R

R kodiert Fehlwerte mit 'NA'

bei Funktionen:

> *a* <- *c(1, 3, 4, NA, 5)*

> *sum(a)*

[1] NA ...man muss angeben wie mit NAs umgehen will mit **na.rm**

> *sum(a, na.rm=TRUE)* ...z.B. na.rm (NA remove) TRUE

[1] 13 ... NAs werden ignoriert

auch für *mean()*, *sd()*, *apply()*, *aggregate()*,...

beim Daten einlesen:

> *x* <- *read.table('data.txt', na.strings=...)* ..angeben wie NAs im
einzulesenden Datensatzkodiert sind z.B. „-999“, „NA“...

oder nach einlesen mit *replace(x, x==...,NA)*

Umgang mit Fehlwerten in R

ABER: Abfrage von Fehlwerten mit „is.na“

```
> a <- c(1, 3, 4, NA, 5)
```

```
> is.na(a)
```

```
[1] FALSE FALSE FALSE TRUE FALSE
```

...nicht a==NA

...logischer Vector

```
> which(is.na(a))
```

```
[1] 4
```

...Vektor mit Elementnummer

umgekehrt

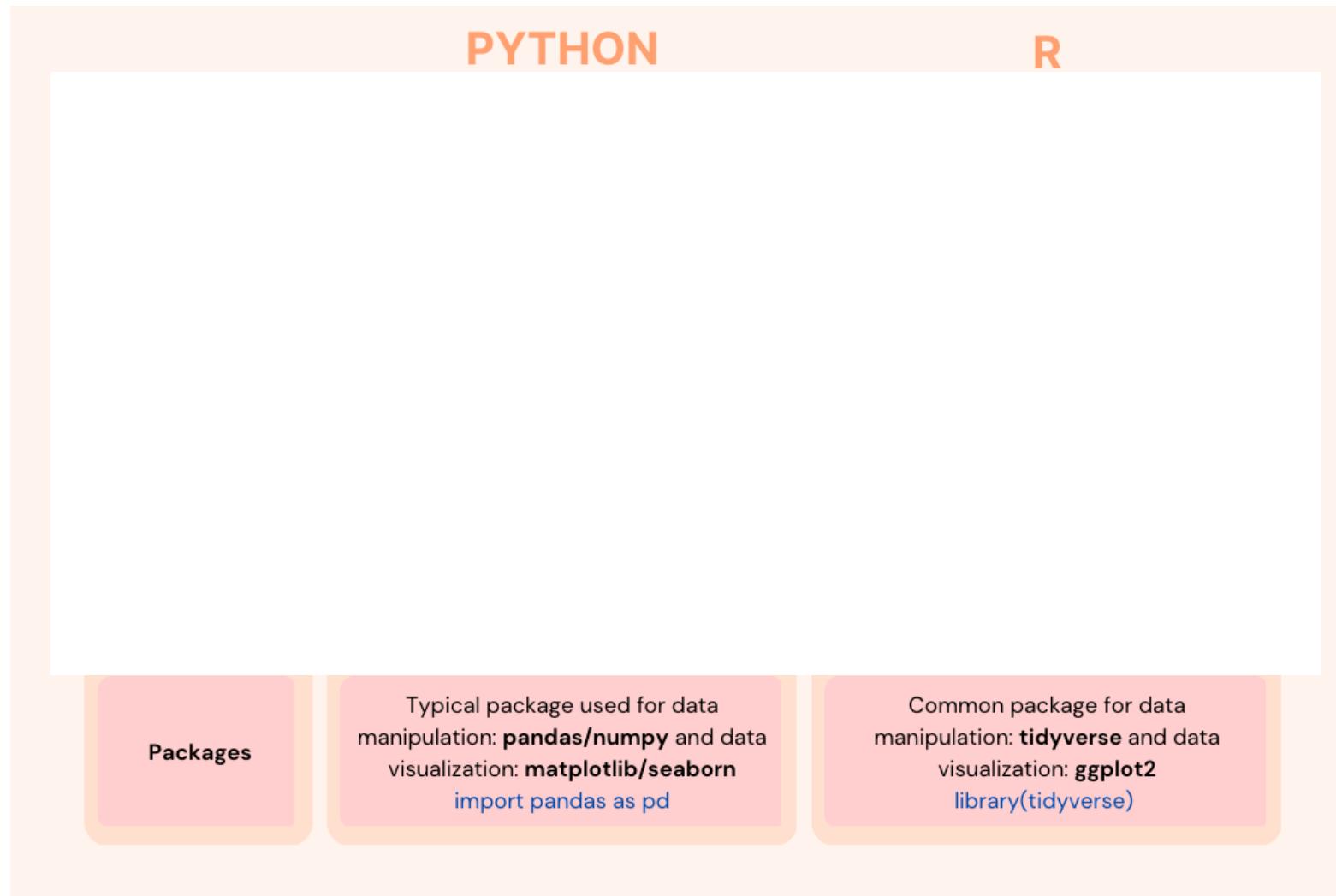
```
> !is.na(a)
```

```
[1] TRUE TRUE TRUE FALSE TRUE
```

...! bedeutet NICHT

```
> which(! is.na(a))
```

Python vs R



<https://towardsdatascience.com/the-starter-guide-for-transitioning-your-python-projects-to-r-8de4122b04ad>

Packages/Libraries (Pakete/Bibliotheken) installieren und laden

Am Beginn ins Script schreiben welche Pakete benötigt und geladen werden sollen (z.B. `library('lattice')`)

Dazu ist einmalige Installation auf Computer notwendig:

> `install.packages('lattice', dependencies=TRUE)` ... nur 1 x, dann auf Computer installiert

Dann am Anfang von jedem Skript laden, welches die Funktionen der Bibliothek nutzen soll:

> `library(lattice)` ... Paket laden (in jeder Session!)

Hilfe zur Bibliothek:

> `library(help=lattice)` ... Hilfe/Information zu package
> `?lattice` ... Hilfe/Information zu package
> `demo(lattice)` ... einige packages haben demo files

Tidyverse Library

```
> install.packages("tidyverse")
> library("tidyverse")
```

Objekt type “tibble” einlesen, ähnlich data.frame()

```
> saison=read_csv("meteodaten_saison.csv")
```

Datentransformationen

```
> winter <- filter(saison, Saison == "Winter(DJF)") # auswählen
> arrange(saison, desc(Jahr)) # sortieren
> select(saison, Jahr, Saison, Genf_Mitteltemperatur)
```

Pipes

```
> saison = saison %>% mutate(temp_diff_genf_san_bern = Genf_Mitteltemperatur -
  GrStBernhard_Mitteltemperatur)
```

Python vs R

PYTHON	R
<p>Chaining operations</p> <p>Method chains with the "." operator <code>df.head(n)</code> or <code>df.describe()</code></p>	<p>Can chain operations using following symbol: "%>%" <code>df %>% head(n)</code> or <code>df %>% summary()</code></p>

<https://towardsdatascience.com/the-starter-guide-for-transitioning-your-python-projects-to-r-8de4122b04ad>

ggplot()

```
> ggplot(data = winter) +  
  geom_point(mapping = aes(x = Jahr, y = Genf_Mitteltemperatur))  
  # Fügt Punkte hinzu
```

Weitere “aes” Parameter:

color = Saison ; size = ... ; alpha = ... ; shape = ...

Mehrere Abbildungen kombinieren:

```
> ggplot(data = saison) +  
>   geom_point(mapping = aes(x = Jahr, y = Genf_Mitteltemperatur, color=Saison)) +  
>   facet_wrap(~ Saison, nrow = 2, ncol = 2)
```

ifelse – Bedingte Anweisung für Vektoren

> **ifelse** (Abfrage , wenn TRUE, wenn FALSE)



Operation für alle Elemente die Abfrage erfüllen

Operation für alle Elemente die Abfrage NICHT erfüllen

z.B.:

> **a <- c(5, 7, 10)**

> **ifelse (a==7, a+1, a-1)**

> [1] 4 8 9

if – Bedingte Anweisung für einzelne Elemente, keine Vektoren

> *if* (Bedingung) { was ausführen wenn Bedingung TRUE }

> *if* (Bedingung) {
 was ausführen wenn Bedingung TRUE
} *else* {
 was ausführen wenn Bedingung FALSE
}

z.B.:

> *b* <- 5
> *if* (*b*==10) { *b* <- *b*+1 } *else* { *b* <- *b*-1 }
> *b*
[1] 4

Schleifen (loops): z.B. *for()* – Schleife

wenn gleiche Operationen mehrmal vorgenommen werden
müssen z. B. in Algorithmen

> *for* (beliebiger Name *in* Vektor angeben) { *Operationen* }

für alle „beliebiger Name“ (z.B. „i“) aus dem angegebenen „Vektor“ wird
nacheinander eine Operation durchgeführt

z.B.:

> *for* (*i* *in* 3:5) { *print* (*i*+1) }

[1] 4

[1] 5

[1] 6

Schleifen (loops): z.B. *for()* – Schleife mit Laufindex

z.B.:

```
> x <- c(8, 10, 12)
```

```
> xx <- vector() # leeren Vektor für Ergebnisse definieren
```

```
> j=1 # laufender Index (ausserhalb Schleife!)
```

```
> for ( i in x) {  
  xx [j] <- ( i+1 )  
  j=j+1  
}
```

```
> xx  
[1] 9 11 13  
> j  
[1] 4
```

eigene Funktionen I

> *Funktionsname <- function (Argument 1, Argument 2, ...)* Ausdruck

> *Funktionsname <- function (Argument 1, Argument 2, ...){*

mehrere Ausdrücke

...

}

eigene Funktionen II

z.B.: Funktion die 2 Werte multipliziert

```
> myFunc <- function (x, y) {  
  w <- x * y  
  return(w)      # return() gibt Ergebnis aus  
}
```

```
> myFunc(1,2)      # Funktion aufrufen mit Argumenten  
[1] 2
```

eigene Funktionen III

z.B.: Funktion die nur alle positiven Werte eines Vektors ausgibt

```
> myFunc2 <- function (x) {  
  y <- x[x>0]  
  return(y)      # return() gibt Ergebnis aus  
}
```

```
> myFunc2(c(1,2,-1,4))    # Funktion aufrufen mit Argumenten  
[1] 1 2 4
```

!! Funktionen kann man in eigenem R-script speichern ('auslagern') und mit **source()** in Arbeitsskript laden

Python vs R

	PYTHON	R
Assigning a Variable	Use assignment operator "=" <code>x = 1</code>	Use assignment operator "<- " or "=" <code>x <- 1 or x = 1</code>
Data Structure Difference	Python list datatype is equivalent to vector datatype in R <code>ex_list = [1, 2, 3, 4, 5]</code>	Common data type in R is a vector <code>ex_vector <- c(1, 2, 3, 4, 5)</code>
Indexing and slicing	Indexing starts from 0, inclusive of the start index and excludes the end index. <code>ex_list[0] and ex_list[0:2]</code> <code>output: 1 output: [1, 2]</code>	Indexing starts from 1, Indexing is inclusive of both start and end index <code>ex_vector[1] and ex_vector[1:2]</code> <code>output: 1 output: 1, 2</code>
Chaining operations	Method chains with the "." operator <code>df.head(n) or df.describe()</code>	Can chain operations using following symbol: "%>%" <code>df %>% head(n) or df %>% summary()</code>
Packages	Typical package used for data manipulation: pandas/numpy and data visualization: matplotlib/seaborn <code>import pandas as pd</code>	Common package for data manipulation: tidyverse and data visualization: ggplot2 <code>library(tidyverse)</code>

Chatbots

Chatbot Einsatz an Uni allgemein: Was habt ihr schon lernt und selber gemacht?

Chatbots

- > Besser Fragen stellen lassen nachdem ein virtueller Gesprächspartner definiert wurde als Antworten geben lassen
- > Brainstorming
- > Text überarbeiten lassen, auch Sprachniveau, Stil, etc.
- > **immer kontrollieren, ihr seid verantwortlich!**

Chatbots - Prompting

- > • **Kontext:** Konkretisieren Sie, in welchem Zusammenhang Ihre Anfrage steht. Durch zusätzliche Informationen können Genauigkeit und Qualität der vom Chatbot generierten Ausgaben verbessert werden.
- > • **Zielgruppe:** Identifizieren Sie die Zielgruppe und berücksichtigen Sie deren Kenntnisse und Erfahrungen. Geben Sie diese Informationen in klarer und verständlicher Sprache an.
- > • **Zielsetzung:** Beschreiben Sie präzise, welches Ziel mit dem Prompt erreicht werden soll und überlegen Sie sich, welche Art von Informationen der Chatbot hierfür benötigt.
- > • **Datenquelle:** Sollten Sie zusätzliche Daten eingeben, stellen Sie sicher, dass die Daten von hoher Qualität und Präzision sind.

Chatbots - Programmieren

1. Interaktion mit einem Chatbot

- > Fragen zum Programmieren, Funktionen oder Syntax stellen
- > Beispiel: “Wie erstelle ich eine Weltkarte mit Kontinentgrenzen in R?” oder “Was entspricht einem Python dictionary in R?”

2. Chatbots um Fehlermeldungen zu verstehen

- > Fehlermeldungen oder Code der Problem verursacht einfügen und nach Hilfe fragen.
- > Dies führt meist zu leichter verständlichen Fehlerbeschreibungen.

3. Lernen

- > Chatbot nach Empfehlungen für gutes Codieren fragen oder nach effizienterer Variante.

u^b

b
**UNIVERSITÄT
BERN**

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

ÜBUNGEN 3

Übungen 3.1

- > Bereche die Sommer (JJA) Temperaturanomalien zur Referenzperiode 1961 bis 1990 in Bern mit Excel.
- > Schreibe R Code, um die gleiche Berechnung durchzuführen.
- > Lass Dir mit einem Chatbot den R Code schreiben,
- > Neuerdings können Chatbots direkt Datenanalysen ohne Programmierkenntnisse durchführen:
<https://help.openai.com/en/articles/8437071-data-analysis-with-chatgpt>
Überprüfe die Ergebnisse und den generierten Code.
- > Diskutiere die Vorteile, Nachteile und Risiken aller vier Methoden.

Übungen 3

mit Standard R (Beispiele unten) oder
tidyverse und ggplot mit Chatbot Hilfe



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

3.2) Klimadiagramm

- > Ladet den Datensatz meteodataen_tag.csv nach dem Excel Export in R
(ACHTUNG: NA-Werte sind sowohl mit '-' als auch mit 'NA') kodiert, deshalb:
`na.strings = c('-', 'NA')`
- > Mit `str()` ansehen, ob Daten korrekt (z.B. als **numerisch**) gelesen wurden.
- > Erstellt ein Histogramm (`hist()`) mit den Tagestemperaturen mit feinen Abständen
(`breaks=40`).
- > Wie sieht die Verteilung nach Augenmaß aus?
- > Berechnet die Monatsmittelwerte der Temperatur und der Bewölkung über alle Jahre (also Mittel über alle Jan, alle Feb,... wie in Klimadiagrammen). Achtung: Fehlwerte vorhanden!
- > Erstellt in eine Abbildung mit zwei Barplots der Ergebnisse übereinander
(`par(mfrow=c(2,1))`). Was erwartet ihr?

Übungen 3

3.3) Boxplots

- > Wählt den Zeitraum **2000-2001** in den täglichen Daten.
- > z.B. `zeit <- meteodaten_tag[,1] >= 2000 & meteodaten_tag[,1] <= 2001`
- > Stellt die Temperaturen dieses Zeitraumes als Funktion der Bewölkung in einem `boxplot()` dar (je einen Boxplot pro Bewölkungsklasse).
- > Beschriftet die Achsen und vergeb einen Titel.
- > Unter welchen Bewölkungs- bedingungen ist die Spannweite und Varianz der Temperatur am grössten?
- > Findet heraus welcher Monat im Mittel der bewölkungsärmste und der -reichste Monat ist (im Mittel der 2 Jahre). Wieviel Bewölkung gibt es im Mittel in diesen Monaten (in Octas)?

Übungen 3

3.4) R als GIS Ersatz

- > Installiert das Paket 'maps' und ladet es in R (z.B. library(maps)). Findet die x,y-Koordinaten von Genf und dem Gr. S. Bernhard heraus (ungefähr).
- > Versucht eine Europakarte herzustellen und Genf und G.S.Bernhard als Punkte auf der Karte zu plotten und die Punkte mit Stationsnamen zu versehen:

```
# zuerst einen leeren plot erstellen mit
# Koordinaten von Europa.
> plot(x=c(-5,30), y=c(35,60),
       type='n', xlab='lon', ylab='lat')
# type=' n" heisst es wird nichts geplottet
> map("world",add=TRUE)
# jetzt die Stationen als Punkte dazu (evtl.
# in verschiedenen Farben, mit
# unterschiedlichen Symbolen und mit
# Text!? Verwendet Google!!)
> points(...)
```

u^b

^b
**UNIVERSITÄT
BERN**

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Take-home messages

- > Insbesondere für grosse Datensätze viel besser als Tabellenkalkulation
- > Erzeugt mit einer Zeile Code grundlegende statistische Auswertungen und Abbildungen wie Histogramme und Test, die mit Excel kompliziert oder gar nicht möglich sind
- > Bei Fragen Internetsuchmaschine oder Chatbot gebrauchen
- > Programmieren muss gar nicht so kompliziert sein
- > ACHTUNG: man bekommt (fast) immer etwas heraus, stellt sicher, dass es auch das Richtige ist!
- > In der Prüfung müsst ihr nicht alle Funktionen und deren Syntax in R kennen, aber einfachen Code verstehen können und beschreiben können wie ein Problem lösen würdet, z.B. Scheife über alle Stationen, um statistischen Test an allen Orten durchzuführen.

Beispiel Prüfungsfrage

- > Was berechnet folgende R Funktion (Modus, Mittelwert, Median, Minimum oder Maximum):

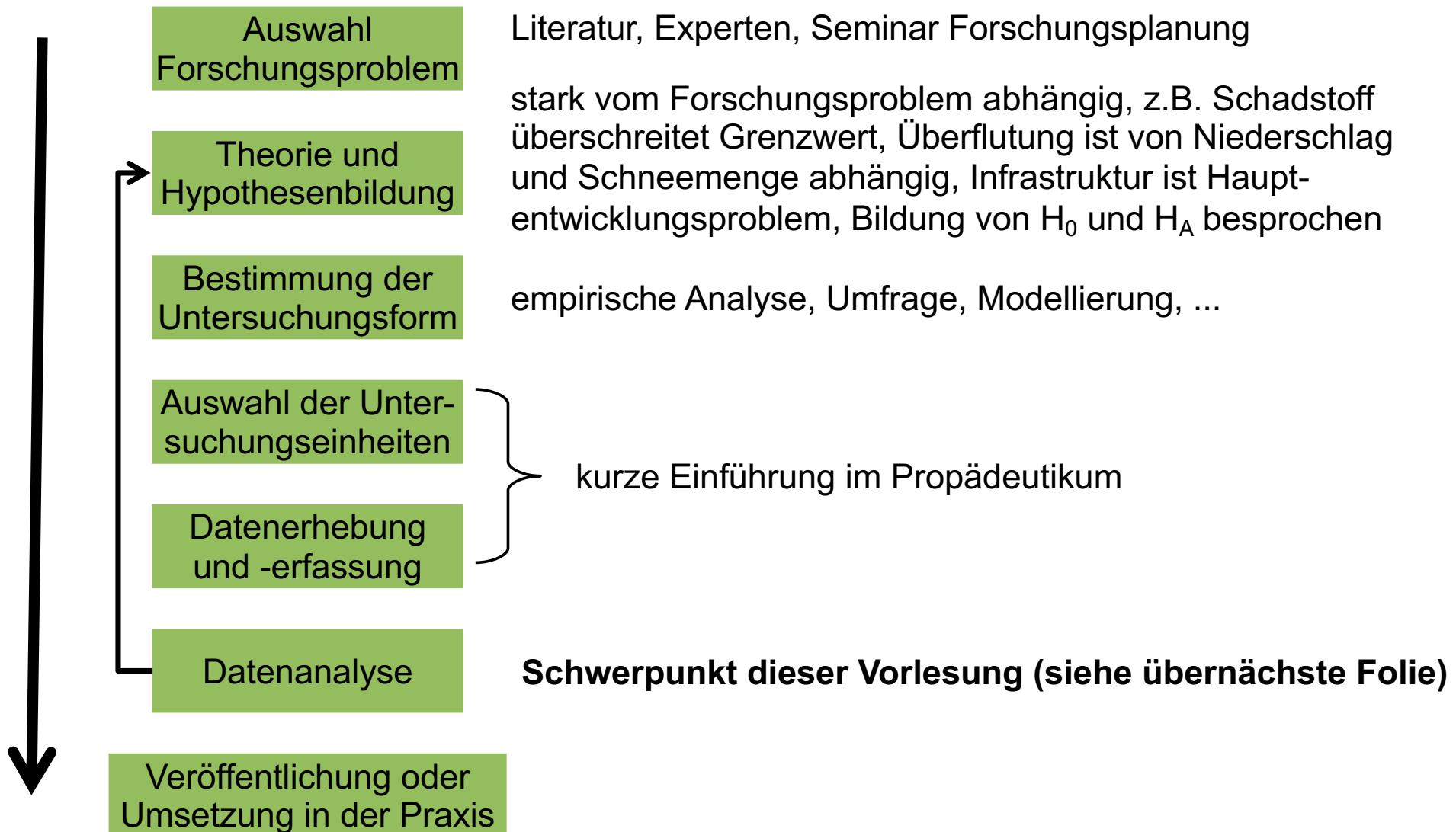
```
# Funktion zur Berechnung des ??? eines Vektors
berechne_? <- function(vektor) {
  # Vektor sortieren
  sortierter_vektor <- sort(vektor)
  # Länge des Vektors
  n <- length(sortierter_vektor)
  # ? berechnen
  if (n %% 2 == 1) { #
    ergebnis <- sortierter_vektor[(n + 1) / 2]
  } else {
    ergebnis <- (sortierter_vektor[n / 2] + sortierter_vektor[n / 2 + 1]) / 2 }
  return(ergebnis)
}
```

Quantitative Methoden in der Geographie

Herbstsemester 2024

Jörg Franke

Übersicht Forschungsprozess



Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
Fallen der Statistik	Statistische Tests Konfidenzintervalle	Korrelation	Regression
weiterführende Methoden	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen, Modellvalidierung
Daten zusammenfassen	Hauptkomponenten-analyse	Clusteranalyse	Extremwertstatistik
			Zeitreihenanal. etc.

Deskriptive Verfahren

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Ziel: Daten strukturieren, beschreiben

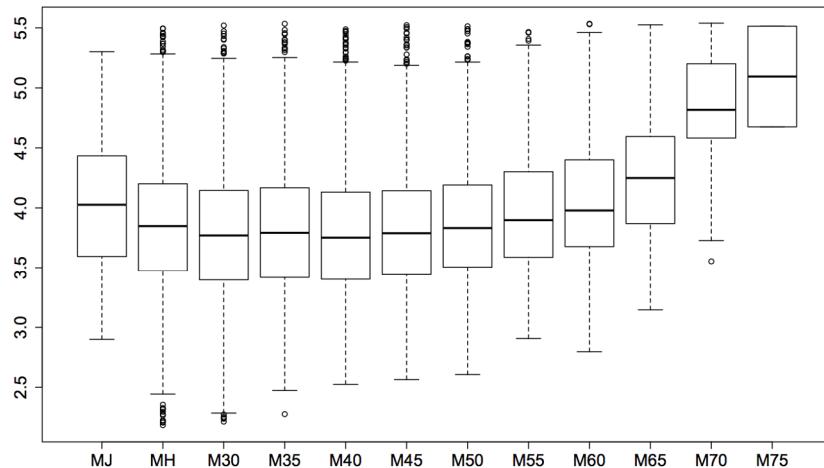


Abbildung 2.11: Box-Whisker-Plot von Laufzeit versus Altersklasse (Männer, Beispiel 2.4).

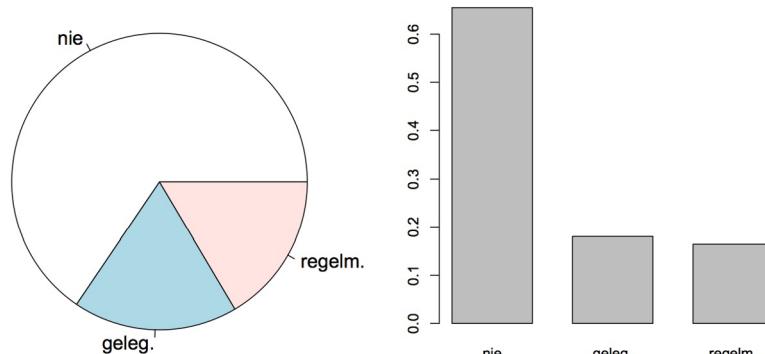
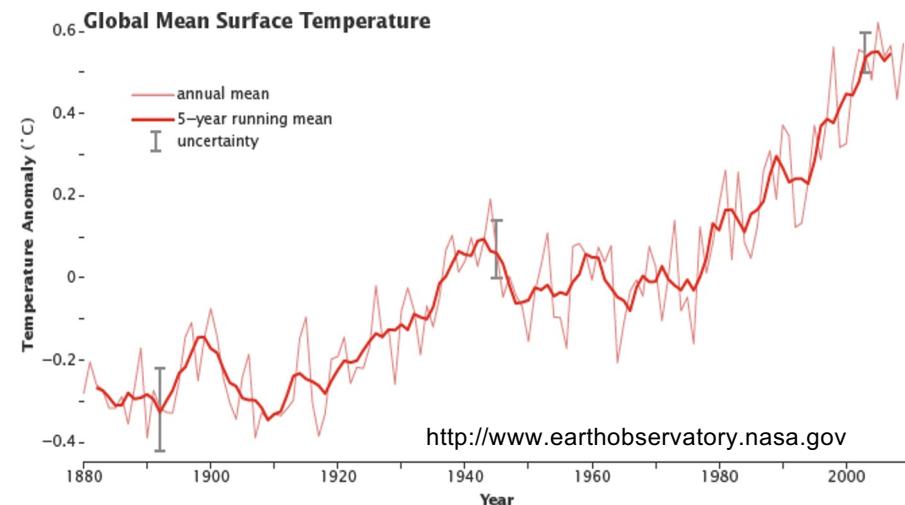


Abbildung 2.1: Kuchen- und Stabdiagramm der Variable Rauchen für Beispiel 2.1.

Das Auge erkennt sehr gut Strukturen
in den Daten!



Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik

Rohdaten
visualisieren

Datenqualität
prüfen

statistische
Masszahlen

Schliessende Statistik

Unterschiede
identifizieren

Zusammenhänge
identifizieren

Abhängigkeiten
modellieren

Statistische Tests
Konfidenzintervalle

Korrelation

Regression

Wie wahrscheinlich
sind die Daten der
Stichprobe, wenn
die Nullhypothese
zutrifft?

Gibt es gemein-
same gleich- oder
entgegengerichtete
Variationen

Kausalzusammen-
hänge für Vorher-
sagen oder Inter-
polationen nutzen,
Modellvalidierung

Fallen der Statistik

weiterführende
Methoden

Daten
zusammenfassen

Extremwertstatistik

Hauptkomponenten-
analyse

Clusteranalyse

Zeitreihenanal. etc.

Grundgesamtheit vs. Stichprobe

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Grundgesamtheit = Alle Elemente für die eine Aussage gemacht werden soll

- > Untersuchung der Grundgesamtheit meist nicht durchführbar (unendlich gross oder zu hohe Kosten oder ...)
- > Griechische Buchstaben für **Grundgesamtheit** (μ für Mittelwert)

Stichprobe = repräsentative Teilmenge der Grundgesamtheit

- > Keine Regel für Mindestgrösse (meist >30 angestrebt)
- > Standard Römisches Alphabet für **Stichprobe** (m für Mittelwert)

Deskriptive/Beschreibende Verfahren lassen sich auf Grundgesamtheit und Stichprobe anwenden

Schliessende/Analytische Verfahren leiten aus Daten einer Stichprobe Eigenschaften einer Grundgesamtheit ab

Schliessende Verfahren

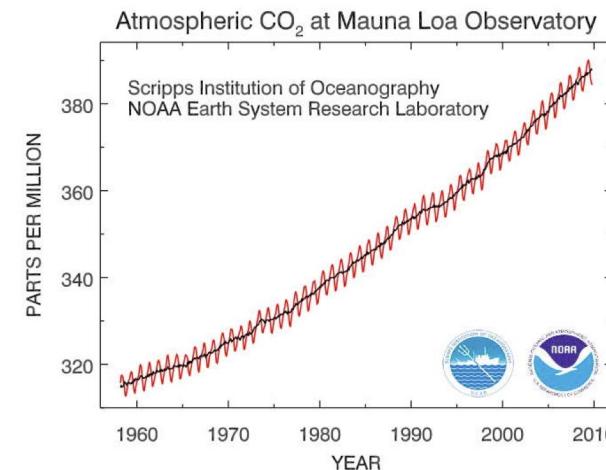
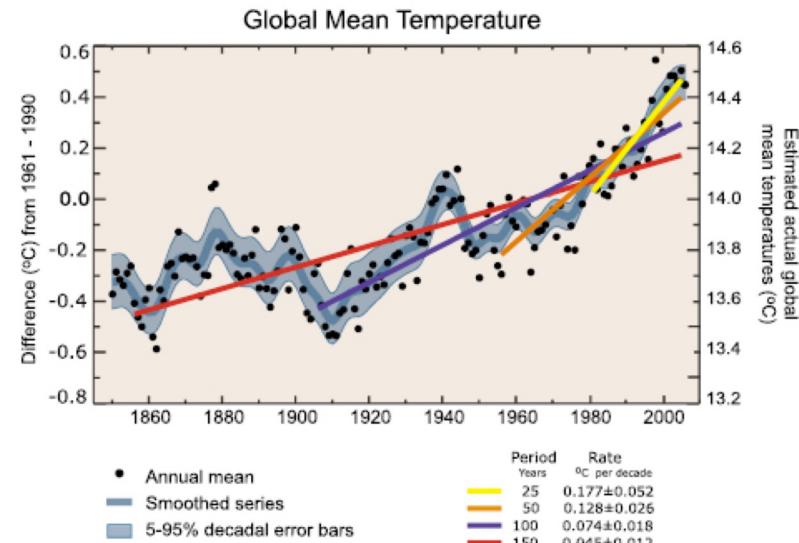
u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Ziele:

- > Hypothesen testen, z.B. Es gibt einen Erwärmungstrend über die Zeit
- > Parameter schätzen (Vorhersagen machen), z.B. Trend: die Temperatur nimmt um 1K pro 100 Jahre zu
- > Abhängigkeiten untersuchen, z.B. Temperatur und CO₂
- > Konfidenz-/Vertrauensintervalle angeben, z.B. mit 95%iger Sicherheit nimmt die Temperatur um 0.9 bis 1.1K pro 100 Jahre zu



Skalen

Nominalskala (keine Rangordnung)

- > z.B. Farben (grün, rot, blau), Wohnort
- > =/≠

Ordinalskala (Rangordnung)

- > z.B. A, B, C; Rangliste
- > =/≠, >/<

Kardinalskala

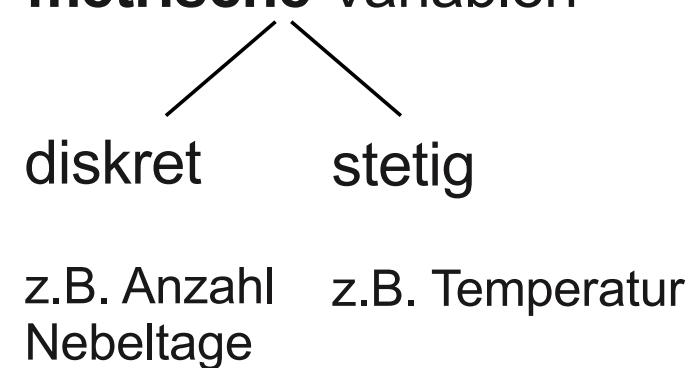
- > **Intervallskala** (x....y)
 - z.B. Temperatur [°C]
 - =/≠, >/<, +/-
- > **Verhältnisskala** (0....z)
 - z.B. Niederschlag in mm
 - =/≠, >/<, +/-, ×/÷



kategoriale Variablen



metrische Variablen

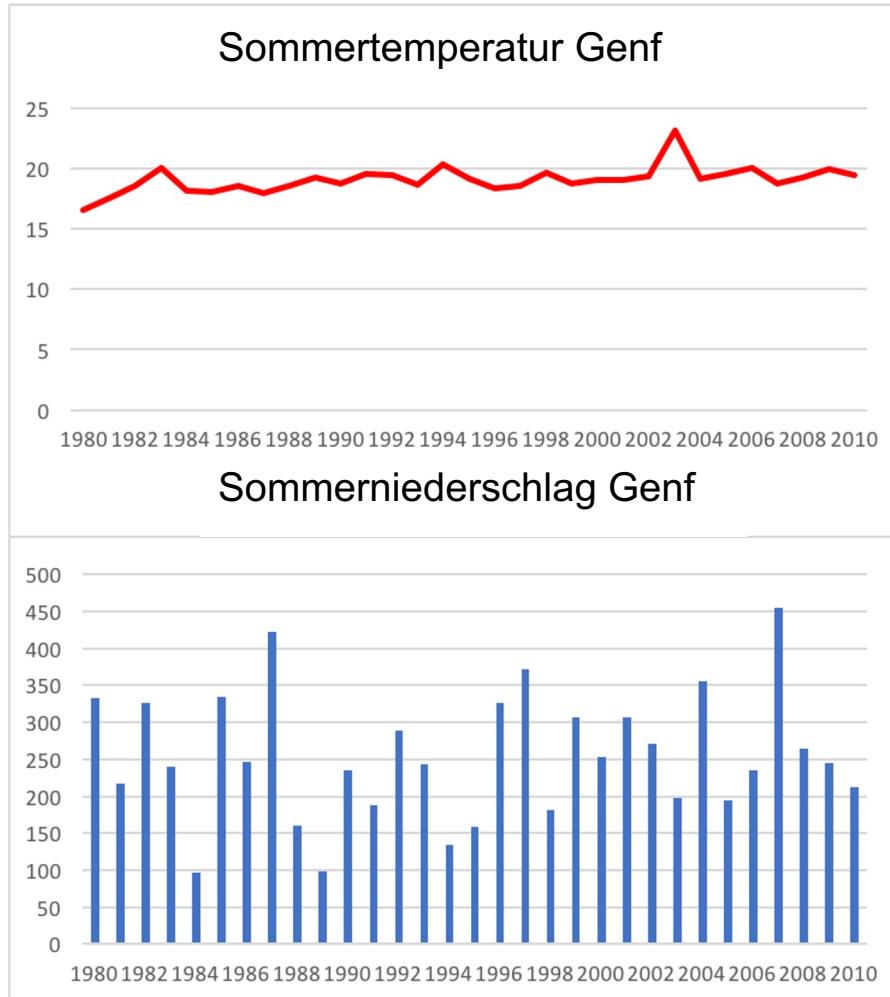


Deskriptive Verfahren

Was erkennt ihr in den Daten?

1980	16.53333	332.2
1981	17.53333	218
1982	18.56667	326.3
1983	20.03333	240.2
1984	18.13333	96.6
1985	18.06667	335.2
1986	18.56667	246.1
1987	17.93333	421.5
1988	18.53333	160.2
1989	19.2	97.9
1990	18.76667	234.3
1991	19.5	188.6
1992	19.4	289.2
1993	18.66667	243.5
1994	20.36667	133.7
1995	19.16667	157.8
1996	18.3	326.1
1997	18.5	372.5
1998	19.63333	181.2
1999	18.73333	307.3
2000	19.06667	252.9
2001	19.03333	307.2
2002	19.36667	271.5
2003	23.13333	197
2004	19.13333	355.2
2005	19.5	193.9
2006	20	234.9
2007	18.73333	455.6
2008	19.26667	264.1
2009	19.96667	244.1
2010	19.46667	212.8

Deskriptive Verfahren



1980	16.53333	332.2
1981	17.53333	218
1982	18.56667	326.3
1983	20.03333	240.2
1984	18.13333	96.6
1985	18.06667	335.2
1986	18.56667	246.1
1987	17.93333	421.5
1988	18.53333	160.2
1989	19.2	97.9
1990	18.76667	234.3
1991	19.5	188.6
1992	19.4	289.2
1993	18.66667	243.5
1994	20.36667	133.7
1995	19.16667	157.8
1996	18.3	326.1
1997	18.5	372.5
1998	19.63333	181.2
1999	18.73333	307.3
2000	19.06667	252.9
2001	19.03333	307.2
2002	19.36667	271.5
2003	23.13333	197
2004	19.13333	355.2
2005	19.5	193.9
2006	20	234.9
2007	18.73333	455.6
2008	19.26667	264.1
2009	19.96667	244.1
2010	19.46667	212.8

Lageparameter / Masse der Zentraltendenz

- > Mittelwert, Median, Modus, ...

Streuungsparameter

- > Spannweite, Varianz, Standardabweichung, Quantile, ...

Häufigkeitstabellen

Sonstiges

- > absolute Werte vs. Anomalien
- > Standardisierung/Transformationen
- > Freiheitsgrade

Lageparameter

Mittelwert, Median, Modus

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Modus

- > Wert der am häufigsten Auftritt (mehrere Werte, wenn multimodal)
- > Ordinal- und Nominaldaten

Median

- > Wert an mittlerer Stelle, wenn man Werte nach Grösse sortiert
- > Metrische und ordinale Daten
- > **verteilungsunabhängig**

arithmetischer Mittelwert

- > Metrische Daten, symmetrische Verteilung
- > Empfindlich gegenüber Ausreisern

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Kenntnis wird in der Prüfung vorausgesetzt

Streuungsparameter Quantile, Quartile, Median

u^b

b
UNIVERSITÄT
BERN

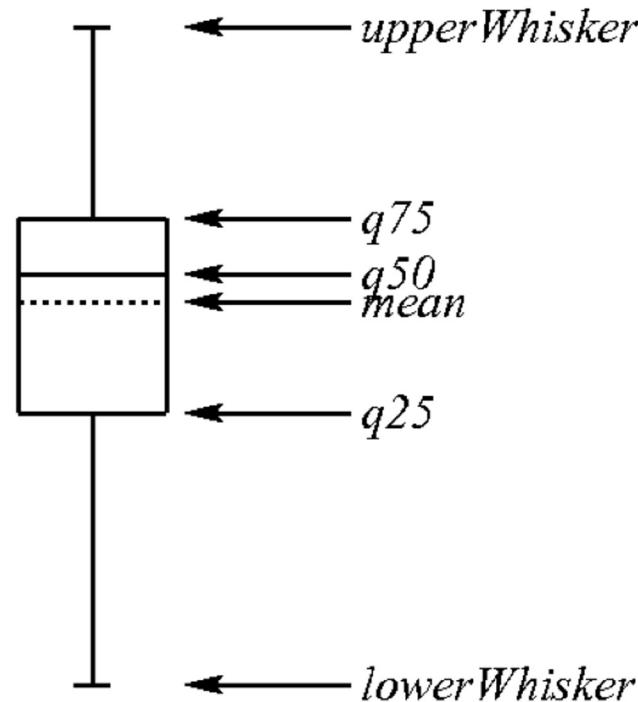
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Quantile

- > q% Quantil ist der Wert in einer geordneten Datenreihe, unterhalb dessen q% der Variablenwerte liegen
- > Metrische, ordinale Daten
- > Unempfindlich gegenüber Ausreisern
- > Auch für NICHT-symetrisch verteilte Daten

Spezielle Quantile

- > Median ist 50%-Quantil ($Q_{0.5}$)
- > Quartile ($Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$)
- > Whisker im Boxplot sind uneinheitlich definiert



Streuungsparameter Spannweite, Varianz, Standardabweichung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Spannweite

- > Metrische Daten
- > Stark extremwertabhängig

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

Kenntnis wird in der Prüfung vorausgesetzt

Varianz

- > Mittlere quadratische Abweichung von arithmetischen Mittelwert
- > Sinnvoll, wenn arithmetischer Mittelwert sinnvoll (metrische Daten, symmetrische Verteilung)
- > Starker Einfluss von Extremwerten durch das Quadratieren

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

Kenntnis wird in der Prüfung vorausgesetzt

Standardabweichung (Quadratwurzel der Varianz)

- > Gleiche Einheiten wie Ausgangsdaten

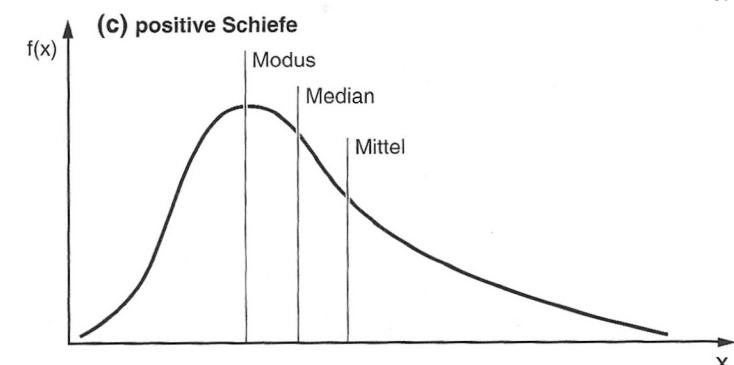
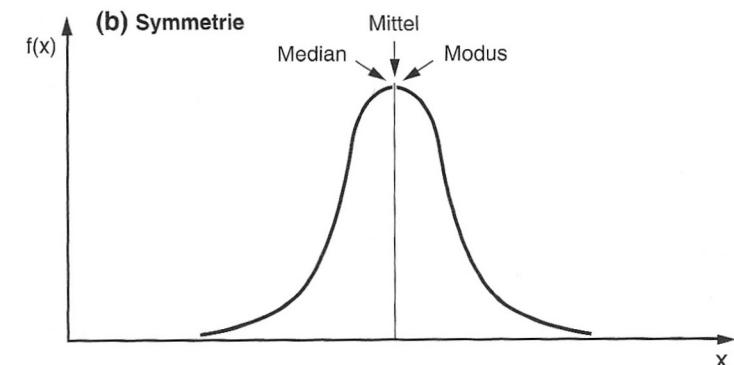
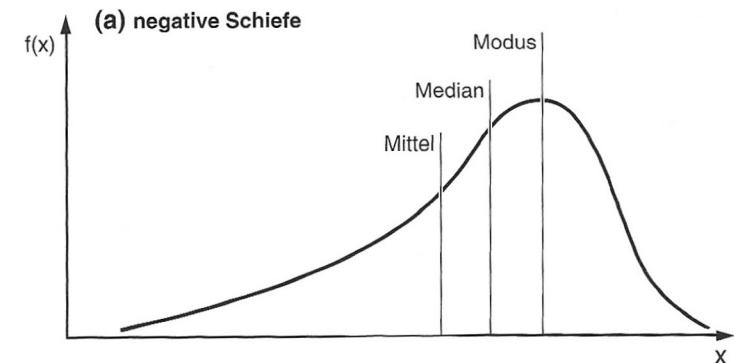
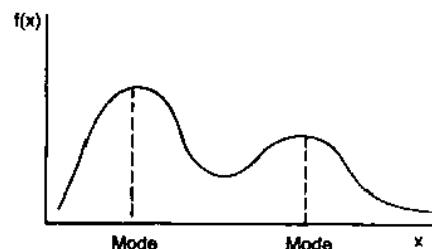
$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Kenntnis wird in der Prüfung vorausgesetzt

Schiefe

Einfaches Mass für die Schiefe:

- > $\text{Schiefe} = \frac{\text{arithm. Mittel} - \text{Median}}{\text{Standardabweichung}}$
- > Negative Schiefe = linksschief, rechtssteil
- > Positive Schiefe = rechtsschief, linkssteil
- > VORSICHT bei bi-, multi-modalen Daten!



Kreuztabelle / Kontingenztafel

Häufigkeitstabelle

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Für nominale Daten
- > Ordinale und metrische Daten können in nominale Daten transformiert werden (z.B. Grenzwertüberschreitung ja/nein)
- > Beispiel: Es werden 2000 Personen darüber befragt, ob sie Produkt A oder B bevorzugen. Das Ergebnis wird nach Geschlecht des Befragten ausgewertet.

Produkt/Geschlecht	weiblich	männlich	Summe
Produkt A	660	440	1100
Produkt B	340	560	900
Summe	1000	1000	2000

Freiheitsgrade = Anzahl der Werte, die man frei ändern kann, ohne das Ergebnis zu ändern.

Beispiel 1 Mittelwert:

- > Ein Produkt kostet in einem über 30 Geschäfte gemittelten Durchschnitt 19.63 CHF.
- > Die Preise von 29 Geschäften können nun frei variiieren und nur beim 30. Geschäft muss der Preis so angepasst werden, dass der Mittelwert 19.63 ergibt. Das arithmetische Mittel hat also n-1 Freiheitsgrade.

- > Anzahl Beobachtungen abzüglich Anzahl geschätzter Parameter.
- > Beispiel: Standardabweichung aus Stichprobe mit n Beobachtungen.

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

- > Der Mittelwert wurde bereits aus den Beobachtungen geschätzt. Wenn man den Mittelwert und alle Beobachtungen ausser der letzten kennt ($n-1$) dann kann man diese berechnen, es besteht also keine "Freiheit" mehr (oder: die letzte Beobachtung ist "überflüssig").

Freiheitsgerade

R Beispiel

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

```
> # Grundgesamtheit
> gg=rnorm(100000,mean=0,sd=10)
> sqrt(sum((gg-mean(gg))^2)/100000)

> # Stichprobe ziehen
> st=sample(gg,30)
> sqrt(sum((st-mean(st))^2)/30)
> sqrt(sum((st-mean(st))^2)/(30-1))

> # Ergebnis von Stichprobenziehung abhängig,
> # deshalb viele Stichproben ziehen
> s=vector()
> sm1=vector()
> for (i in 1:1000) {
>   st=sample(gg,30)
>   s[i] <- sqrt(sum((st-mean(st))^2)/30)
>   sm1[i] <- sqrt(sum((st-mean(st))^2)/(30-1))
> }
> mean(s)
> mean(sm1)
```

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

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

Take-home messages

Metrische und symetrisch verteilte Daten

- > Mittelwert und Standardabweichung sind aussagekräftig

NICHT-symetrisch verteilte Daten

- > Median und Quantile sind robustere Lage- und Verteilungsmasse

Nomiale Daten

- > Können mit Modus und Kontingenztabellen beschrieben werden

Absolute vs. relative Zahlen?

Wichtige Begriffe

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Objekt	Untersuchungseinheit	Beispiel: Person
Merkmal	Theoretische Dimension, die mehreren Objekten gemeinsam ist	Körpergrösse, Haarfarbe
Konstante	Merkmal mit nur einer beobachteten Ausprägung	schwarz
Variable	Merkmal mit mehreren beobachteten Ausprägungen	schwarz, braun, blond, ...
Wert	Ausprägung einer Variablen	180cm Körpergrösse
Skala	Systematische Zuordnung von Zahlen oder Symbolen zu den Ausprägungen einer Variablen	Nominal, ordinal, kardinal/metrisch
Messen	Einordnen von Objekten auf einer Skala (umfasst auch Zählen)	32

Häufige Abkürzungen

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Skalare gekennzeichnet durch nicht-fett gedruckte Kleinbuchstaben (x),
- > Vektoren durch fett gedruckte Kleinbuchstaben (**x**) und
- > Matrizen durch fett gedruckte Grossbuchstaben (**X**)
- > n: Stichprobengrösse
- > H_0 : Nullhypothese (beinhaltet normalerweise das „=“ Zeichen)
- > H_1 oder H_A : Alternativhypothese (beinhaltet normalerweise das „≠“, „>“ oder „<“ Zeichen)
- > α : Signifikanznivau, maximal zulässige Irrtumswahrscheinlichkeit für Fehler 1. Art (α -Fehler)
- > p-Wert: Probability, also Wahrscheinlichkeit, diese Stichprobenwerte zu erhalten, wenn H_0 zutrifft
- > x/μ : Mittelwert der Stichprobe/Grundgesamtheit
- > s/σ : Standardabweichung der Stichprobe/Grundgesamtheit
- > s^2/σ^2 : Varianz der Stichprobe/Grundgesamtheit
- > s_x : Standardfehler zeigt die theoretische Streubreite des Stichprobenmittelwerts $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$
- > $s_{\widehat{\beta}_1}$: Standardfehler des Regressionskoeffizienten
- > s_{xy} : Kovarianz zwischen x und y
- > r/ρ : Korrelationskoeffizient
- > y: Beobachtungen von y
- > \hat{y} : Das „Dach“ steht für eine Schätzung hier von y, z.B. aus Regressionmodell
- > β_0 : Regressionskonstante (Schnittpunkt mit der y-Achse bei $x=0$)
- > β_1 : Regressionskoeffizient (Steigung der Regressionsgeraden)
- > R^2 : Bestimmtheitsmass bei der Regression ($r^2=R^2$)
- > ε : error, also Fehler, bei der Regression die Residuen ($y - \hat{y}$)

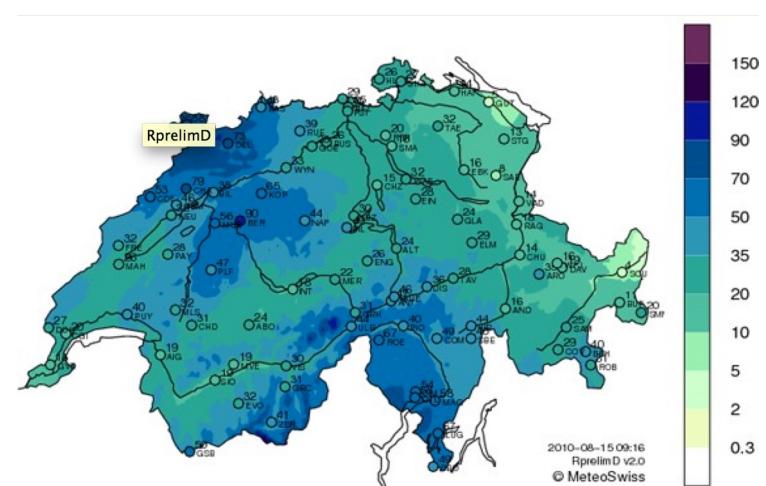
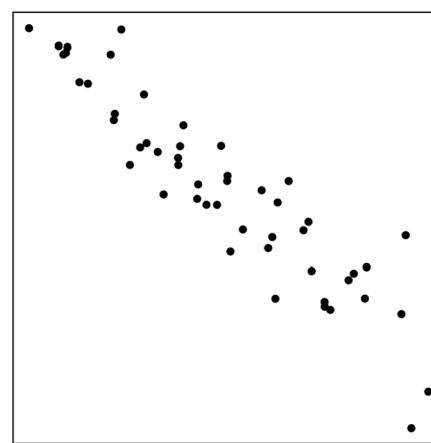
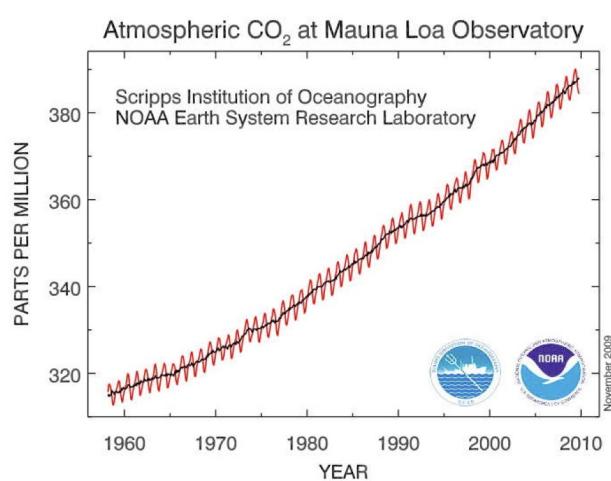
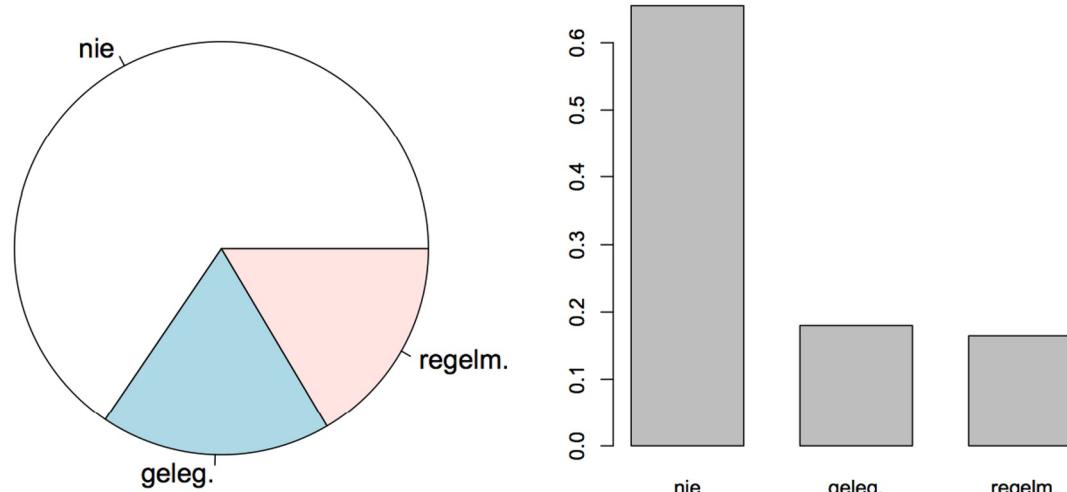
Visualisierung

Das Auge erkennt sehr gut Strukturen in den Daten!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Graphiken sollten benutzt werden, wenn:

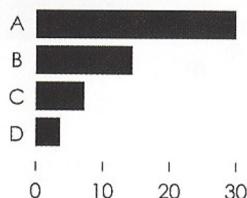
- > für schnelle Übersichten unübersichtlicher Zahlen
- > genaue Zahlen nicht interessieren
- > als Argumente in Vorträgen, da Graphiken mehr Information in kurzer Zeit vermitteln und gut im Gedächtnis bleiben

Balken- und Kreisdiagramme

Categories

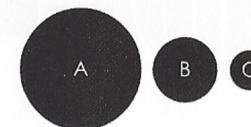
When your data is straightforward, with a value for each category, these are easy to read and create.

Bar graph



With length as visual cue, useful for straightforward comparisons

Symbol plot

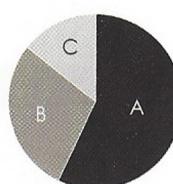


Can be used in place of bars, but can be hard to see small differences

Parts of a whole

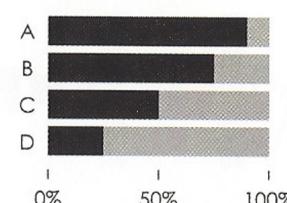
The categorical breakdown within a population can be interesting, and you might want to keep the groups together, although often not essential.

Pie chart



Parts add to 100 percent, typically sorted clockwise for readability

Stacked bar chart

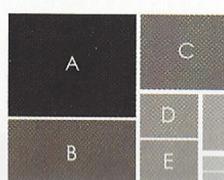


Often used to show poll results and can also be used for raw counts

Subcategories

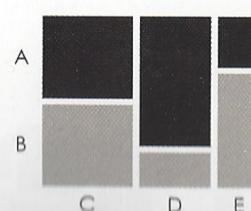
Data can have a hierarchical structure, which can be important in data interpretation and it often allows for different points of view.

Treemap



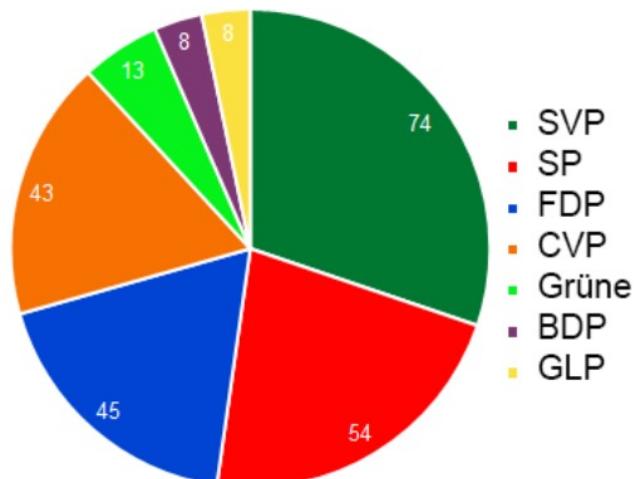
Shows hierarchical structure in a compact space, area often combined with color

Mosaic plot



Allows comparison across multiple categories in one view

Kuchendiagramme



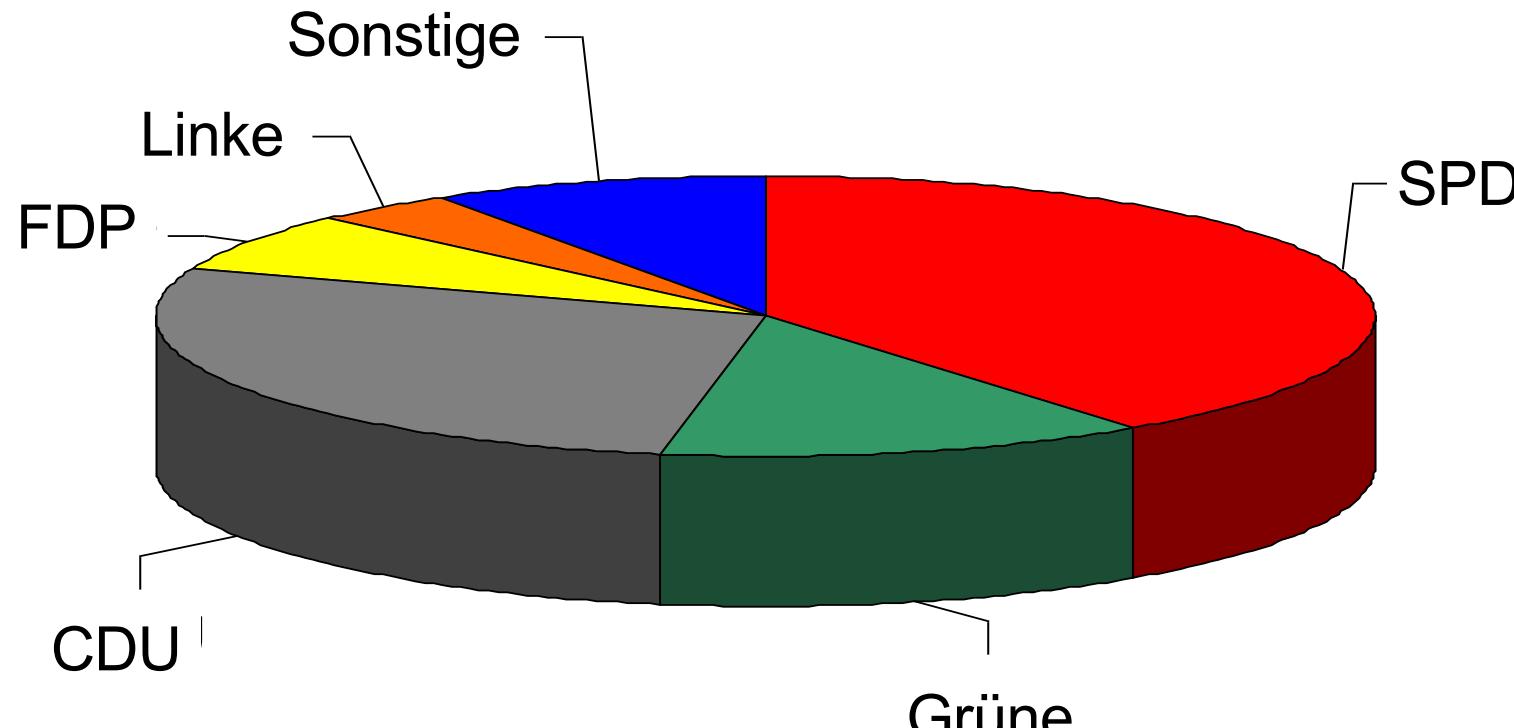
Anzahl Ratsmitglieder
pro Fraktion (21.6.19)

18.7

Man liest im Uhrzeigersinn,
beginnend auf 12 Uhr.
Wichtige Tortenstücke
auf die 12-h-Position setzen.

Kuchendiagramme

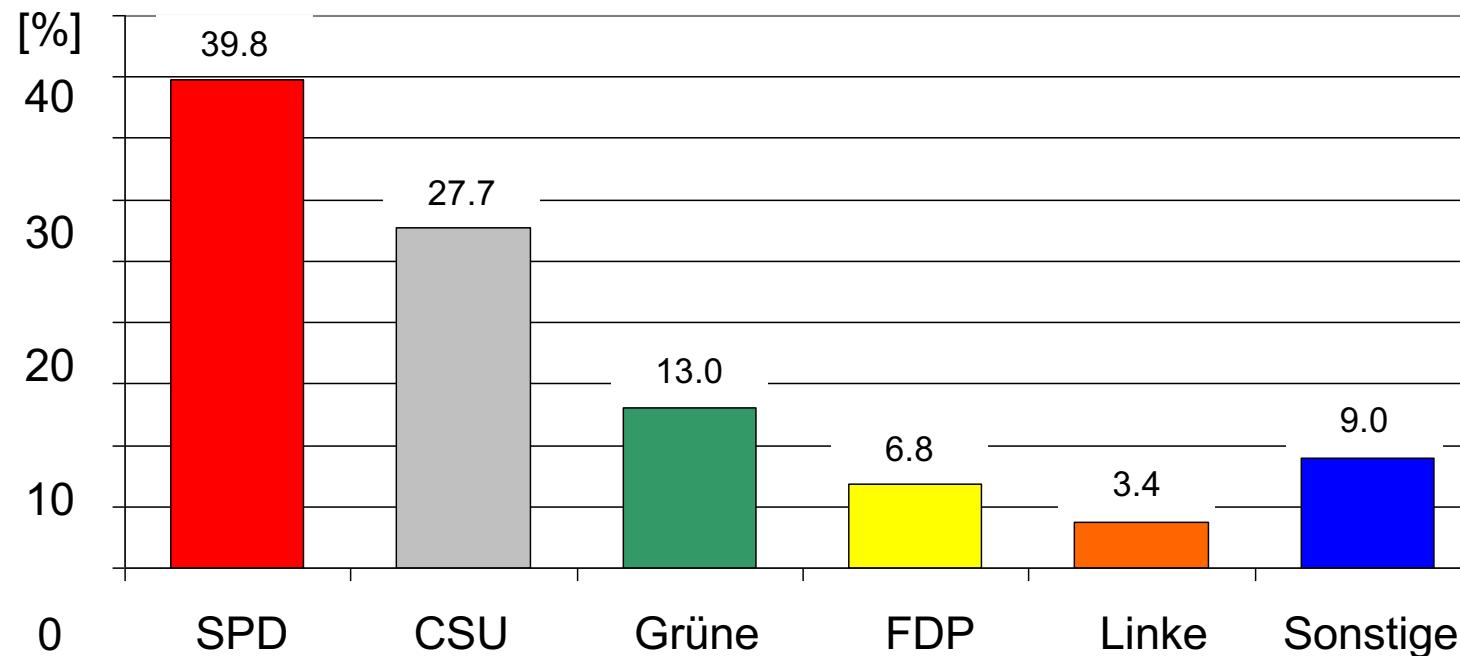
Was zusammengehört,
gehört nebeneinander!



Winkel bei 3D-Darstellung
Schwer zu interpretieren

Säulendiagramme

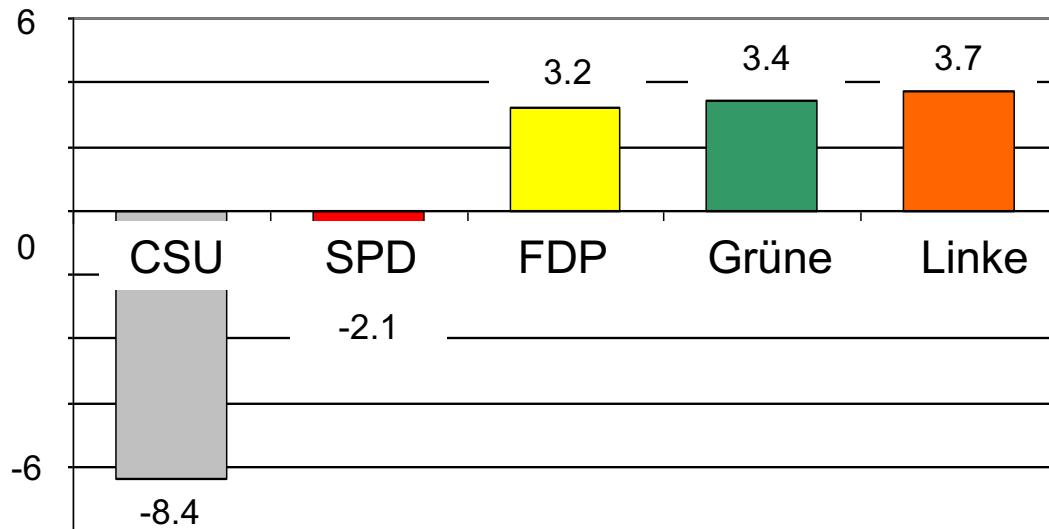
- > Wenn der Vergleich von Anteilen im Vordergrund steht
- > Für Torten mit zu vielen Anteilen



- > Säulen betonen die Anteilsunterschiede und Rangfolge
- > Torten die Aufteilung eines Ganzen auf Teile

Säulenvariationen

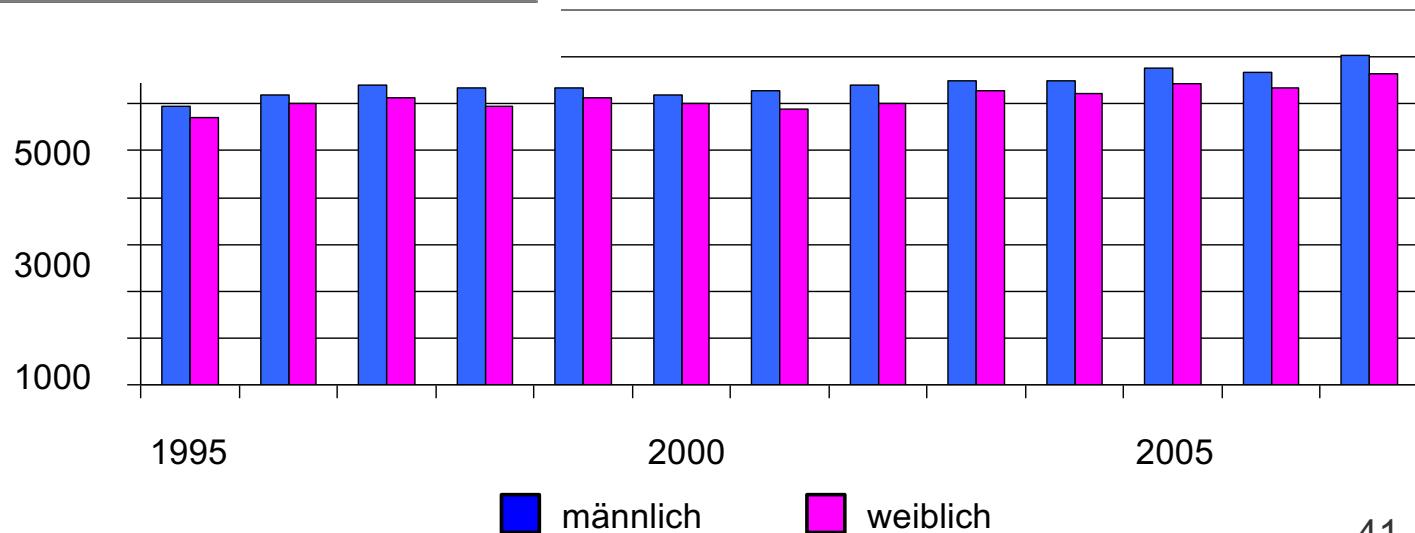
Gewinne und Verluste bei den Wahlen



Auch negative Werte
darstellbar (nicht bei
Torte)

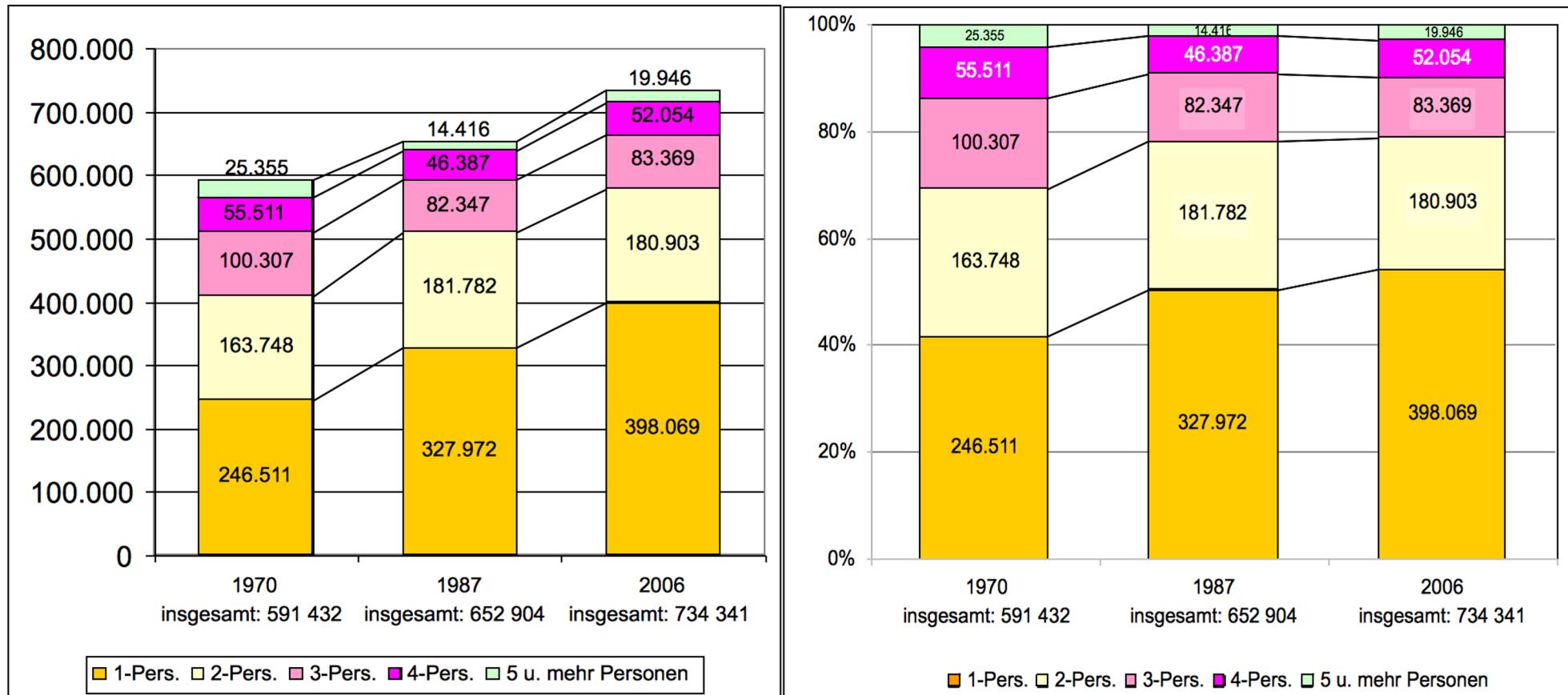
Geburten 1995-2007

Doppel-
Säulen-
Diagramm



Säulenvariationen

Privathaushalte in München 1970, 1987, 2006



Additives Säulendiagramm

100%-Säulendiagramme stellen
Verschiebung der Anteile heraus

Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Beispiel: Geschlechterverteilung 1974 bis 2004 in %

Jahr	Frauenanteil	Männeranteil
1974	6.8	93.2
1984	10.8	89.2
1994	23.6	76.7
2004	33.2	66.8

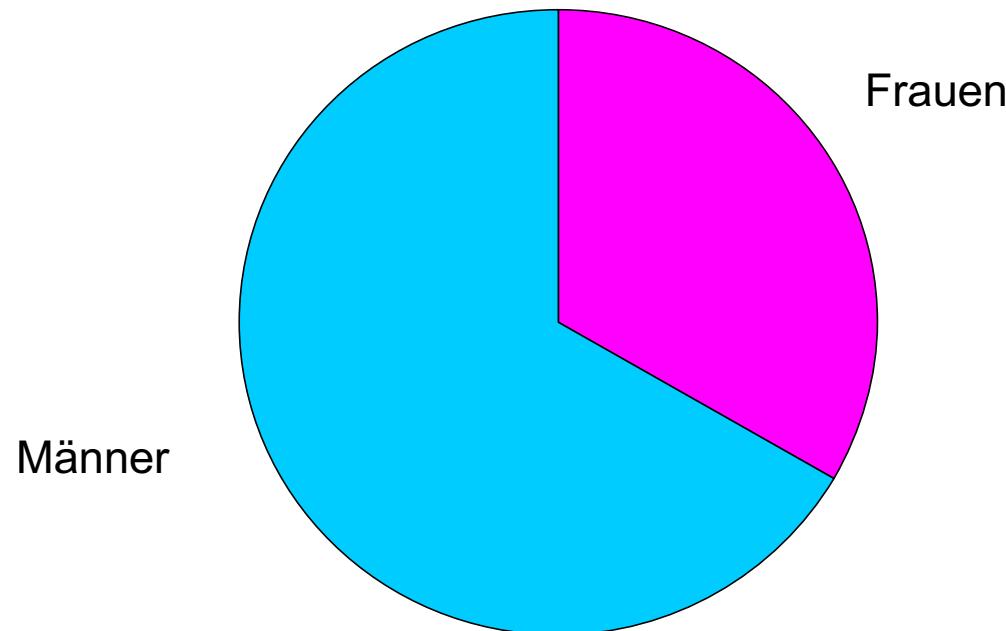
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Tortendiagramme betonen die Anteile



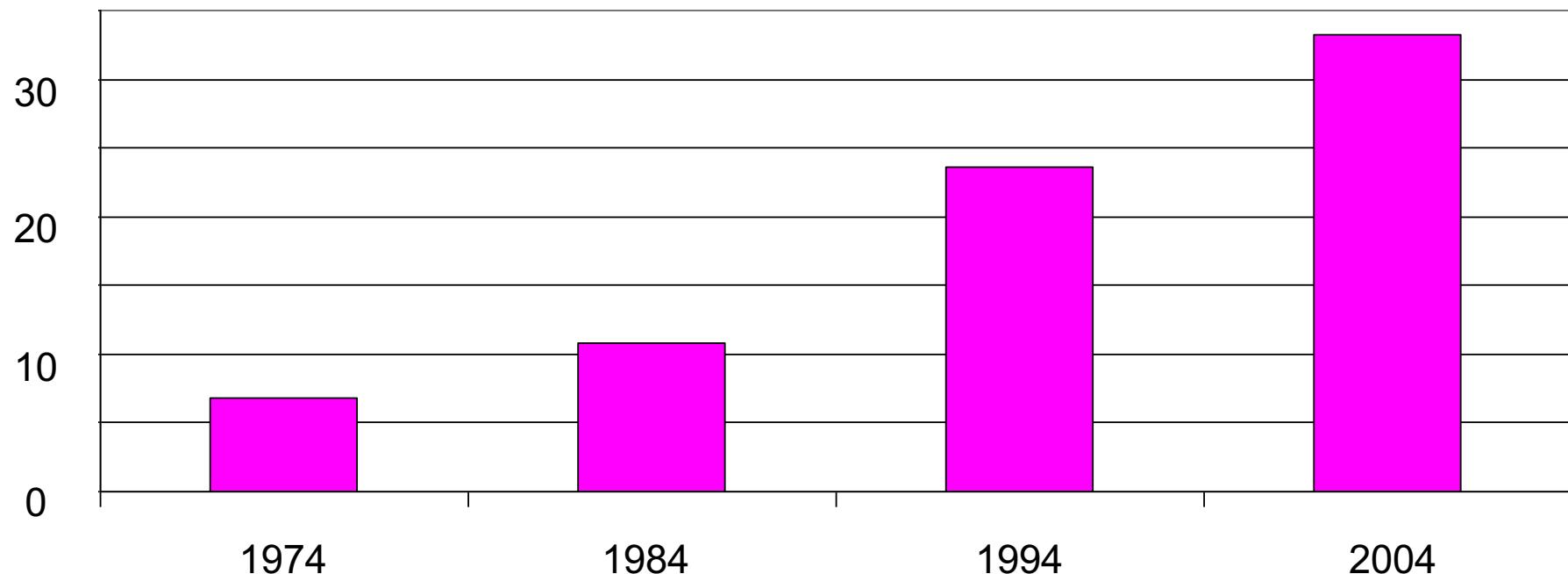
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Säulen betonen die Unterschiede von Anteilen im Verlauf



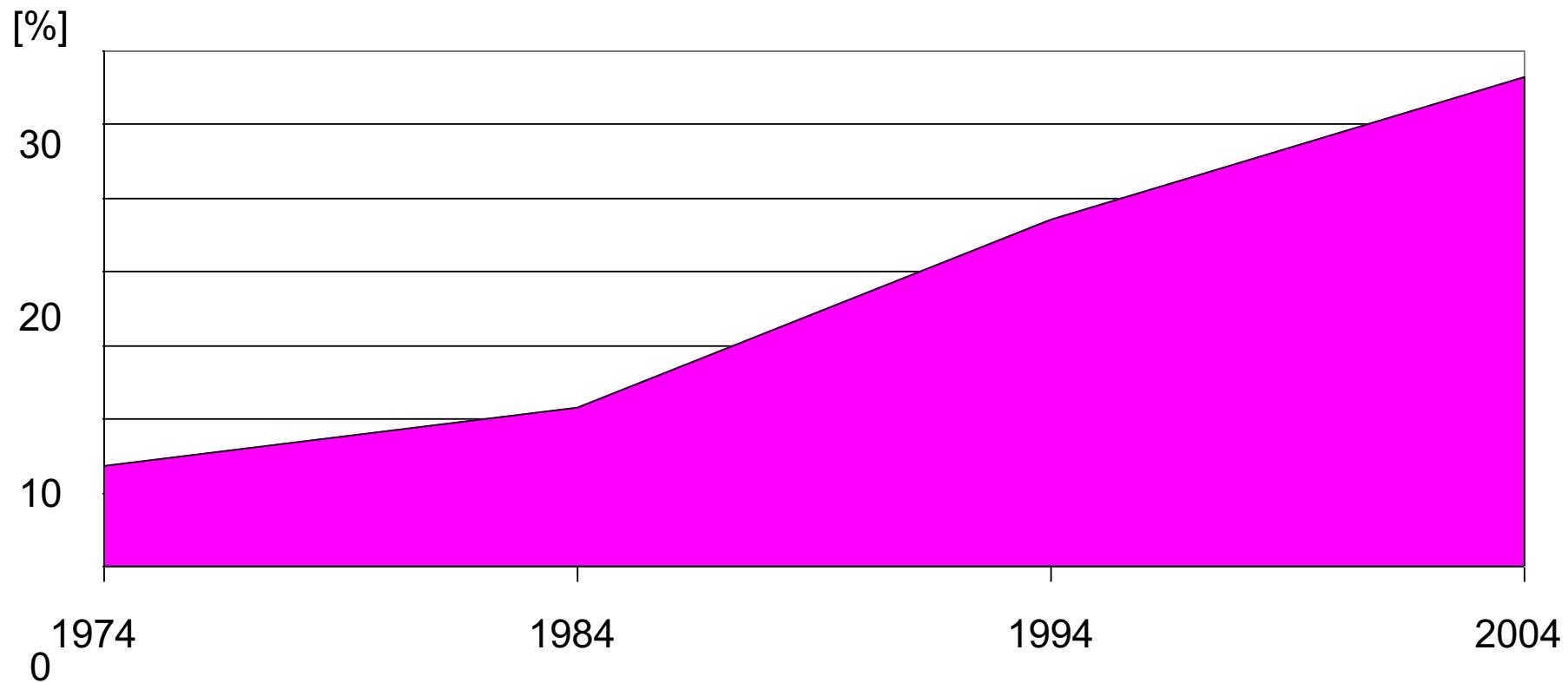
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Kurvendiagramme betonen den Trend



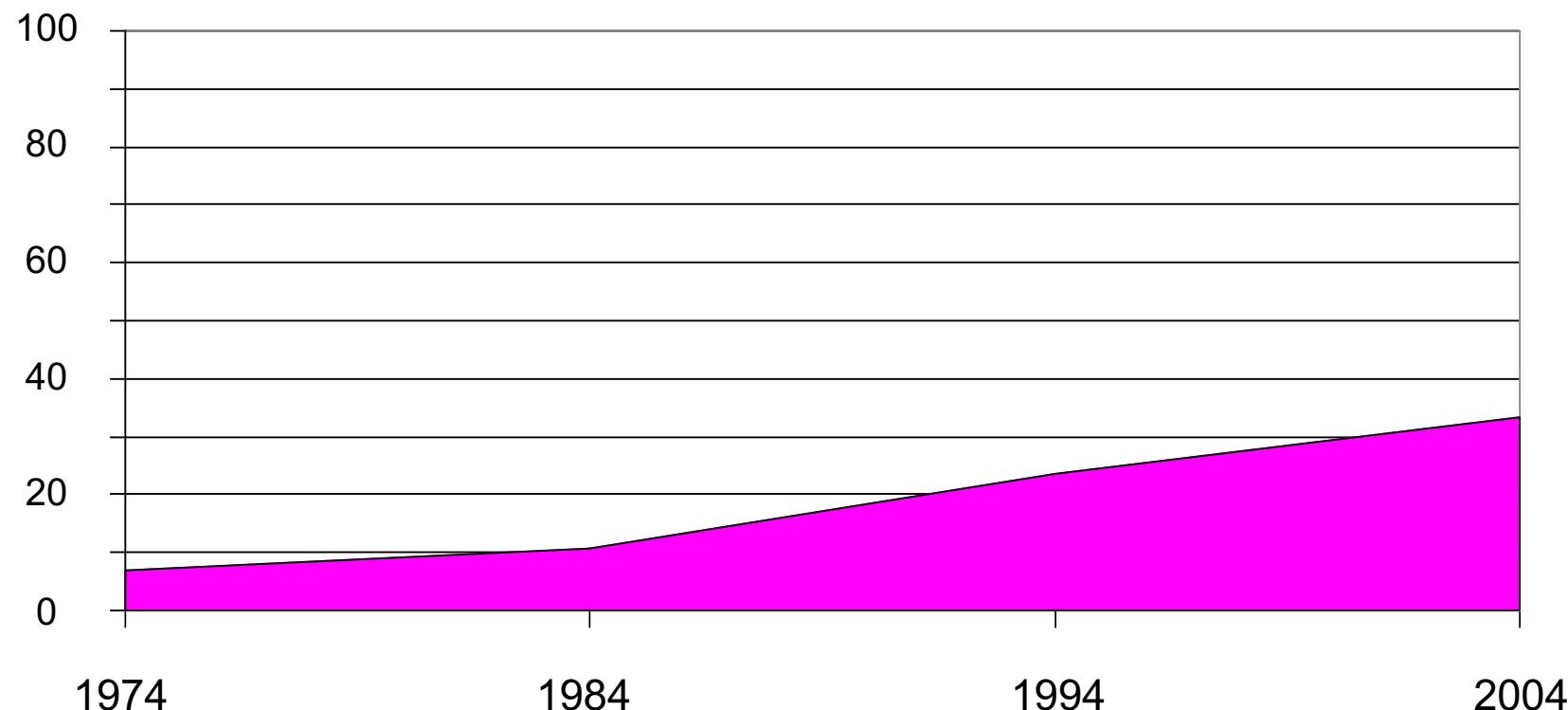
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Diese Kurve betont den Abstand zur 100%-Linie

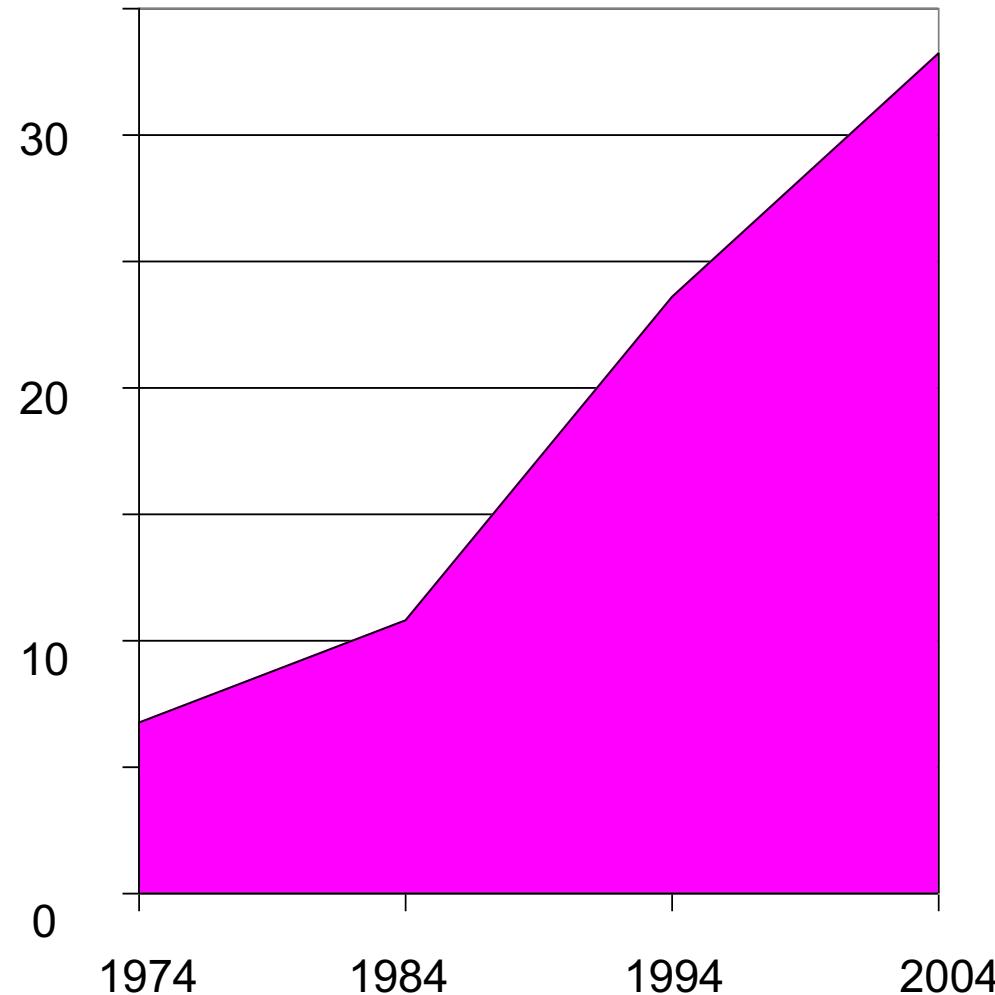


Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



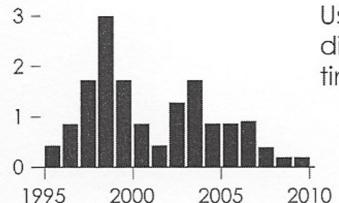
Bei dieser Kurve
wächst der
Frauenanteil
dramatisch!

Zeitreihen

Time series

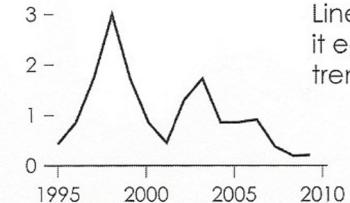
There are a variety of ways to see patterns over time, using cues such as length, direction, and position.

Bar graph



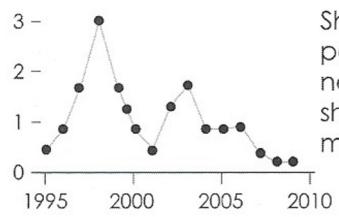
Useful for discrete points in time

Line chart



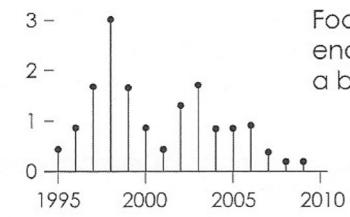
Lines can make it easier to see trends

Dot plot



Shows distinct points but might need line to show trend if not much data

Dot-bar graph

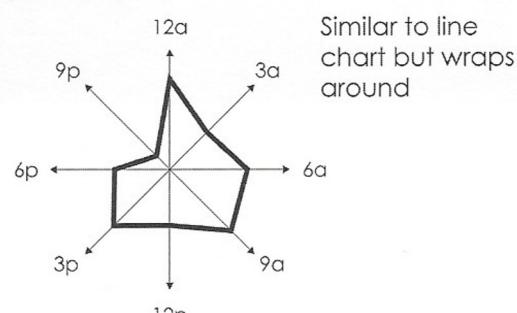


Focuses more on endpoints than a bar graph

Cycles

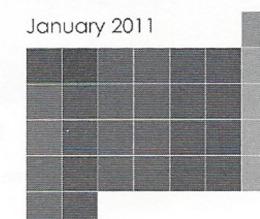
Time of day, day of the week, and month of the year repeat themselves, so it is often beneficial to align the segments in time.

Radial plot



Similar to line chart but wraps around

Calendar



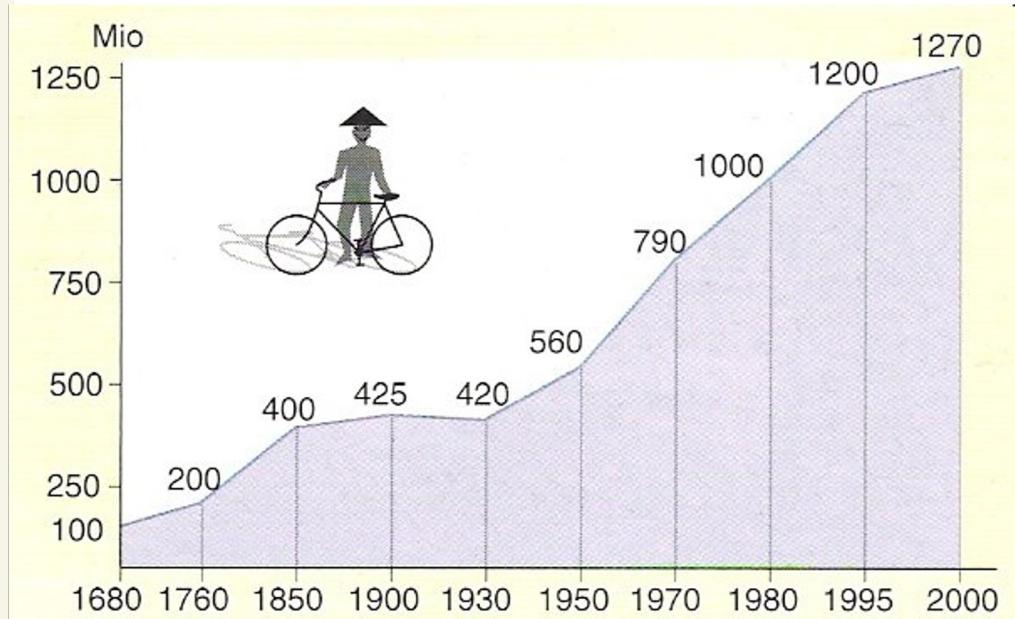
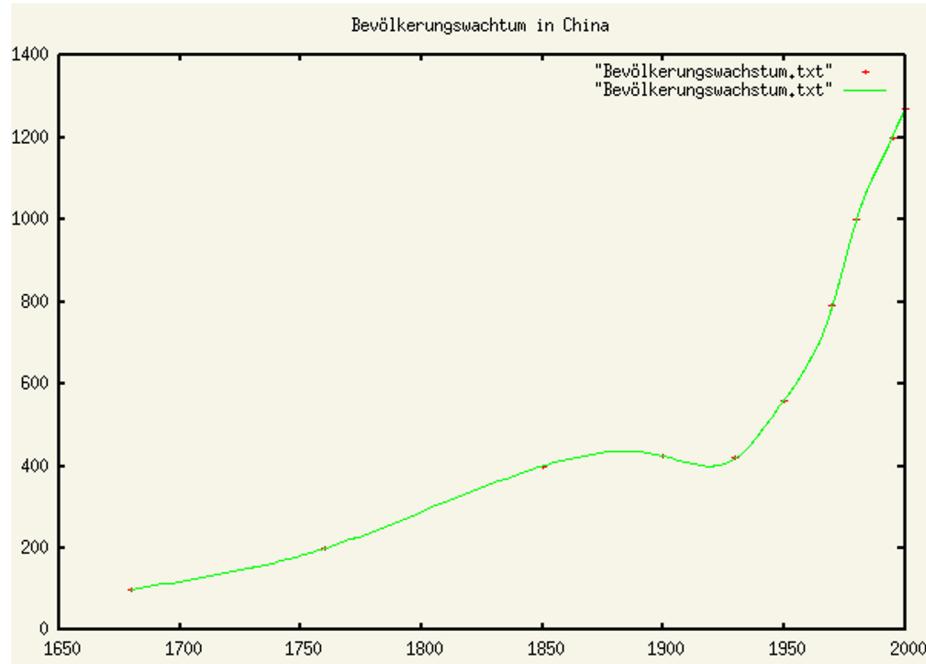
Patterns for days of week seen more easily than views above

Woher kommen die Unterschiede?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

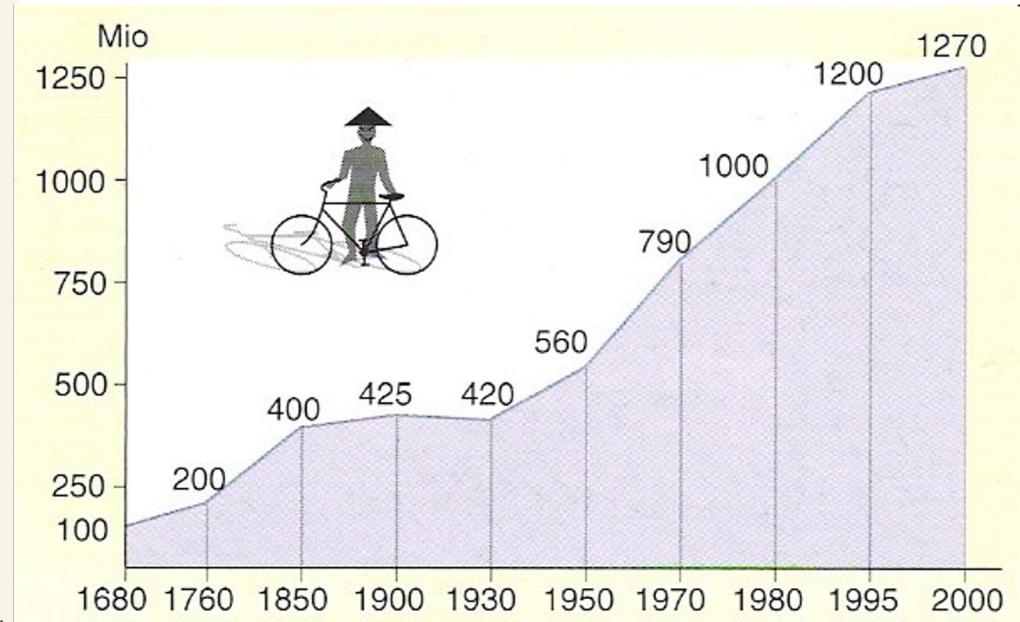
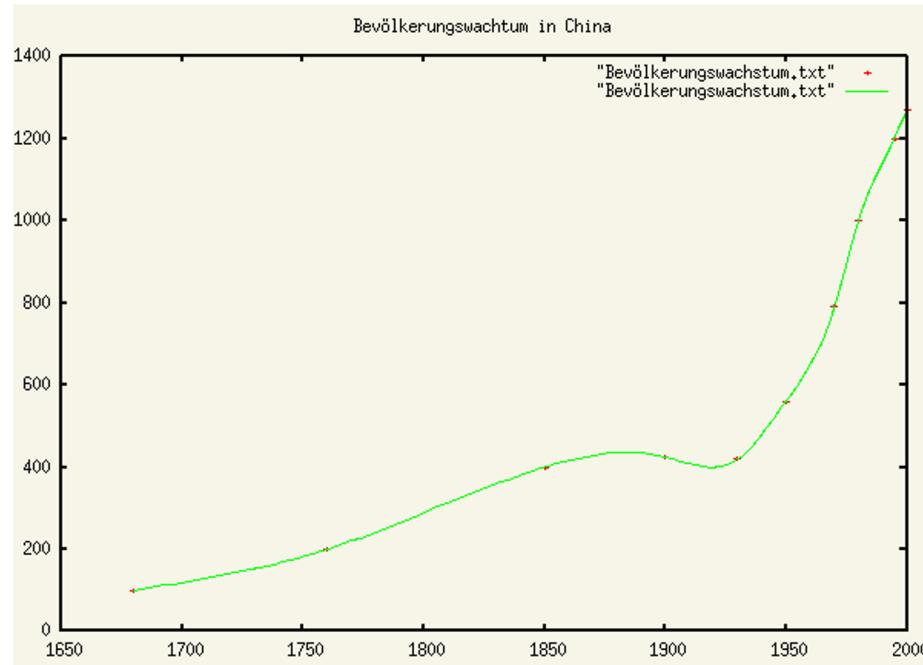


Achtung: Manipulationsgefahr!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



- > Auf Achsenbeschriftung achten
- > Achsen von eigenen Abbildungen IMMER gut beschriften einschliesslich der Einheiten!

Warum nimmt das Flugzeug diese Route?

u^b

b
UNIVERSITÄT
BERN

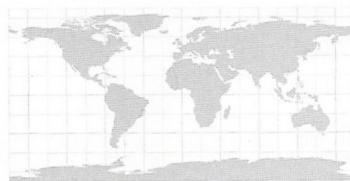
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Map projections

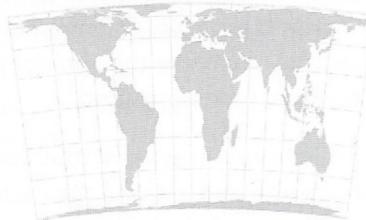
Equirectangular

Typically used for thematic mapping, but doesn't preserve area or angle



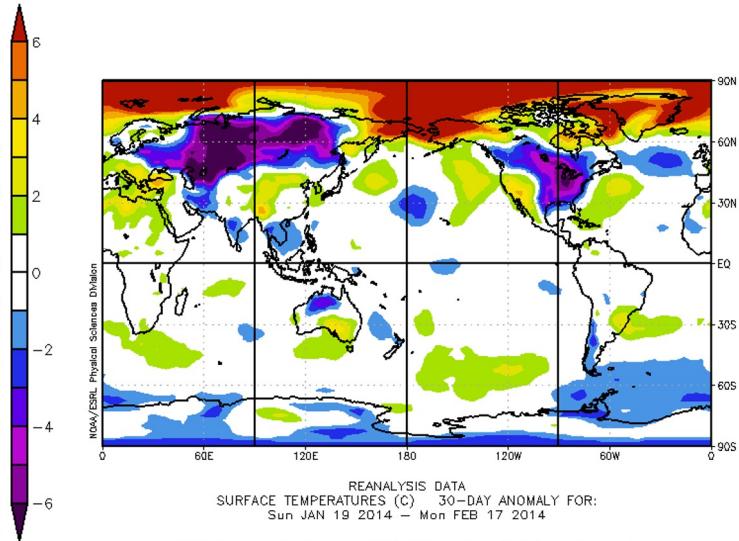
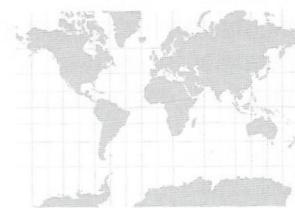
Albers

Scale and shape not preserved; angle distortion is minimal



Mercator

Preserves angles and shapes in small areas, making it good for directions



Lambert conformal conic

Better for showing smaller areas and often used for aeronautical maps.



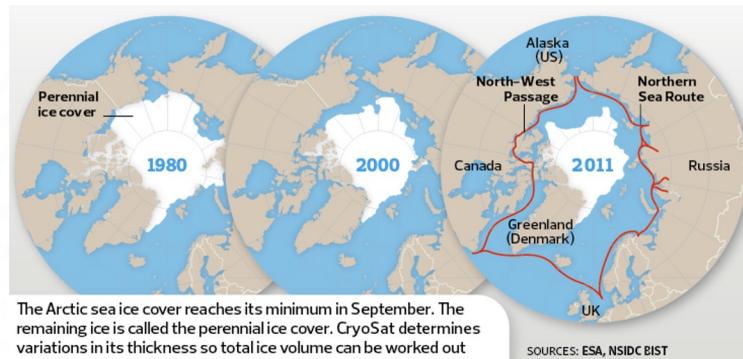
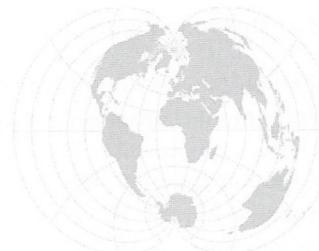
Sinusoidal

Preserves area; useful for areas near the prime meridian



Polyconic

Was often used to show US in the mid-1900s; little distortion in small areas near meridian



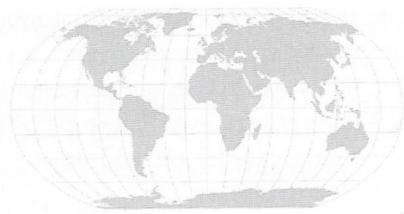
Winkel Tripel

Minimized area, angle, and distance distortion; good choice for world map



Robinson

A compromise between preserving areas and angles; good to show world map



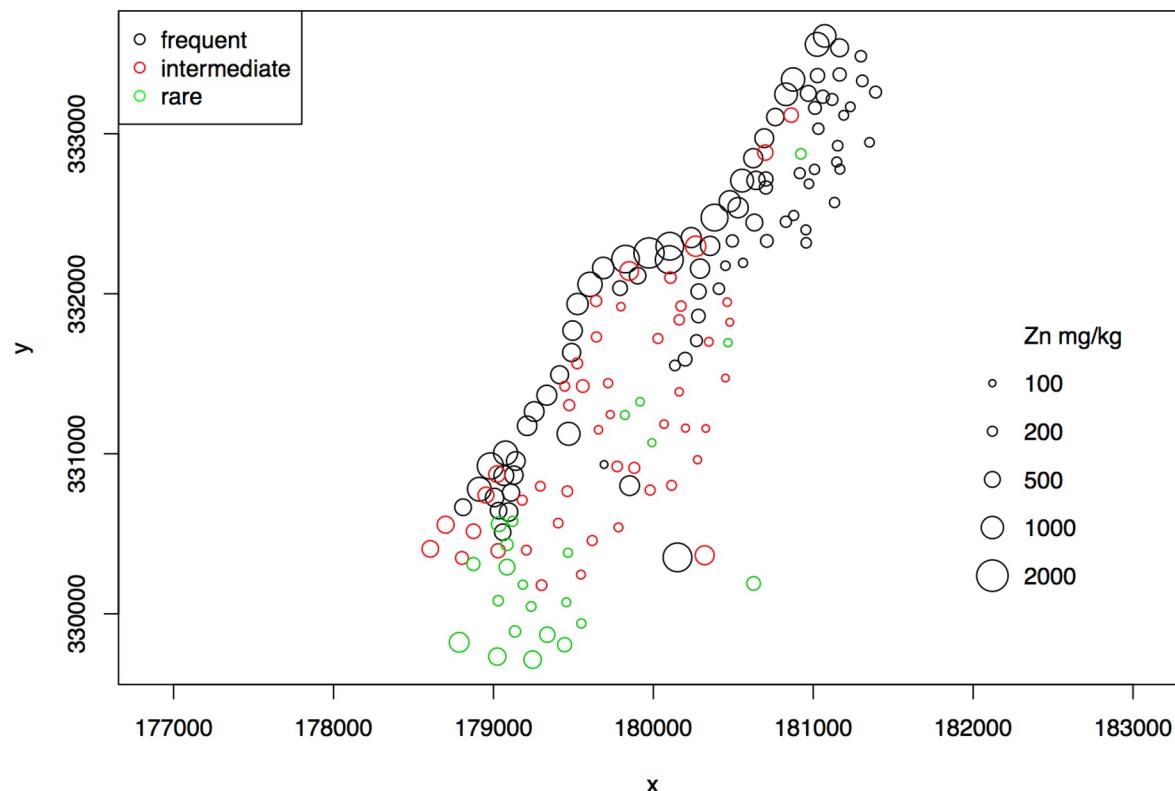
Orthographic

Representing a 3-D object in 2-D, need to rotate to area of interest



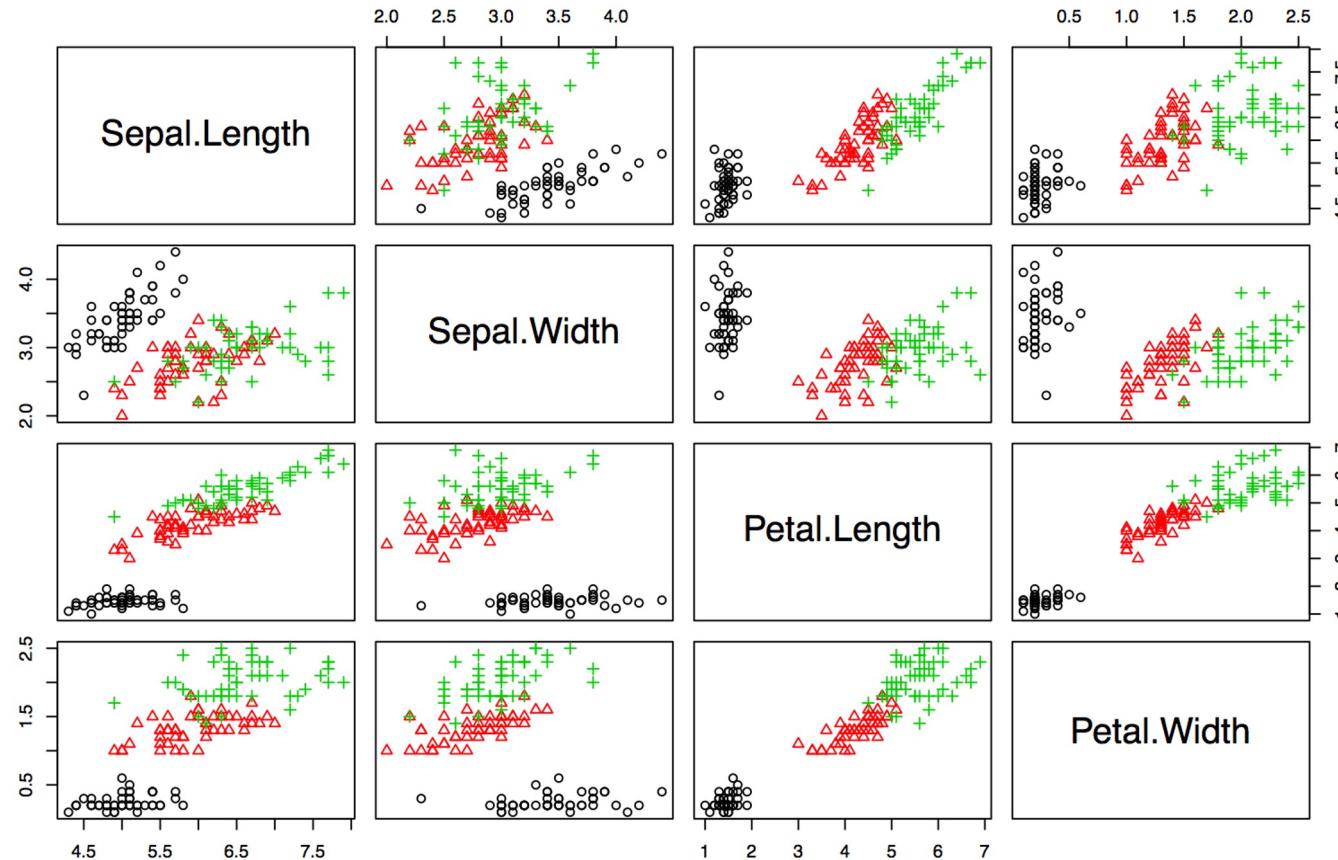
Multivariate Darstellungen

Beispiel Streudiagramm mit Farben und variabler Symbolgrösse in Relation zu Zinkgehalten im Boden

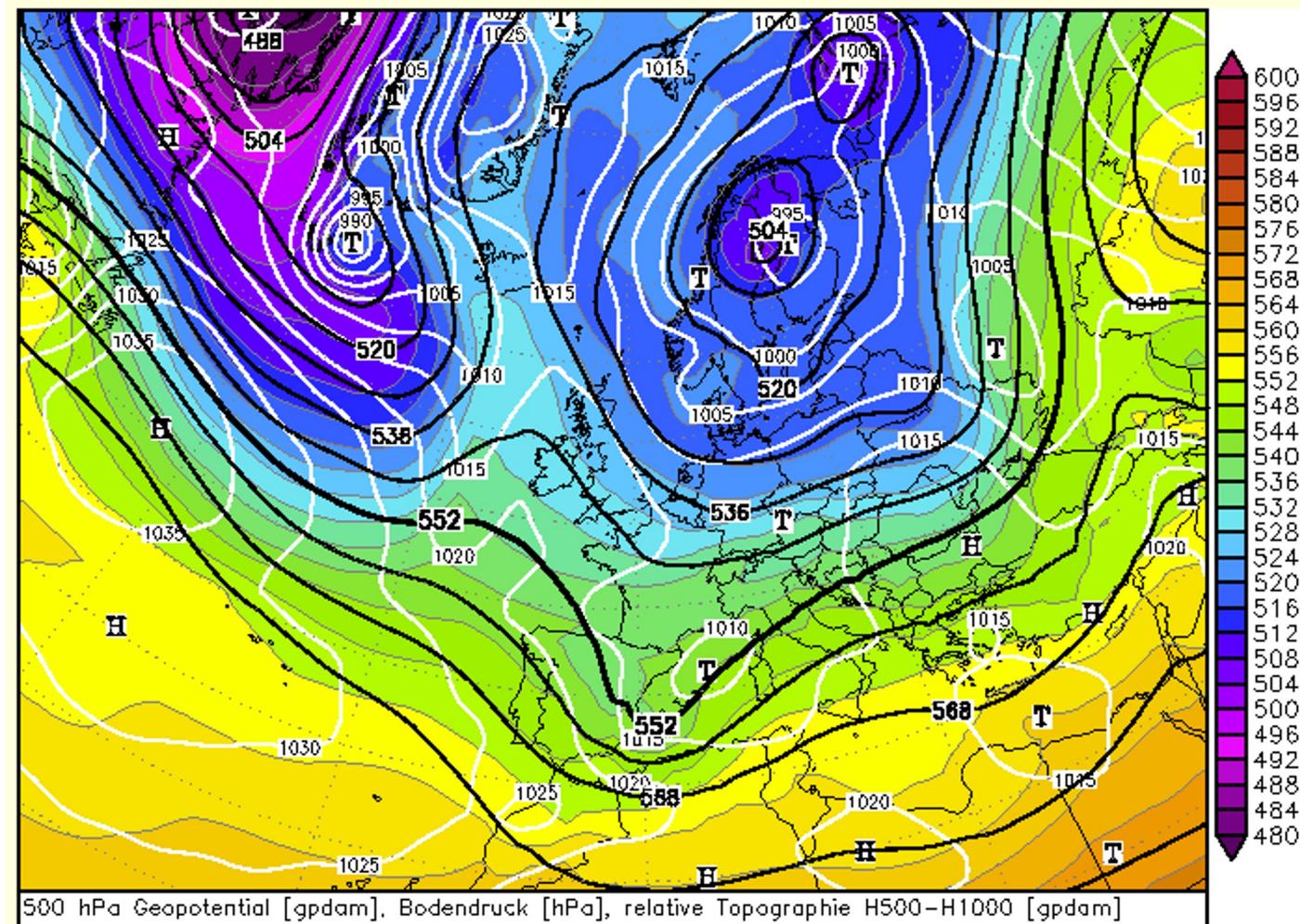


Multivariate Darstellungen

Streudiagrammmmatrix für 4 Variablen und 3 Pflanzenarten

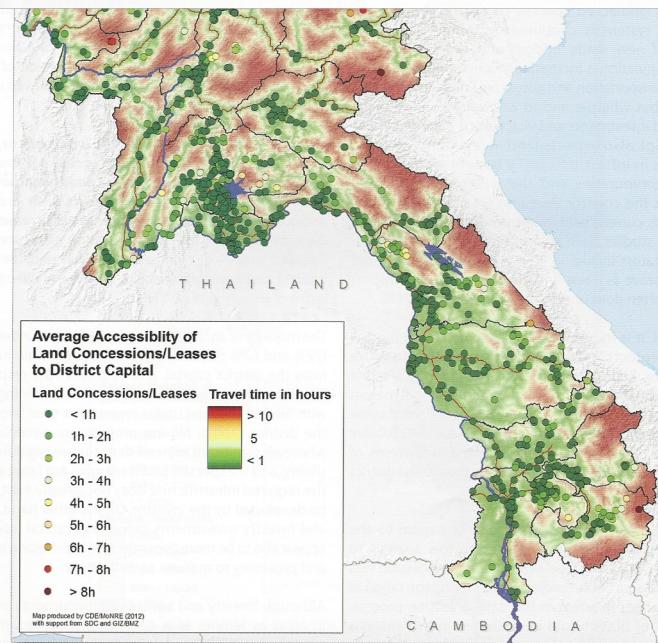
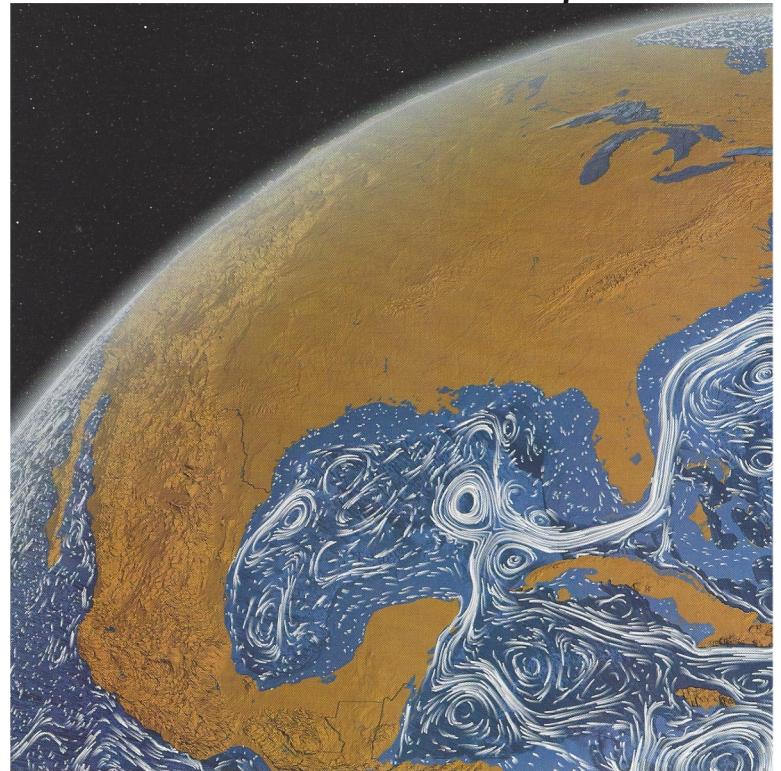
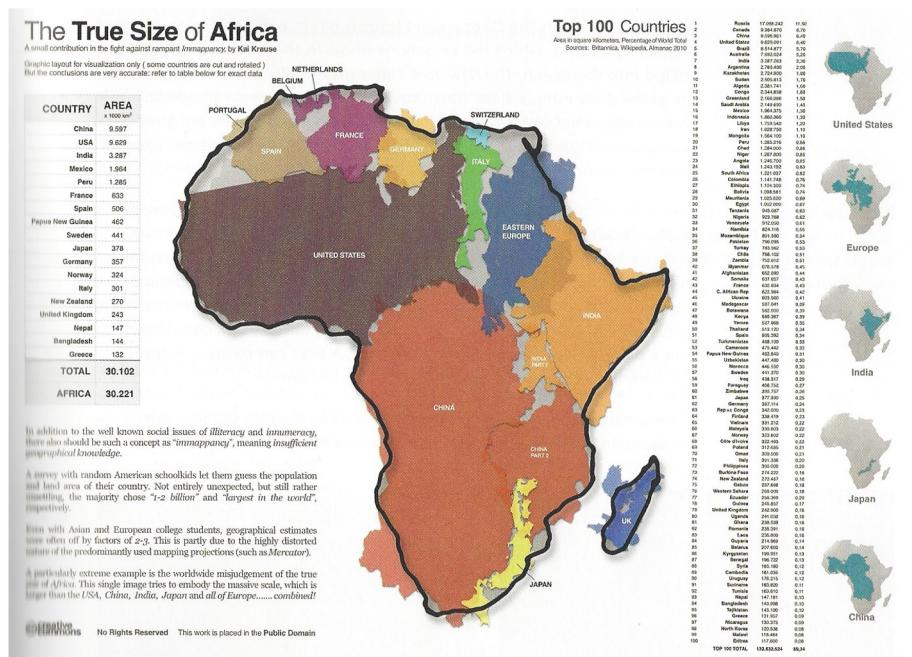


Multivariate Darstellungen



Faktenbezogen oder künstlerisch

thetruesize.com

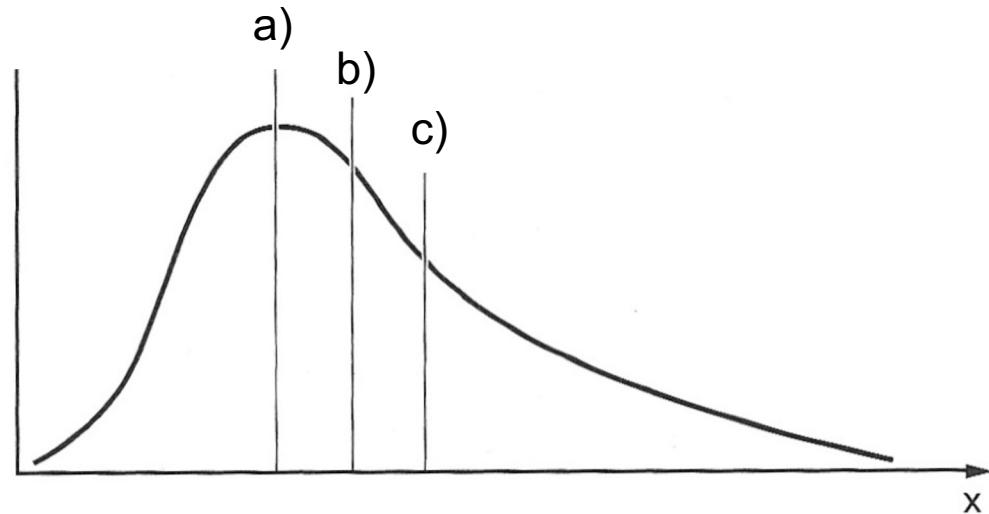


Take-home messages

- > Daten immer zuerst plotten, das Auge erkennt Strukturen, aber auch Fehlwerte sehr gut
- > Art der Abbildung gut auswählen und bei fremden Graphiken vor allem genau auf die Achsen achten

Beispiel Prüfungsfrage

- > Ordne a, b und c zu:
 - Mittelwert
 - Median
 - Modus



- > Welche Streuungsmasse eignen sich für die Daten mit der abgebildeten Verteilung
 - Varianz
 - Standardabweichung
 - Quantile
 - Minimum/Maximum

Würfelexperiment

Gruppen «Würfeln»:

- > 50 Mal mit zwei Würfeln würfeln

Gruppen «Fälschen»:

- > Summe der Augenzahl zweier Würfel ausdenken

Nächste Stunde versuchen wir die Fälschungen zu identifizieren. Dazu bitte:

- 1) Augenzahlen addieren
- 2) Wie oft ist die Summe=2, =3, ...?
- 3) Ergebnisse als ASCII .txt oder Excel mit Gruppennamen (NICHT Gruppe 1 oder 2) per E-Mail an joerg.franke@unibe.ch

Augen-zahlen-summe	Strich-liste	Summe
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

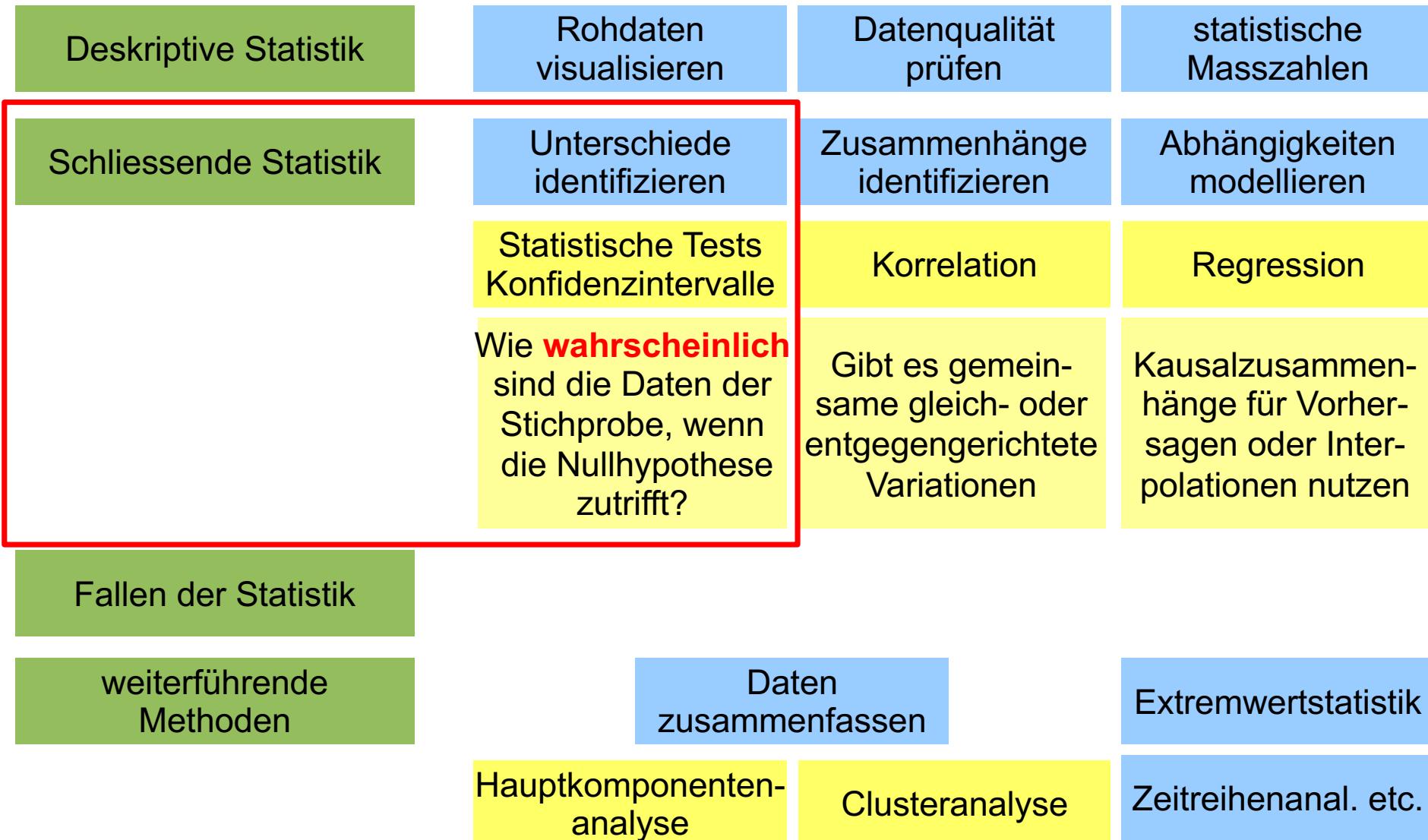
u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

VERTEILUNGEN

Statistische Datenanalyse (Aufbau dieser Vorlesung)



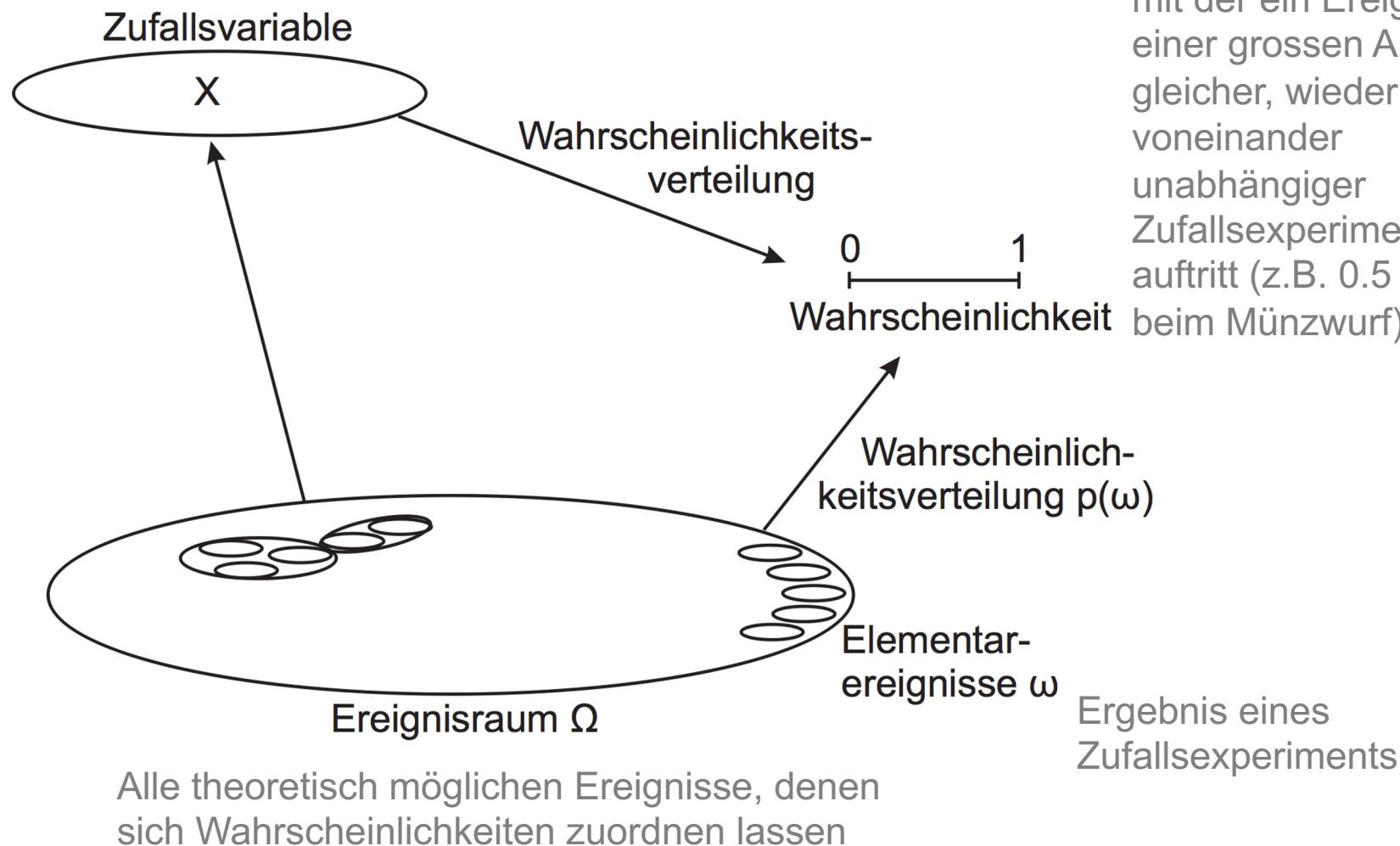
Ereignisraum, Wahrscheinlichkeit, Zufallsvariable

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Variable die den Ergebnissen
eines Zufallsexperiments Werte
(Realisierungen) zuordnet



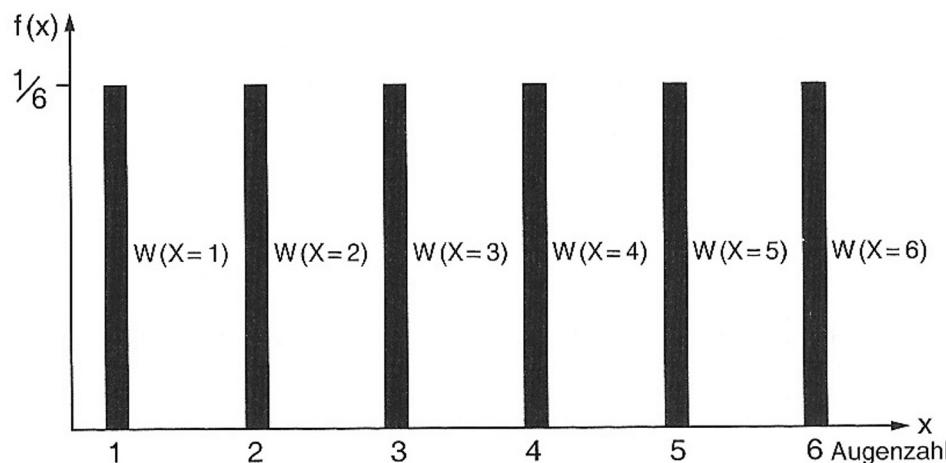
Theoretische Verteilungen diskreter Zufallsvariablen

U^b

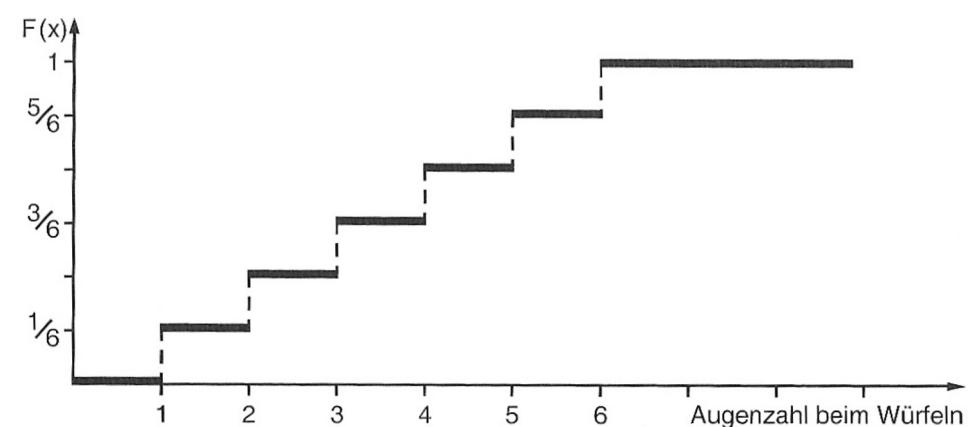
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

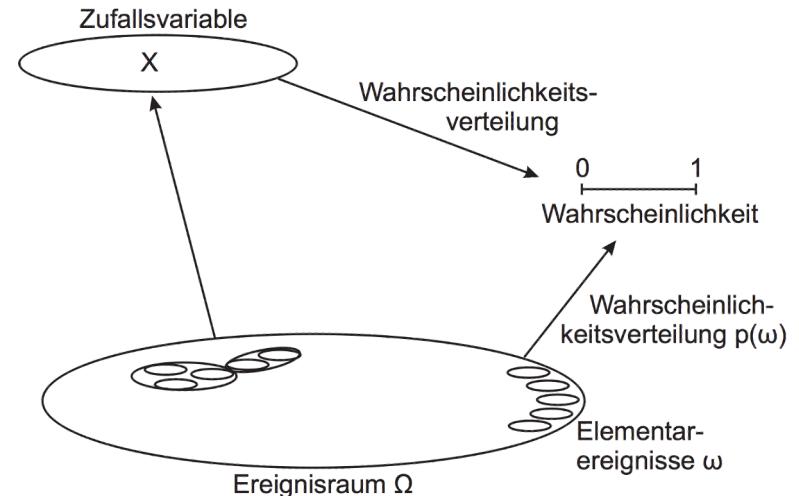
Theoretische **Wahrscheinlichkeitsfunktion** einer diskreten Variablen mit Gleichverteilung



(Kumulative)
Verteilungsfunktion einer diskreten Variablen mit Gleichverteilung



Ereignisraum, Wahrscheinlichkeit, Zufallsvariable



Beispiel: Diskrete Räume

- > Ihr würfelt mit **zwei** Würfeln
- > Der **Ereignisraum** Ω , mit $\omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$
- > **Wahrscheinlichkeitsverteilung** der Elementarereignisse:
- > Jedes Elementarereignis ω hat $p(\omega) = 1/36$ (Gleichverteilung)
- > p : probability
- > **Zufallsvariablen** X können im Ereignisraum definiert werden.
- > Beispiel: Anzahl der insgesamt geworfenen Augen $X = \{2,3,4,5,\dots,12\}$
- > $P(X=x) = (6-|x-7|)/36$
- > P steht für Probability, also Wahrscheinlichkeit

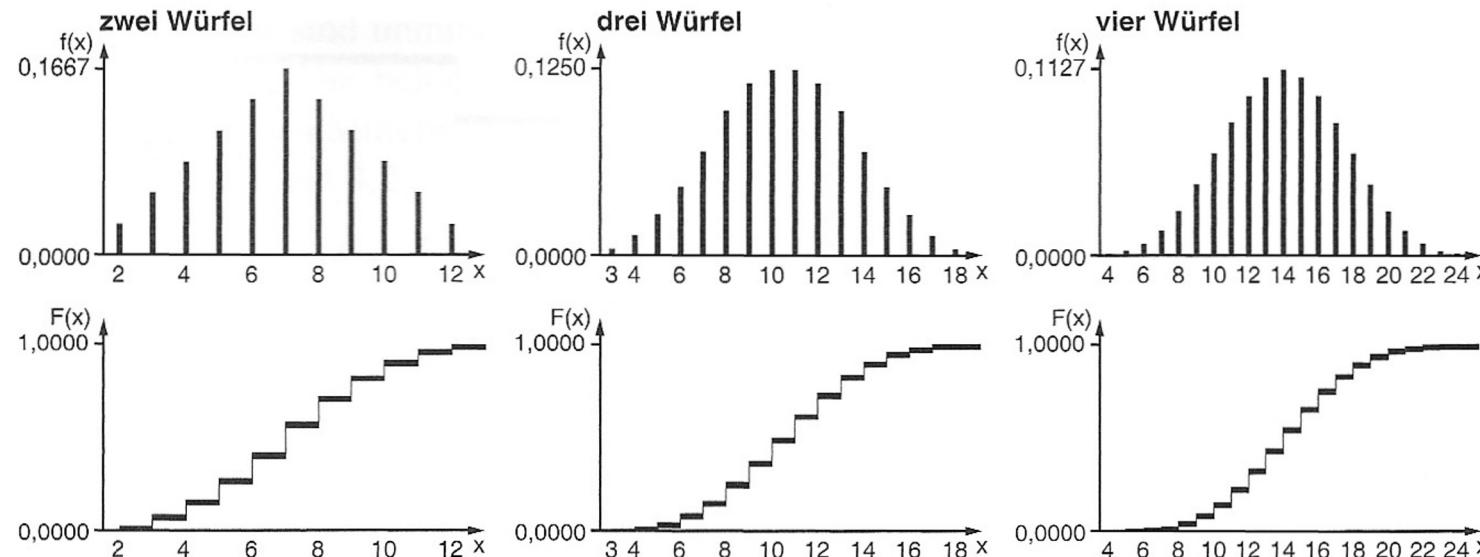
Theoretische Verteilungen diskreter Zufallsvariablen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Wahrscheinlichkeitsfunktion (oben) und Verteilungsfunktion (unten)
der Zufallsvariablen Augensumme



Mittelwert der Wahrscheinlichkeitsfunktion = Erwartungswert

Ereignisraum, Wahrscheinlichkeit, Zufallsvariable

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Beispiel: Stetige/kontinuierliche Räume (z.B. Temperatur)

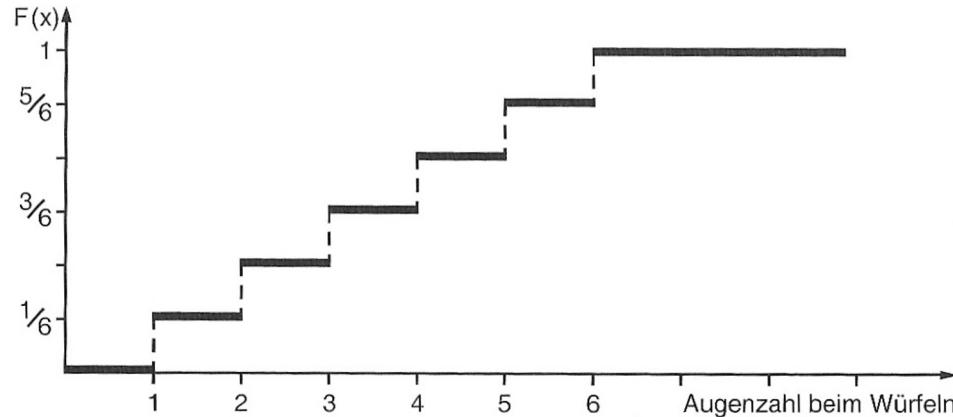
- > Der Ereignisraum Ω ist unendlich gross
- > Zufallsvariablen X
- > Temperatur eines zufällig ausgewählten Zeitpunkts
- > $P(X = x) = 0$
- > Wahrscheinlichkeiten sind definiert für Intervalle
- > $P(X \leq x)$
- > $P(x_1 \leq X \leq x_2)$

Theoretische Verteilungen von Zufallsvariablen

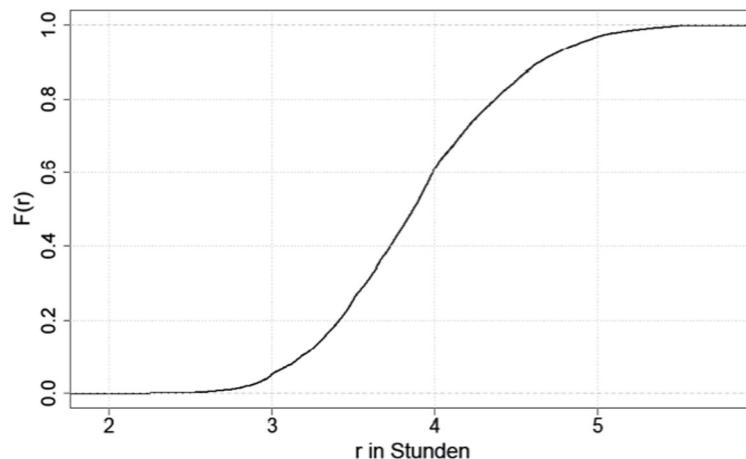
u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



(Kumulative)
Verteilungsfunktion einer
diskreten Variablen mit
Gleichverteilung



Verteilungsfunktion einer stetigen Variablen

$$F(r) = P(X \leq r)$$

Empirische Wahrscheinlichkeitsfunktion einer normal- diskreter Zufallsvariablen

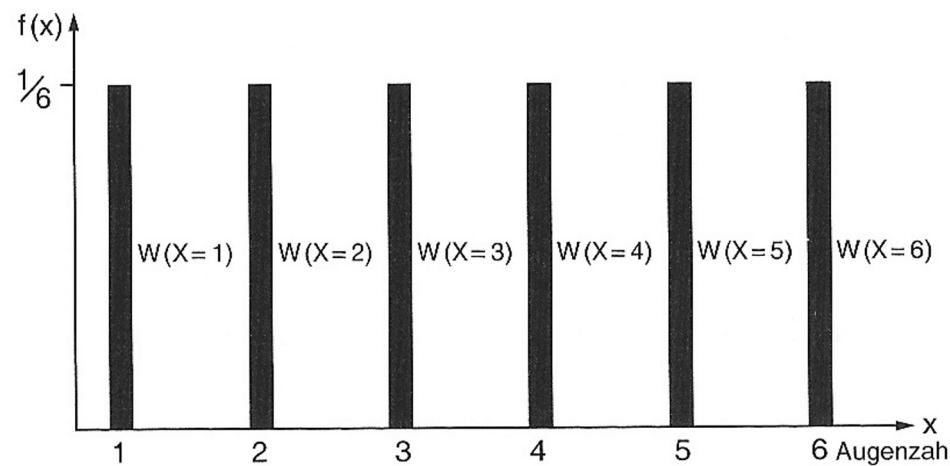
/Theoretische
/ Häufigkeit
/ gleichverteilten,
/ stetiger

u^b

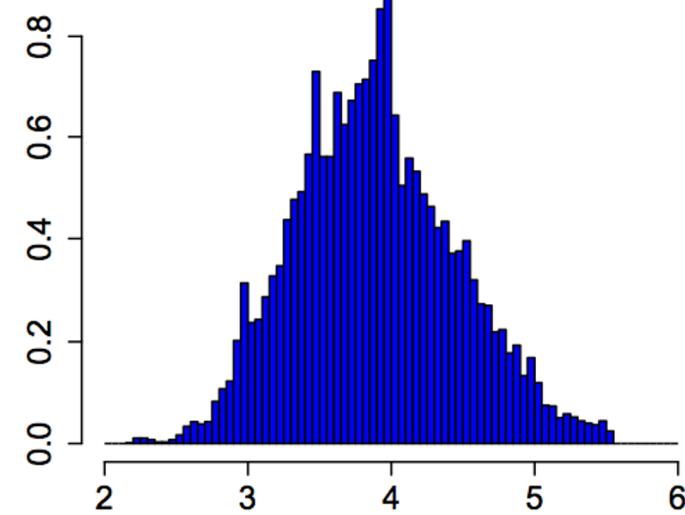
b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

?



?



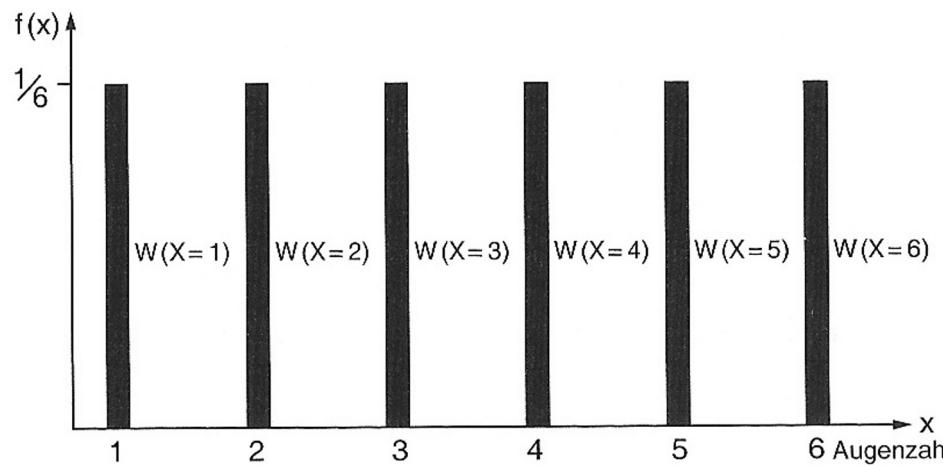
Theoretische/Empirische Wahrscheinlichkeitsfunktion/Häufigkeit einer gleich-/normalverteilten, diskreter/stetiger Zufallsvariablen

u^b

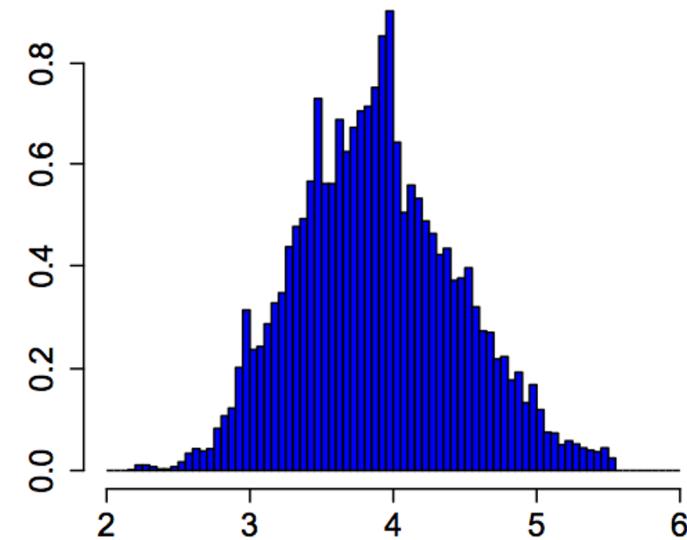
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Theoretische **Wahrscheinlichkeitsfunktion** einer diskreten Variablen mit Gleichverteilung



Empirische **Häufigkeit** einer stetigen Variablen mit Normalverteilung



Histogramm vs. Wahrscheinlichkeitsdichtefunktion

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

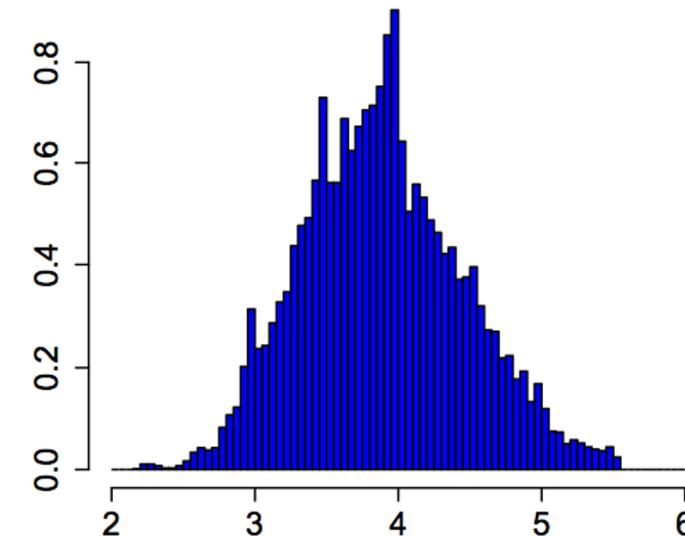
Ein **Histogramm** ist eine graphische Darstellung der **Häufigkeitsverteilung** von stetigen, metrisch skalierten Merkmalen (Kardinalskala). Dies erfordert eine Einteilung der Daten in Klassen.

Daumenregel für die Anzahl an Klassen

Anzahl der Messungen	Balkenzahl
<50	5 bis 7
50 bis 100	6 bis 10
100 bis 250	7 bis 12
>250	10 bis 20

Histogramm einer empirischen Stichprobe:

- **absolute** Anzahl n der Elemente pro Klasse auf der y-Achse oder
- **relative** Anzahl der Elemente auf der y-Achse, d.h. Fläche aller Säulen summiert sich zu 1 auf



Histogramm vs. Wahrscheinlichkeitsdichtefunktion

u^b

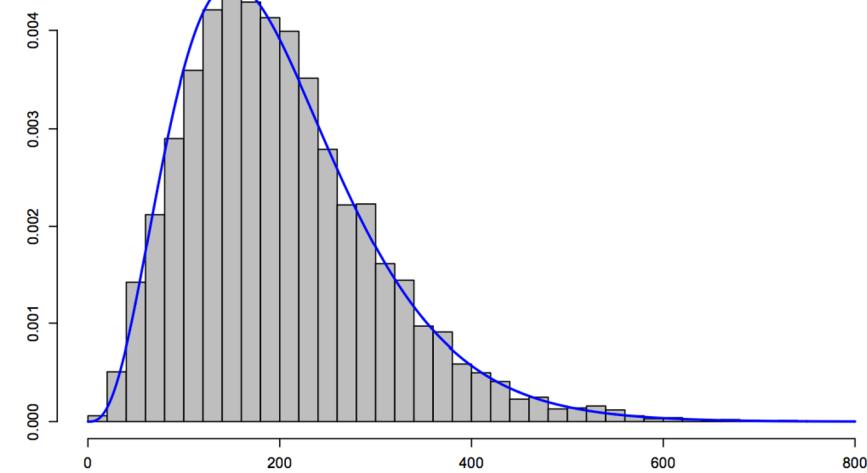
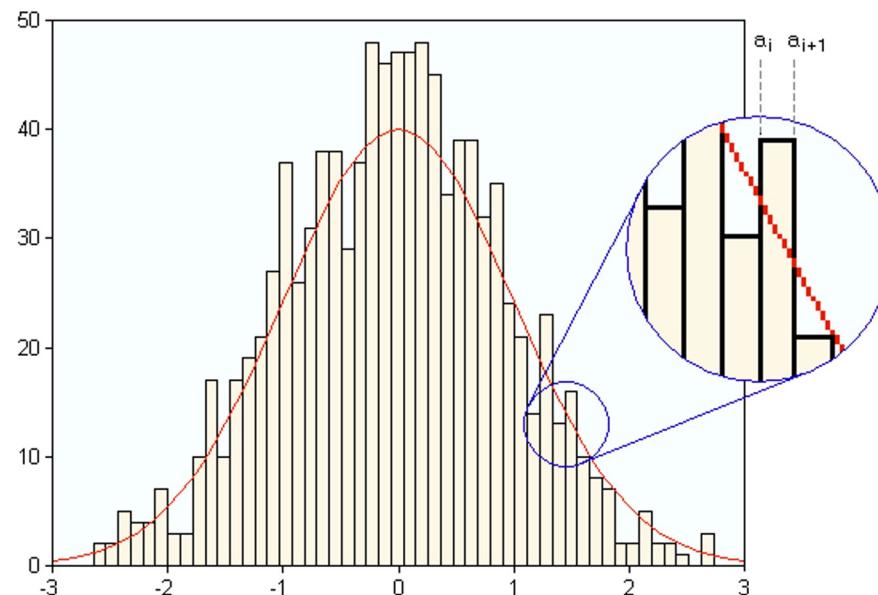
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Histogramm zeigt die
Häufigkeitsverteilung, hier von
empirischen Daten.

Es kann genutzt werden, um die
Wahrscheinlichkeitsdichte der
Grundgesamtheit abzuschätzen

Wahrscheinlichkeitsdichtefunktion
und „idealisiertes“ Histogramm einer
theoretischen Verteilung



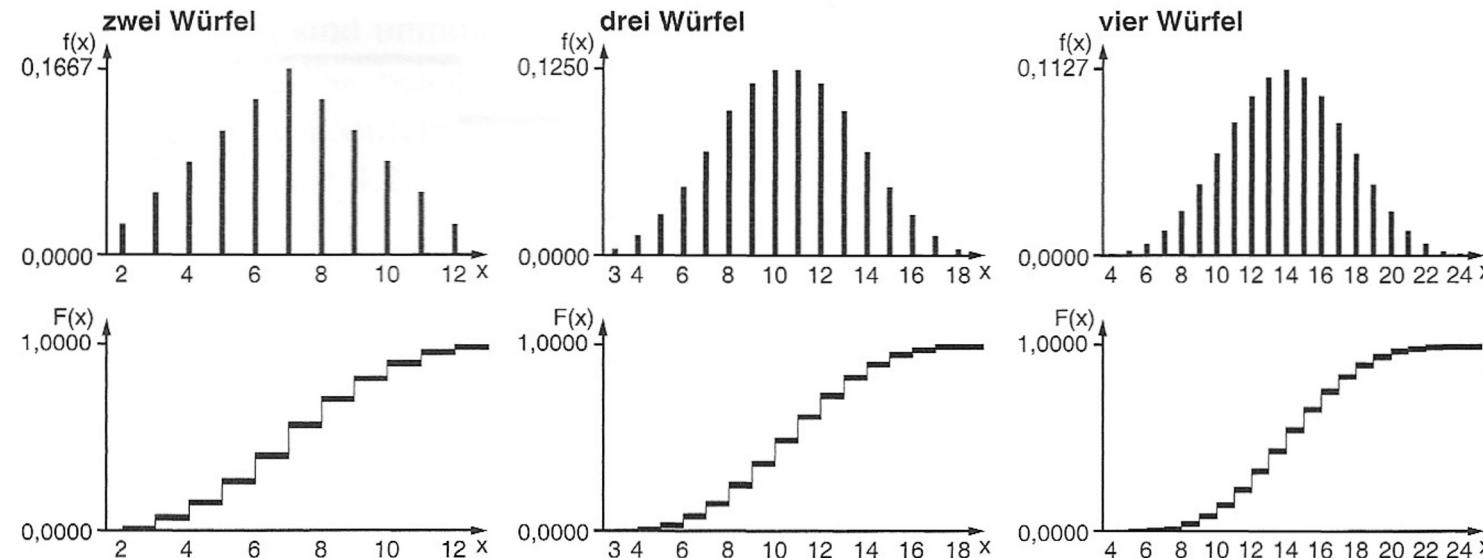
Theoretische Verteilungen diskreter Zufallsvariablen

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Wahrscheinlichkeitsfunktion (oben) und Verteilungsfunktion (unten)
der Zufallsvariablen Augensumme



Mittelwert der Wahrscheinlichkeitsfunktion = **Erwartungswert**

Nähert sich **Normalverteilung** mit steigender Anzahl an Würfeln

Zentraler Grenzwertsatz



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Der zentrale Grenzwertsatz ist ein Hauptsatz in der theoretischen Statistik und besagt im Allgemeinen, dass die Summe von stochastisch unabhängigen Zufallsvariablen annähernd normalverteilt ist.

- > n sollte nach einer groben Faustformel mindestens **30** sein, damit die Summenformel als so gut wie normalverteilt angesehen werden kann.
- > Daher die Daumenregel, dass Stichproben möglichst mindestens einen Umfang von 30 haben sollten.
- > Der Verteilungstyp muss nicht bekannt sein, die Zufallsvariablen müssen nicht normal oder symmetrisch verteilt sein,
- > allerdings muss eine Varianz existieren

Beispiel:

- > Die Augenzahlensumme vieler gleichverteilter Würfelwürfe ist normalverteilt.

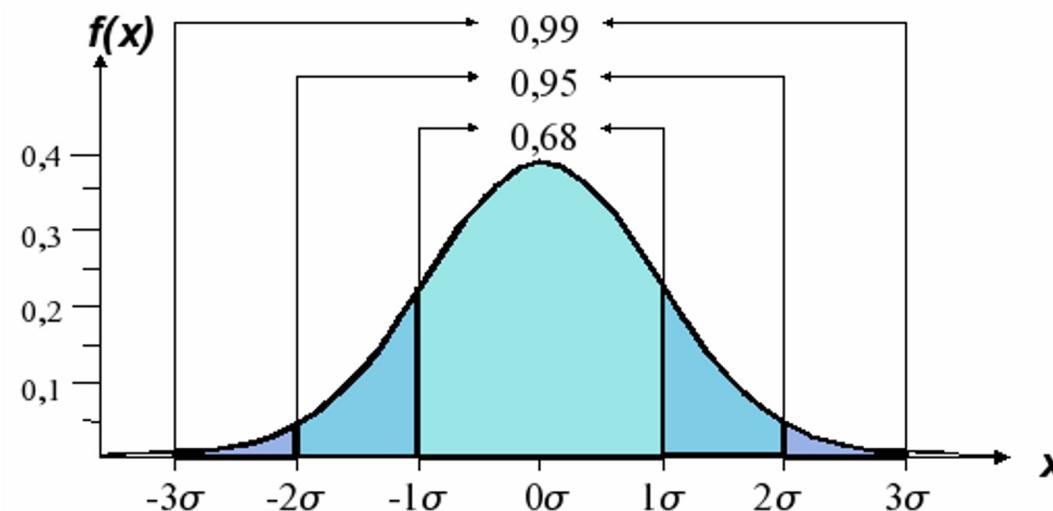
Zentraler Grenzwertsatz

```
> # 1000000 gleichverteilte Zufallszahlen zwischen 0 und 10  
erzeugen  
> r <- runif(1000000,min=0,max=10)  
> hist(r)    # Histogramm der Zufallszahlen  
> mean(r)   # arithmetisches Mittel der Zufallszahlen  
> # erzeuge leere Datenmatrix für 1000 Stichproben  
> s <- matrix(NA,nrow=1000,ncol=50)  
> for (i in 1:nrow(s)) { # Schleife alle Zeilen (1 bis 1000) von s  
  # schreibe 50 zufällige Stichprobenwerte in Zeile i  
>   s[i,] <- sample(r,50)  
> }  
> head(s)      # Kopf der Matrix s  
> hist(s[1,],breaks=10) # Histogramm der ersten Stichprobe (Zeile)  
> hist(s[2,],breaks=10) # Histogramm der zweite Stichprobe (Zeile)  
> m <- apply(s,1,mean) # Mittelwert aller Stichproben (Zeilen)  
> hist(m)           # Histogramm der Stichprobenmittelwerte
```

Normalverteilung

- > Normalverteilungen bzw. Gaussverteilungen oder gaussische Glockenkurve genannt:
- > Die Normalverteilung mit Mittelwert μ und Standardabweichung $\sigma > 0$ (Varianz σ^2) ist definiert als die Verteilung mit Dichtefunktion

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

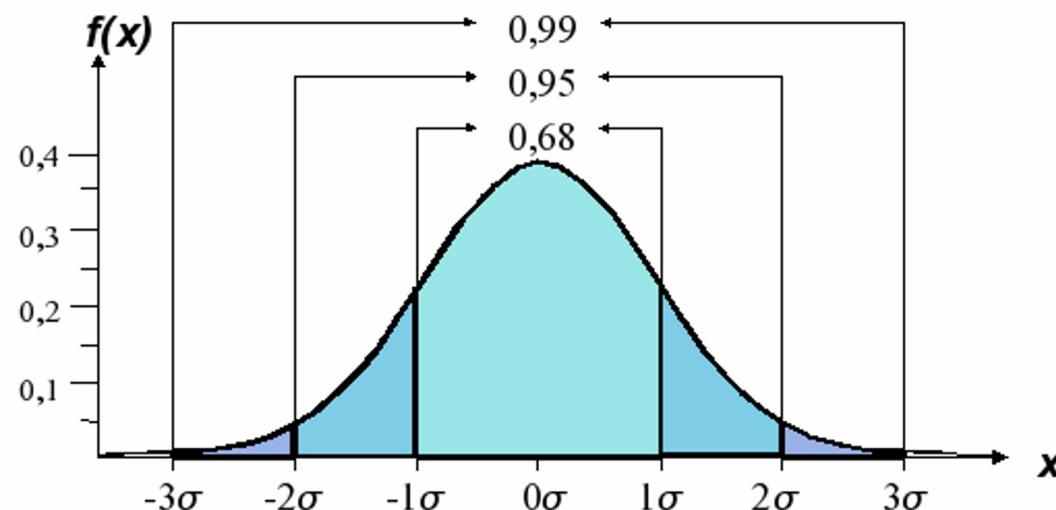


Wahrscheinlichkeitsdichte der Normalverteilung

Normalverteilung

- > Ein/e schweizer Frau/Mann ist im Durchschnitt 165/178cm gross. Wie gross sind 68/95% der Frauen und Männer wenn eine Standardabweichung 6cm beträgt?

Wir können anhand der Verteilungen die Wahrscheinlichkeiten bestimmen



Wahrscheinlichkeitsdichte der Normalverteilung

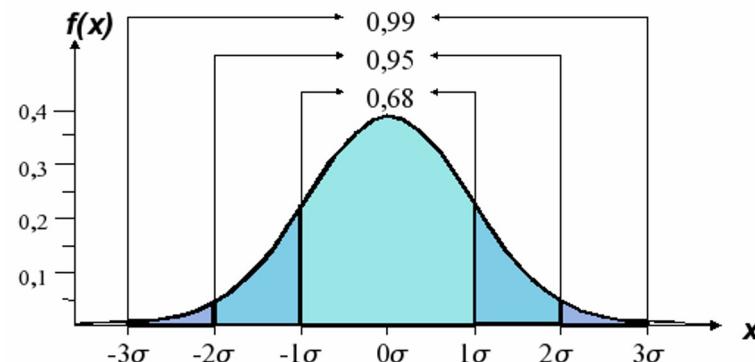
Normalverteilung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > symmetrisch zur Achse $x = \mu$
- > unimodal mit Maximum bei $x = \mu$
- > Wendepunkte bei $x = \mu \pm \sigma$
- > asymptotisch gegen 0
- > Standardnormalverteilung hat Mittelwert $\mu=0$ und Standardabweichung $\sigma = 1$ durch Transformation (Standardisierung)
 $Z = (X - \mu) / \sigma$



Standardisierung / z-Transformation

u^b

b
UNIVERSITÄT
BERN

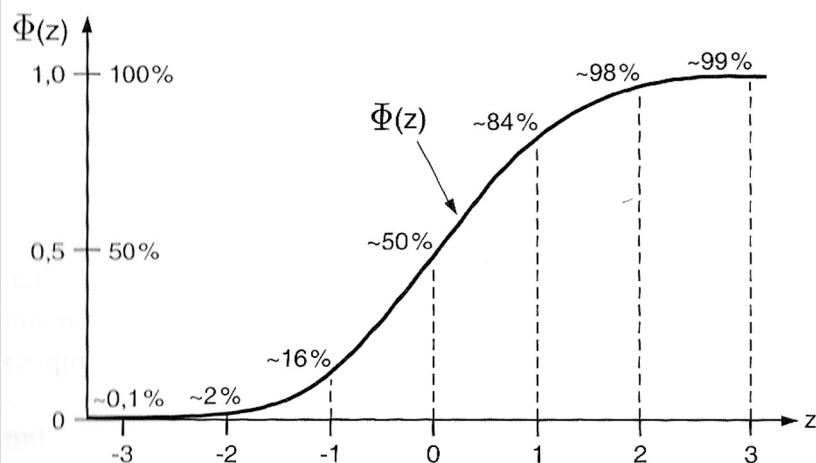
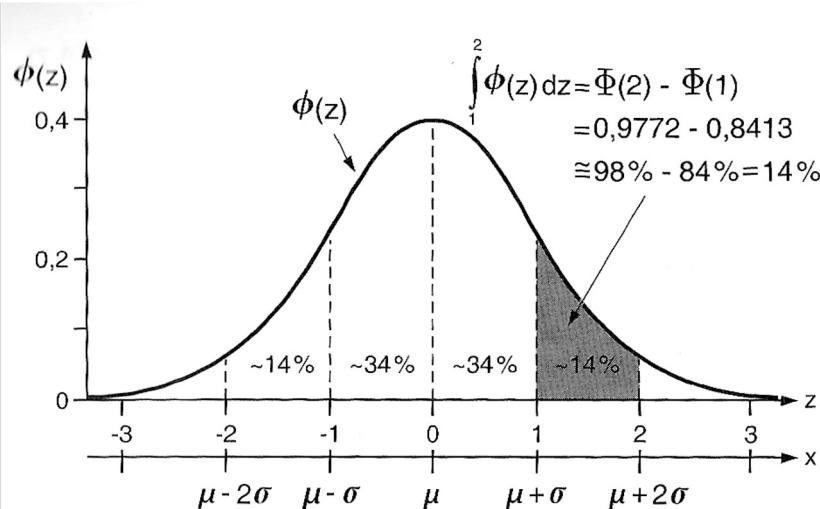
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Zum Vergleich verschiedener Elemente (d.h. Daten mit Bias oder unterschiedlichen Einheiten oder unterschiedlicher Varianz, etc.)
- > Um Mittelwert und Standardabweichung korrigieren, d.h. der Mittelwert von z ist gleich 0 und Standardabweichung gleich 1

$$z = \frac{x - \mu}{\sigma}$$

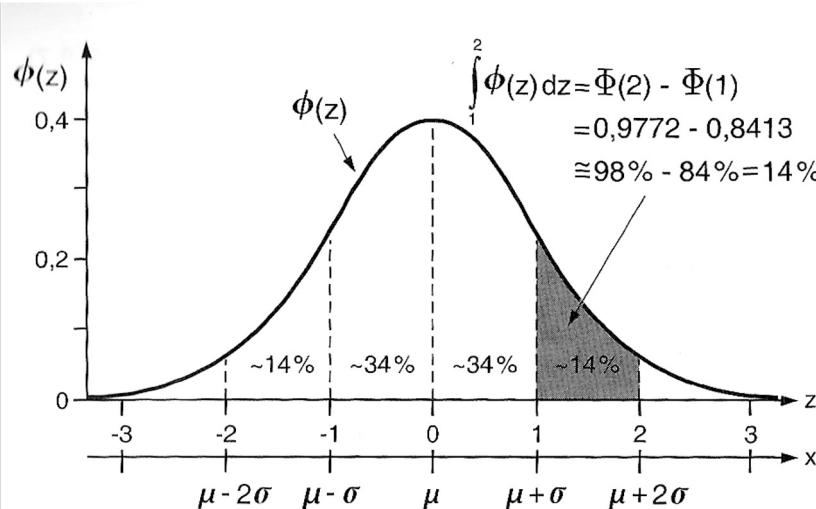
- > Dimensionslos
- > Für symmetrische, unimodale Variablen, da auf Mittelwert und Standardabweichung beruhend

Normalverteilung und Verteilungsfunktionen

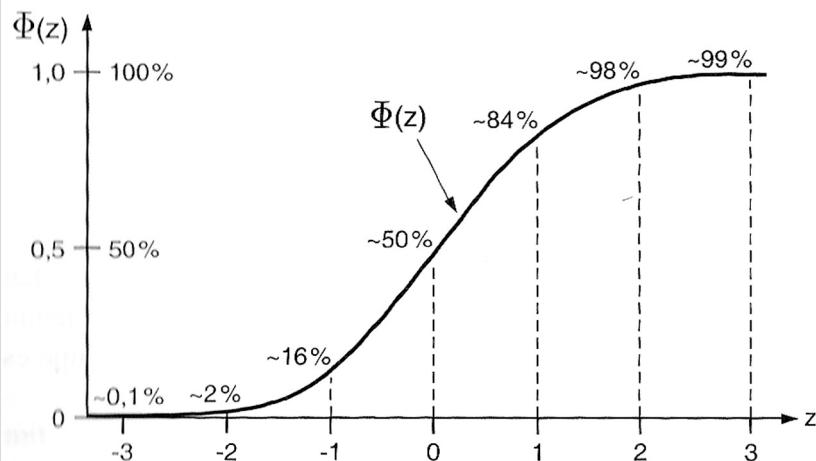


- > Die Verteilungsfunktion $\varphi(z)$ gibt die Unterschreitungswahrscheinlichkeit von z an und entspricht der Fläche unterhalb der Kurve links von z .
- > Überschreitungswahrscheinlichkeit von z ist $1 - \text{Unterschreitungswahrscheinlichkeit}$

Normalverteilung



- > Durchschnitt 165/178cm (Frau/Mann)
gross. Standardabweichung bei beiden 6cm.
- > 99% der Frauen/Männer sind kleiner als?



Take-home messages



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Der zentrale Grenzwertsatz ist ein Hauptsatz in der theoretischen Statistik und besagt im Allgemeinen, dass die Summe von stochastisch unabhängigen Zufallsvariablen annähernd normalverteilt ist.
- > Aus bekannten Verteilungen wie beispielsweise der Normalverteilung lassen sich Wahrscheinlichkeiten ablesen.
- > Schliessende Statistik beruht meist auf Wahrscheinlichkeiten und liefert uns KEINE 100% sicheren Ergebnisse! (D.h. mit welcher Wahrscheinlichkeit trifft das Ergebnis der Stichprobe auf die Grundgesamtheit zu.)

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

STATISTISCHE TESTS

TEIL 1

Bahrenberg I: Kap. 5;
Ernste Anh. B;
Ewing I: Kap. 10

Statistische Datenanalyse (Aufbau dieser Vorlesung)

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen
Fallen der Statistik			
weiterführende Methoden	Daten zusammenfassen	Extremwertstatistik	
	Hauptkomponenten-analyse	Zeitreihenanal. etc.	
	Clusteranalyse		

Grundgesamtheit: die Menge aller Objekte/Untersuchungseinheiten

Liegt eine Grundgesamtheit vor, erübrigt sich die meiste schliessende Statistik und zählen oder messen reicht.

- > "Die Grösse von Kantonen ist nicht normalverteilt."
- > "Die letzten 10 Jahre waren an der MeteoSchweiz Messstation für Bern wärmer als die Jahre 1901-1910».

Stichprobe: Teilmenge der Grundgesamtheit (in diesem Kurs: Zufallsauswahl; es gibt aber auch systematische Stichproben)

Ziel der schliessenden Statistik:

- > Aus Stichproben Aussagen über die Grundgesamtheit machen
- > Schätzen von Kennzahlen der Grundgesamtheit
- > Testen von Hypothesen über die Grundgesamtheit

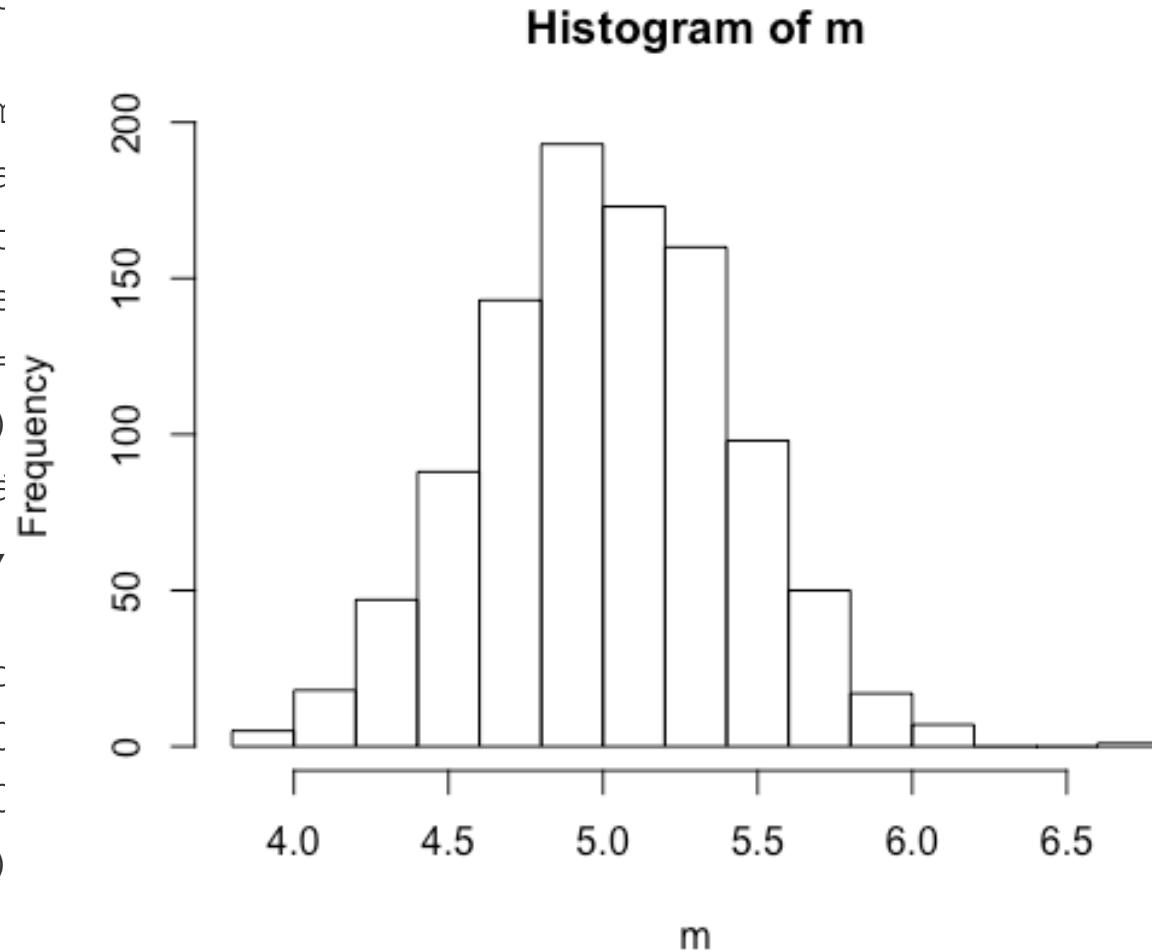
- > Schätzen von Kennzahlen der Grundgesamtheit
(Mittelwert, Streuungsmasse, Stärke von Abhängigkeiten, etc.)
- > Definieren einer **Schätzfunktion** der Stichprobe,
deren Erwartungswert die Kennzahl der Grundgesamtheit ist

Beispiel:

- > Die Stichprobenvarianz (mit $n-1$) ist ein erwartungstreuer Schätzer der Varianz der Grundgesamtheit (mit n ; siehe Folien zur deskriptiven Statistik)
$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$
- > Der Stichprobenmittelwert ist ein erwartungstreuer Schätzer des Mittelwerts der Grundgesamtheit (siehe nächste Folie)

Zentraler Grenzwertsatz

```
> # 1000000 gleichverteilen erzeugen
> r <- runif(1000000,n)
> hist(r)      # Histogramm
> mean(r)      # arithmetische Mittelwerte
> # erzeuge leere Datei
> s <- matrix(NA,nrow=1000000)
> for (i in 1:nrow(s)) {
+   # schreibe 50 zufällige Werte
+   s[i,] <- sample(r, 50)
+ }
> head(s)       # Kopf einer Tabelle
> hist(s[1,],breaks=100)
> hist(s[2,],breaks=100)
> m <- apply(s,1,mean)
> hist(m)
```



Stichprobe vs. Grundgesamtheit

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Wie können wir aus relativ kleinen Stichproben Informationen über die Grundgesamtheit ziehen, z.B.:

- > Aus einer Befragung von 10000 Personen, das Ergebnis einer Wahl
- > Aus 50 Proben, ob ein ganzer Schlachthof frei von Salmonellen ist

Konzept:

Beim Start zum Engadiner Skimarathon wird ein Bus vermisst. Bei der Suche findest du einen Parkplatz einen Bus. Du schaust in den Bus und stellt fest, dass das durchschnittliche Alter der Personen vermutlich bei ca. 80 Jahren liegt.

Denkst ihr dies ist der vermisste Bus oder sucht ihr weiter?

Warum?

Stichprobe vs. Grundgesamtheit

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Konzept:

Beim Start zum Engadiner Skimarathon wird ein Bus vermisst. Bei der Suche findest du einen Parkplatz einen Bus. Du schaust in den Bus und stellt fest, dass das durchschnittliche Alter der Personen vermutlich bei ca. 80 Jahren liegt.

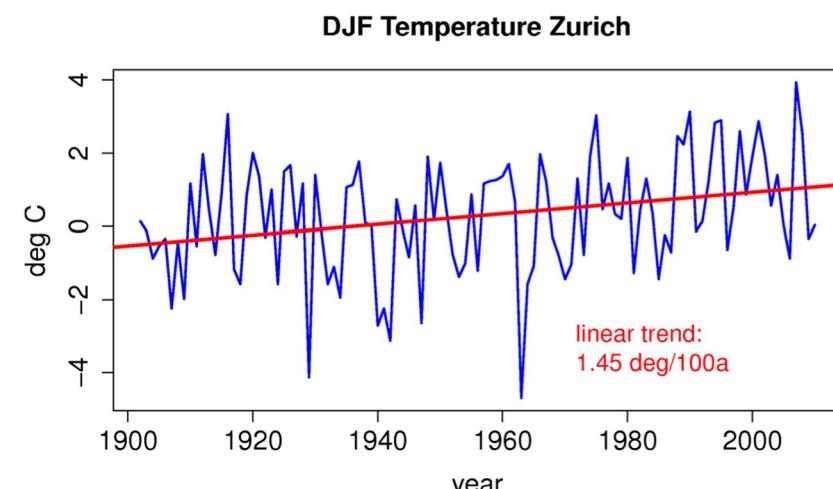
Ihr denkt vermutlich, es ist sehr unwahrscheinlich. Marathonläufer sind normalerweise eher jünger und es ist unwahrscheinlich, dass so viele von den ältesten Teilnehmern in einem Bus gelandet sind. Daher beschliesst ihr, die Suche fortzusetzen.

Eure Annahme ist, das ein zufällige Stichprobe aus allen Teilnehmern ungefähr die gleichen Merkmale haben sollte wie Grundgesamtheit.

Was kann man mit Hypothesen testen?

- > Sind Parameter (Kennzahlen wie Mittelwert) zweier (mehrerer) Grundgesamtheiten aus denen die Stichproben stammen gleich?
Wirkt z.B. ein neues Medikament besser als ein Placebo?
- > Verteilung der Grundgesamtheit aus der die Stichprobe gezogen wurde ist gleich einer bestimmten Verteilung, z.B. Normalverteilung ($N_{\mu; \sigma}$)
- > Parameter der Grundgesamtheit gleichen vorgegebenen Werten, z.B. μ =Konstante oder Korrelationskoeffizient $\rho=0$

Ist dies ein echter Trend oder nur ein Rauschen?



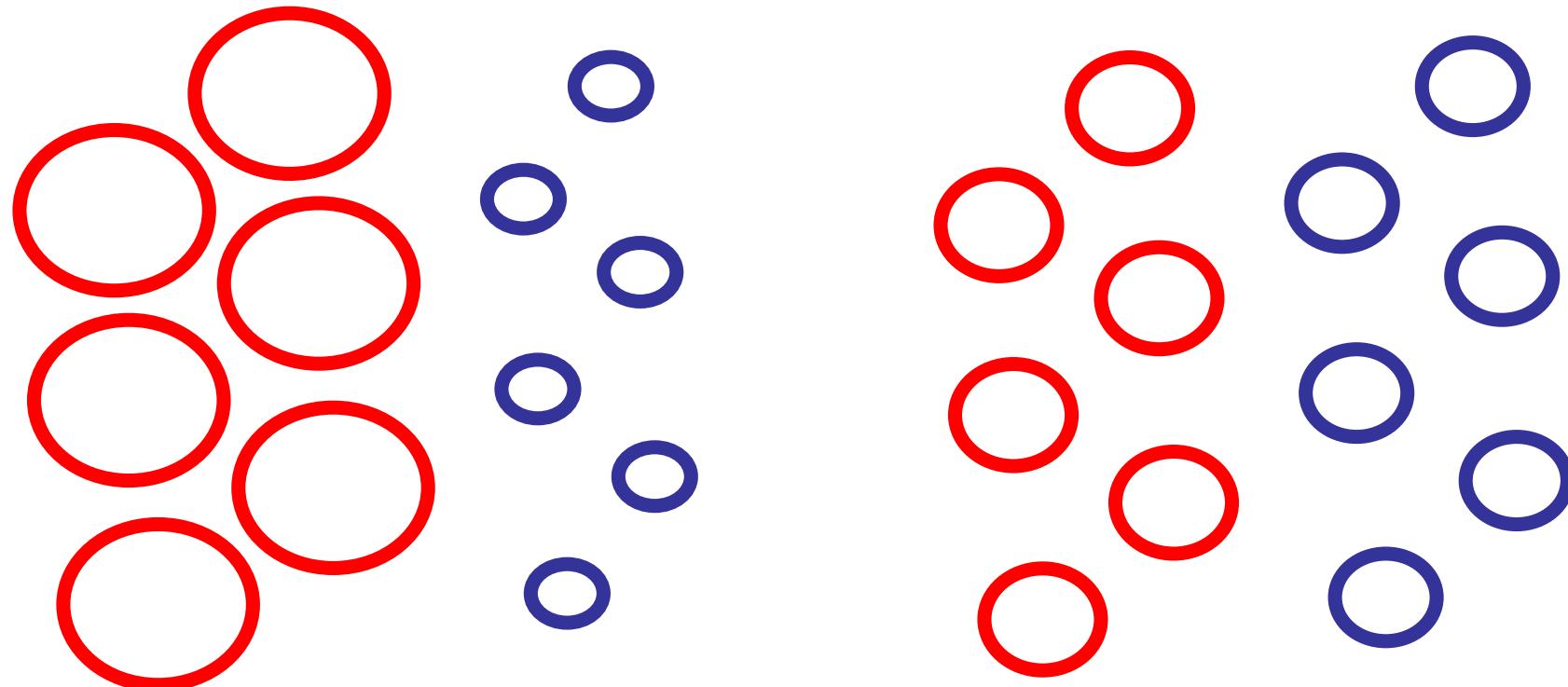
Visualisiert das Problem Statistik muss nicht abstrakt sein!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Ob zwei Gruppen bzw. Stichprobe und Grundgesamtheit statistisch signifikant unterschiedlich sind, hängt an der **Grösse des Effekts**



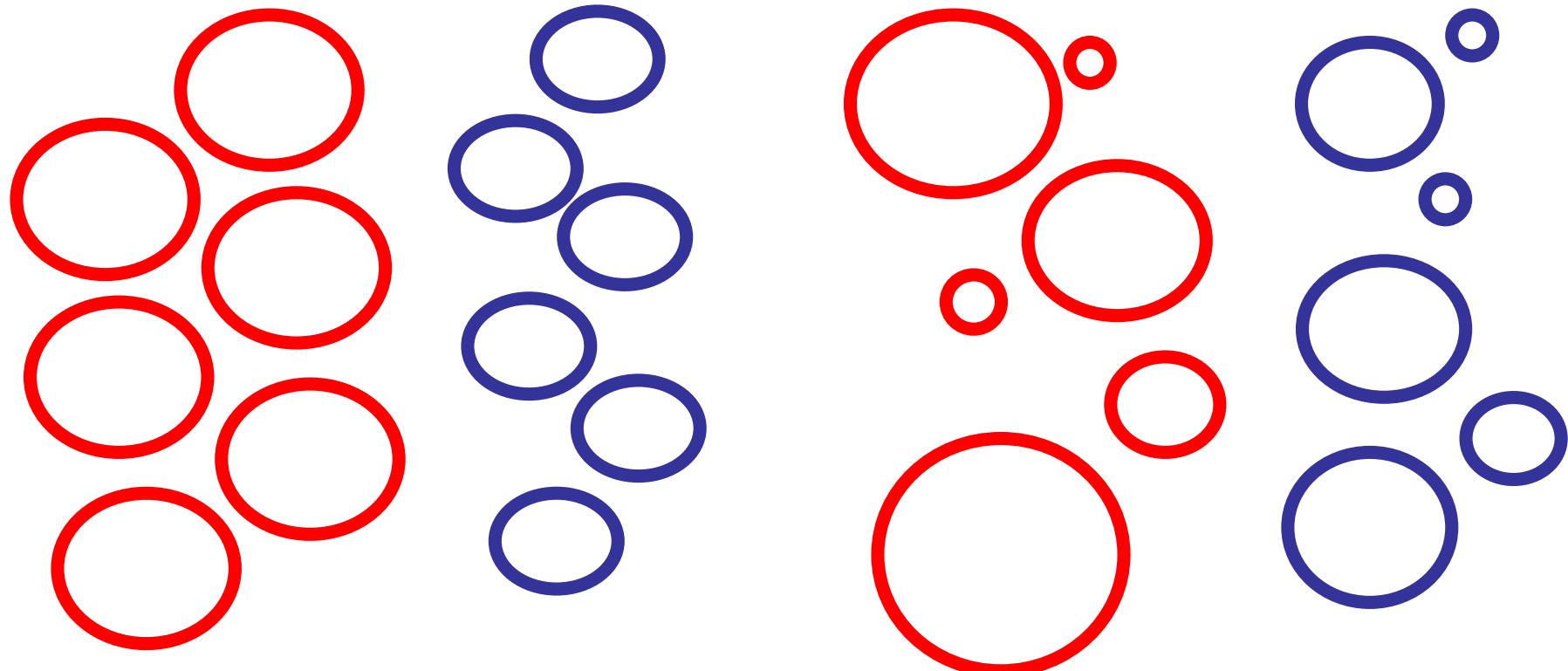
Visualisiert das Problem Statistik muss nicht abstrakt sein!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Ob zwei Gruppen bzw. Stichprobe und Grundgesamtheit statistisch signifikant unterschiedlich sind, hängt an der **Variation der Stichproben**



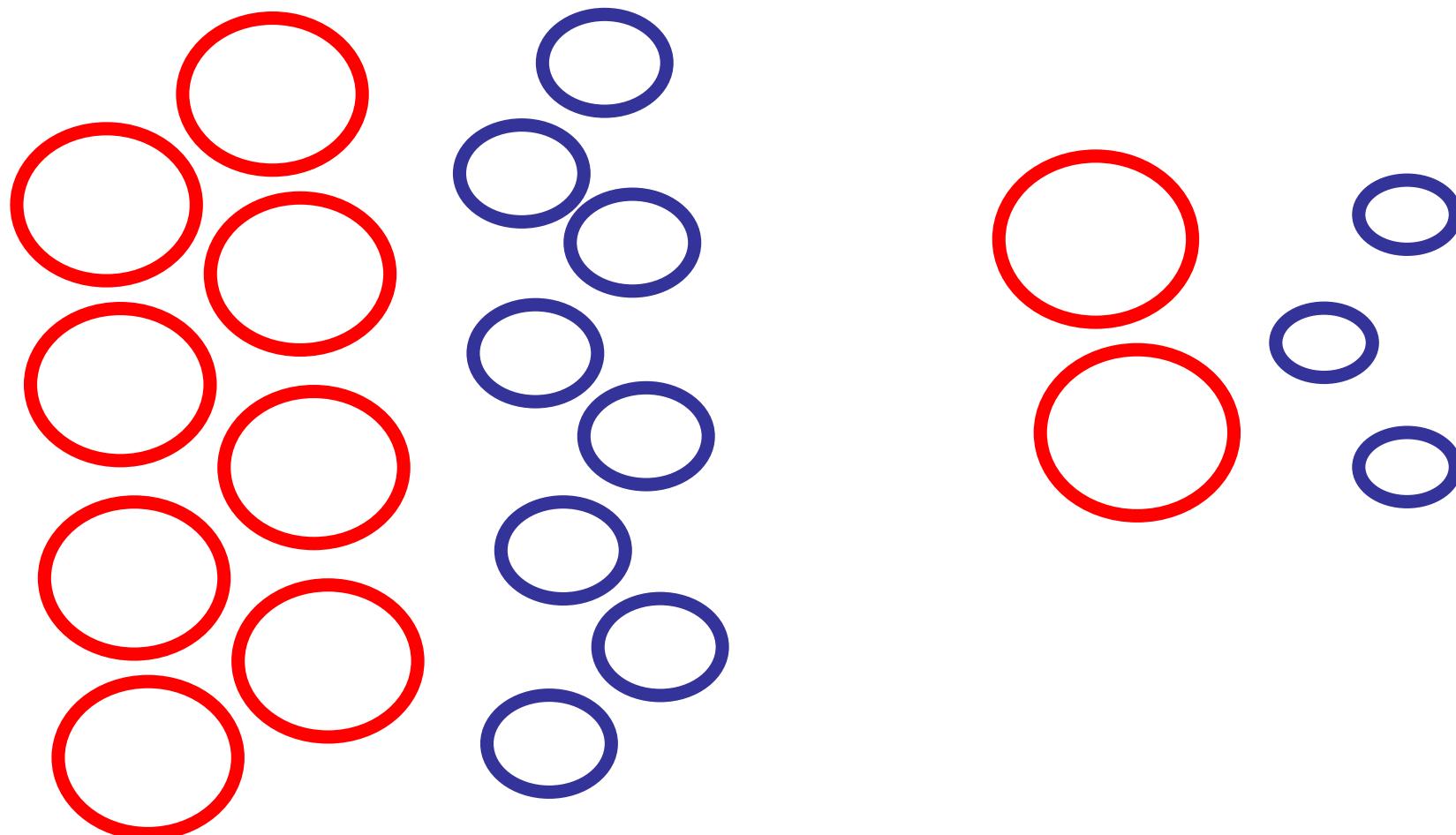
Visualisiert das Problem Statistik muss nicht abstrakt sein!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Ob zwei Gruppen bzw. Stichprobe und Grundgesamtheit statistisch signifikant unterschiedlich sind, hängt an der **Stichprobengröße**



Standardabweichung vs. Standardfehler

- > Die Standardabweichung misst die Variation in der Grundgesamtheit (bzw. deren beste Schätzung aus der Stichprobe), z.B. Die Teilnehmer am Skimarathon haben ein mittleres Alter von 40 Jahren mit einer Standardabweichung von 10 Jahren, wobei wir annehmen, dass die Alter ungefähr normalverteilt ist.
- > D.h. 68% der Läufer sind zwischen ? und ? Jahre und 95% zwischen ? und ? Jahre alt.

Standardabweichung vs. Standardfehler

... mittleres Alter von 40 Jahren mit einer Standardabweichung von 10 Jahren

D.h. 68% der Läufer sind zwischen 30 und 50 Jahre und 95% zwischen 20 und 60 Jahre alt.

Der **Standardfehler** misst die Abweichungen der Stichprobenmittelwerte vom Mittelwert der Grundgesamtheit, z.B. wenn wir Busse mit Läufern beladen, was ist dann die Abweichung des mittleren Alters der Läufer in den Bussen?

D.h. Das mittlere Teilnehmeralter der Grundgesamtheit ist immer noch 40 Jahren. In einem Bus sind die Leute jedoch im Mittel 41 Jahre alt, im nächsten 38 Jahre, ...

Verbindung zwischen beiden:

Der Standardfehler ist die Standardabweichung der Stichprobenmittel, wenn mehrere Stichproben erhoben werden, d.h. ein grosser Standardfehler bedeutet die Stichprobenmittel weichen relativ weit vom Mittel der Grundgesamtheit ab.

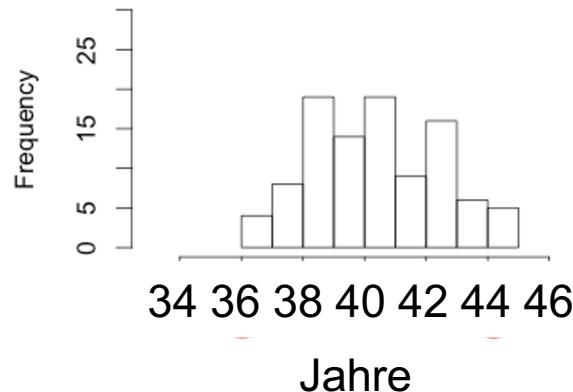
Standardabweichung vs. Standardfehler

U^b

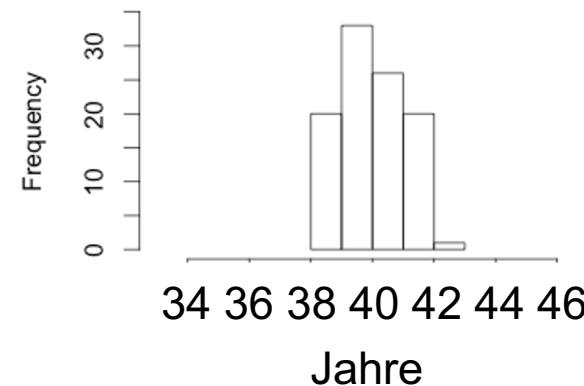
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

100 Stichprobenmittel, n=20
entspricht Bus mit 20 Personen



100 Stichprobenmittel, n=100
entspricht Bus mit 100 Personen



Bei einer grösseren Stichprobe (rechts) ist es unwahrscheinlicher, dass die Stichprobenmittelwerte vom Mittel der Grundgesamtheit abweichen.

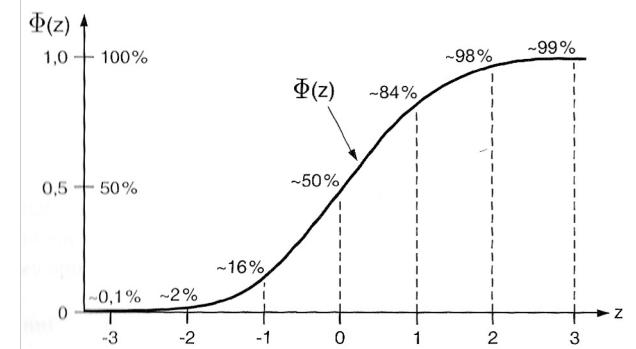
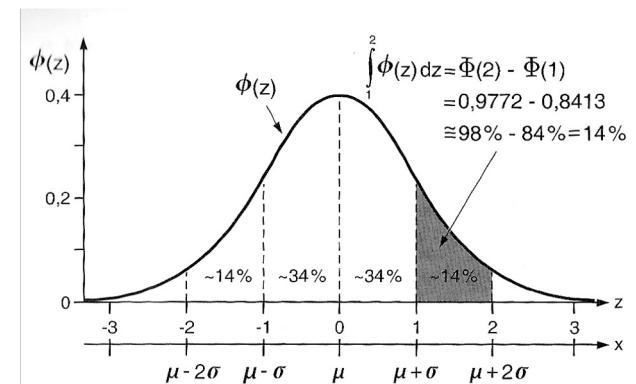
$$\text{Standardfehler: } s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

mit s , der Standardabweichung der Stichprobe, bzw. der Varianz s^2
d.h. der Standardfehler ist gross, wenn die Standardabweichung der Stichprobe gross ist und nimmt mit steigender Stichprobengrösse (n) ab.

Die **Standardfehler** sind dank des zentralen Grenzwertsatzes **normalverteilt!**

Statistisch begründete Entscheidung

- > Mittleres Alter der Marathonteilnehmer beträgt 40 Jahre und die Standardabweichung 10 Jahre.
- > Im gefundenen Bus beträgt das mittlere Alter 80 Jahre. Nehmen wir an im Bus würden 50 Personen sitzen d.h.
 - > der Standardfehler wäre $s=\sqrt{10 \text{ Jahre}^2/50}=1.41 \text{ Jahre}$
 - Die Differenz ist mit $80-40=40$ Jahre grösser als 28 Standardfehler.
- > Aus der Normalverteilung können wir schliessen, dass dies mit >99.9%iger Sicherheit der falsche Bus ist!



Testtheorie

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: **Unschuldig** bis die **Schuld** bewiesen ist

1. Formulierung der **Nullhypothese** H_0 und der **Alternativhypothese** H_A oder H_1

- > Beim Aufstellen der Nullhypothese geht man davon aus, „Alles bleibt beim alten, nichts hat sich geändert“, „Vermutung, die ohne überwältigenden Gegenbeweis weiterhin gelten würde“
- > Was ich zeigen oder beweisen will, gehört normalerweise in die Alternativhypothese
- > Wir nehmen H_A als wahr an, wenn wir Beweise gegen H_0 haben
- > **Alternativhypotesen** können **einseitig (grösser/kleiner)** oder **zweiseitig sein (ungleich)**
- > Das **Gleichheitszeichen** gehört immer in die **Nullhypothese** (kann auch grösser-gleich oder kleiner-gleich sein).

Testtheorie

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: Die Beweise

2. Festlegung des Schwellenwertes / kritisches Signifikanzniveaus,
z.B. $\alpha = 0.05$ (d.h. wir verlangen mindestens 95%ige Wahrscheinlichkeit,
die richtige Entscheidung zu treffen)
3. Ziehung der Stichprobe
4. Zusammenfassung der Daten in einem Wert einer Teststatistik, z.B. t-Wert
des Student t-Test, und überprüfen, ob dieser Wert ein kritisches Niveau
überschreitet.

Testtheorie

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: Beratungen und Überlegungen

ALTERNATIV: Überschreitungswahrscheinlichkeit (p für probability) berechnen.

- > **PROBLEM: p ist von Hand oft komplizierter zu berechnen**
- > p ist zwischen 0 und 1 und gibt die Stärke der Beweise gegen H_0 an
- > Angenommen H_0 ist wahr, wie wahrscheinlich ist es zufällig diesen oder einen grösseren Testwert zu erhalten?
- > Der p-Wert ist die numerische Antwort
- > Je kleiner p, desto stärker die Beweise gegen H_0 (oft 0.05 als Grenze)
- > p-Wert sagt **NICHT**: Wie wahrscheinlich ist es, dass die H_0 wahr ist
- > p-Wert sagt: Wie wahrscheinlich die Daten sind, wenn H_0 wahr ist

Testtheorie

Hypothesen aufstellen

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: Das Urteil

5. Liegt das Ergebnis der Stichprobe innerhalb des Annahmebereichs, wird H_0 angenommen, anderenfalls abgelehnt, d.h.

Option 1:

- > Teststatistik, z.B. t-Wert < kritischer Wert in Tabelle
- > grosser p-Wert ($p > \alpha$, z.B. mit $\alpha=0.05$ für 95%ige Sicherheit)
- > wir schliessen daraus -> Daten sind konsistent mit H_0

wichtig!

Option 2:

- > Teststatistik, z.B. t-Wert > kritischer Wert in Tabelle
- > kleiner p-Wert ($p < \alpha$)
- > wir schliessen daraus -> es gibt ausreichend Beweise, um H_0 abzulehnen und H_A anzunehmen (statistisch signifikant).

Testentscheidung In der GG ist ...	H_0 nicht ablehnen	H_0 ablehnen
H_0 wahr		Fehler 1. Art
H_0 falsch	Fehler 2. Art 	

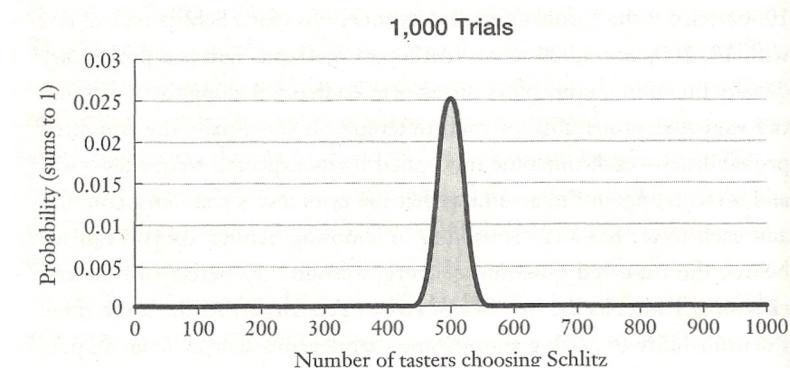
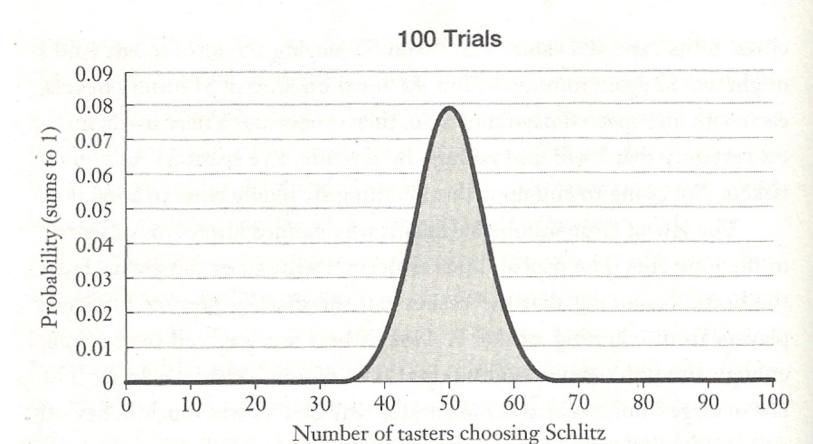
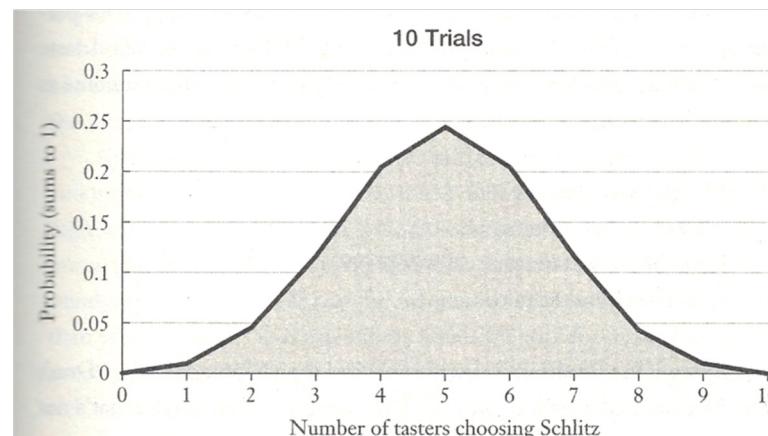
- > Es gibt keine Testverfahren, die gleichzeitig beide Fehlerarten minimieren.
- > Das Signifikanzniveau α legt den Fehler 1. Art fest, den wir in Kauf zu nehmen gewillt sind.
- > Der Fehler 2. Art (β) ist meistens mit nicht so gravierenden Folgen verbunden

Testentscheidung In der GG ist ...	H_0 nicht ablehnen	H_0 ablehnen
H_0 wahr		Fehler 1. Art
H_0 falsch	Fehler 2. Art 	

- > Beispiel SPAM Filter: H_0 = Mail ist kein SPAM.
- > Der Filter schaut, ob es Gründe gibt H_0 zu widerlegen (Auftreten von speziellen Wörtern, etc.)
- > Ein Fehler 1. Art wäre ein Mail in den SPAM Ordner zu verschieben, die gar kein SPAM ist.
- > Der Fehler 2. Art wäre ein Mail durch den Filter zu lassen, die SPAM ist.

Mittelwerte testen

- > μ und σ aus X_{Mittel} und s_x schätzen
- > dies führt bei kleinen Stichproben zu zusätzlichen Unsicherheiten (grosser Standardfehler bei kleinem Stichprobenumfang)
- > es ist unwahrscheinlich, dass die Stichprobe exakt das Mittel der Grundgesamtheit hat
- > dadurch wird Verteilung breiter und flacher (t-Verteilung)

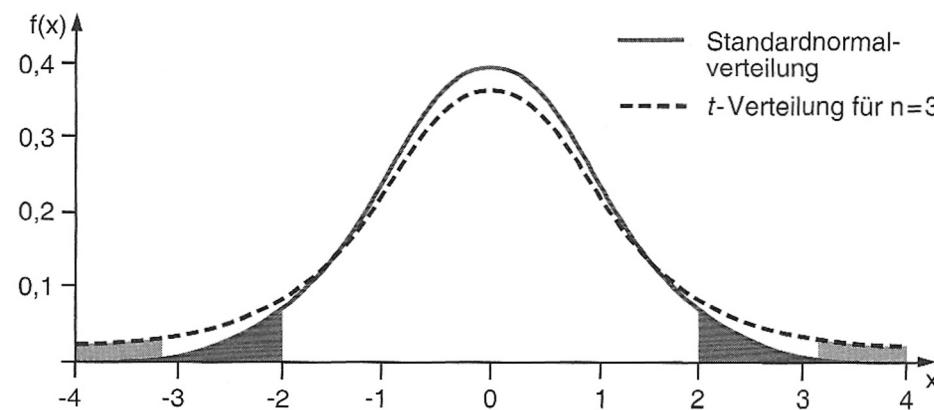


Tests für das arithmetisches Mittel



William Gosset alias Student, 1876-1937

- > Die standardisierte Schätzfunktion des Stichproben-Mittelwerts ist nicht normalverteilt, sondern t-verteilt, wenn die zur Standardisierung des Mittelwerts benötigte Varianz des Merkmals unbekannt ist und mit der Stichprobenvarianz geschätzt werden muss.



- . Symetrisch mit Mittelwert 0
- . Stichprobenmittelwerte folgen t-Verteilung mit $n-1$ Freiheitsgraden
- . bei FG > 30 ist t-Verteilung fast identisch mit Normalverteilung

T-Test für den Mittelwert

(Vergleich von Stichprobemittel und mit fixem Wert,
z.B. bekanntem Mittel der GG oder Grenzwert)

Hypothesen:

H_0 : Erwartete Sept.-Nov. Mitteltemperatur beträgt 9°C ($x = \mu_0$)

H_A : Die Temperatur weicht signifikant von 9°C ab ($x \neq \mu_0$)

Teststatistik: $T = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$ $x=8.75^\circ\text{C}$, $s_x=0.10^\circ\text{C}$, $n=111$, $T=-2.40$

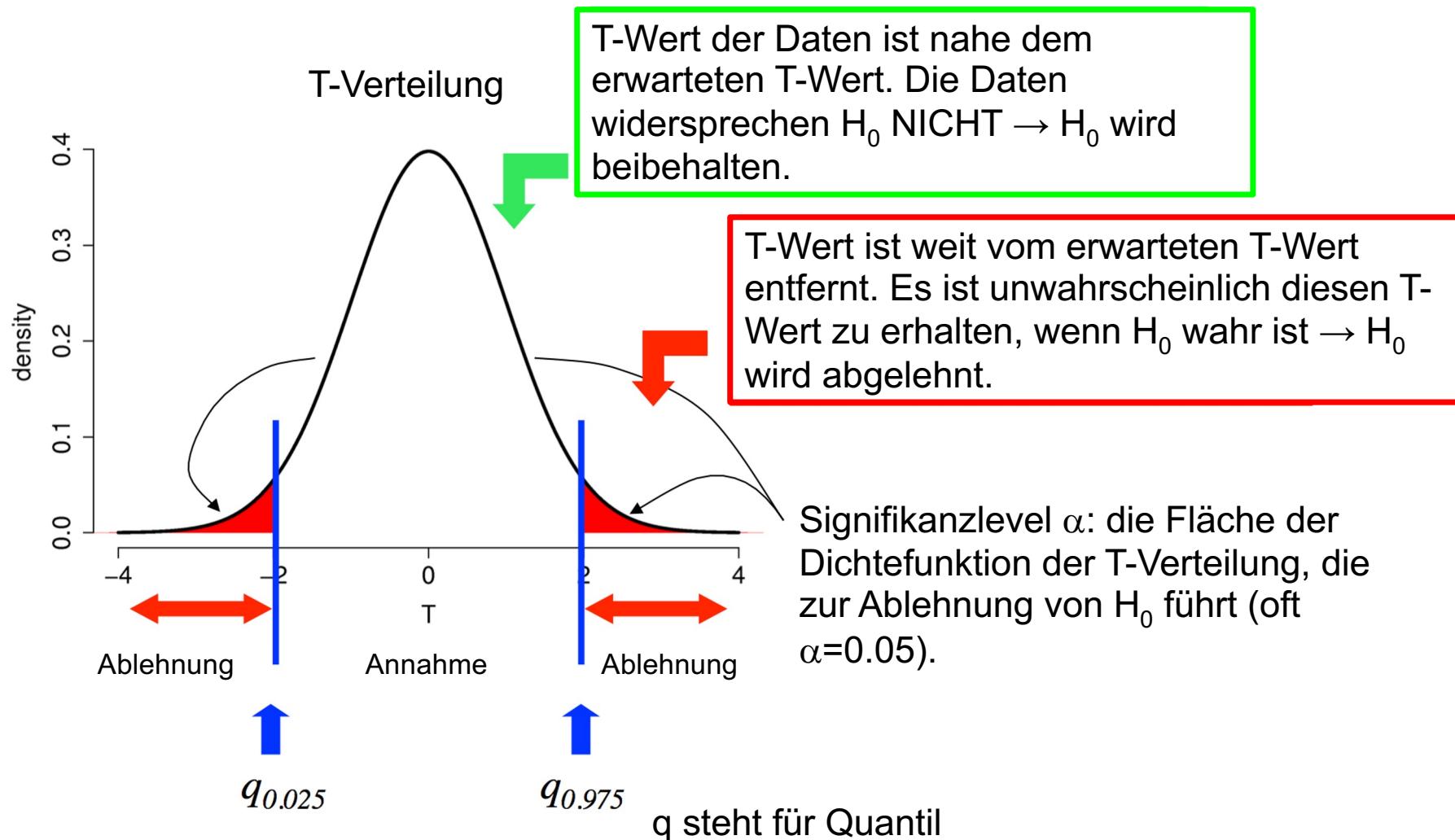
ähnlich der Standardisierung / Z-Transformation / Standardnormalverteilung

mit dem Standardfehler $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$

Der Standardfehler zeigt die theoretische Streubreite des Stichprobenmittelwerts, im Gegensatz zur Standardabweichung, die die reale Streubreite aller Werte der Stichprobe beschreibt. Der Standardfehler wird um so kleiner, je größer der Stichprobenumfang.

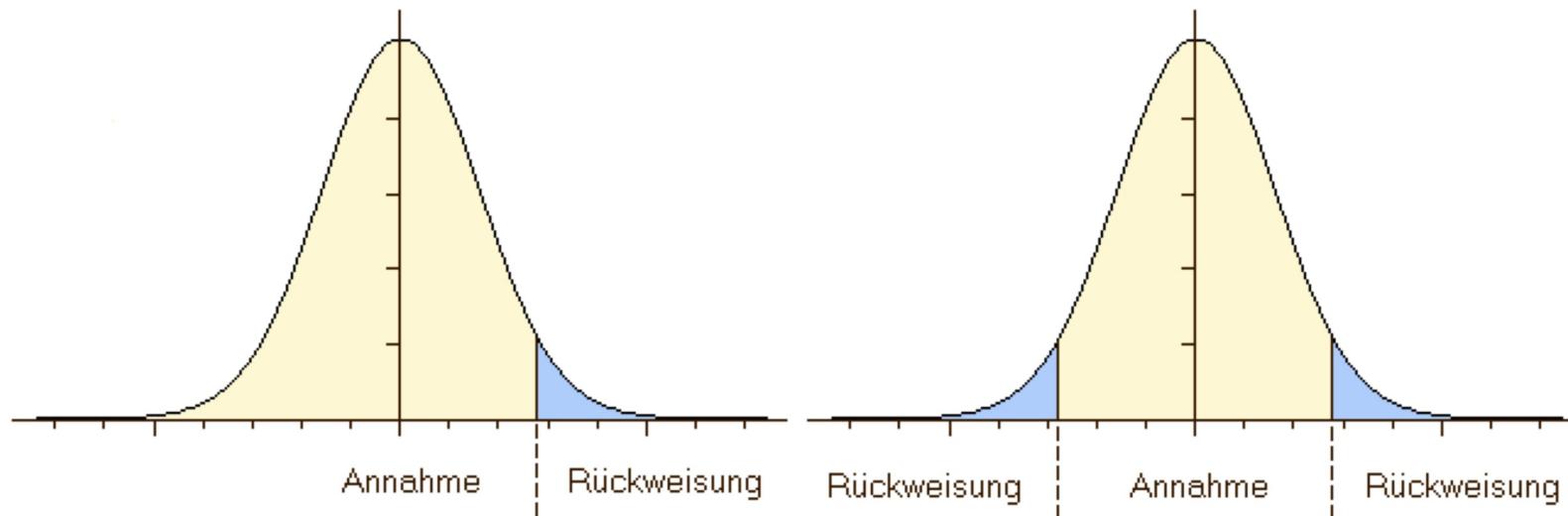
ACHTUNG: s steht für die **Standardabweichung**,
 s^2 für die **Varianz** und
 s_x für den **Standardfehler**.

T-Test für den Mittelwert

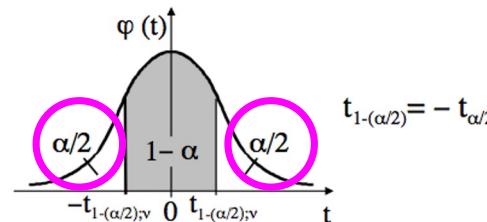
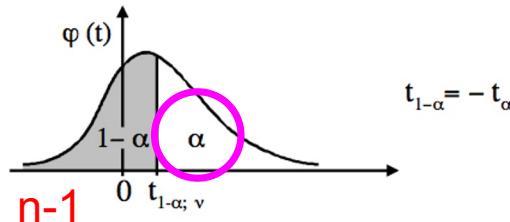


Ein- vs. zweiseitiger Test

- > Bei Parametertests kann H_0 als ein- oder zweiseitiger Test formuliert werden:
- > Zweiseitiger Test: Wir untersuchen Abweichungen vom vermuteten Parameterwert in beide Richtungen ($H_0: \mu = 0$; $H_A: \mu \neq 0$).
- > Einseitiger Test: Wir untersuchen Abweichungen vom vermuteten Parameterwert nur in eine Richtung ($H_0: \mu \geq 0$; $H_A: \mu < 0$ oder $H_0: \mu \leq 0$; $H_A: \mu > 0$)



t-Verteilungs Tabelle

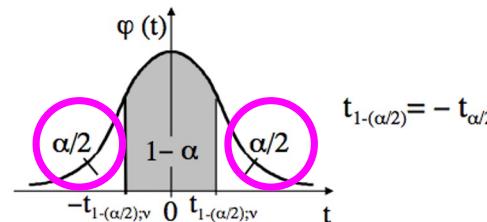
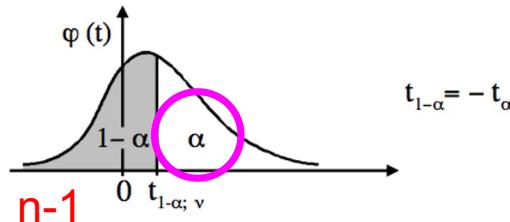


FG

v	Statistische Sicherheit $1-\alpha$					
	0,90	0,95	0,975	0,99	0,995	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3
2	1,886	2,920	4,303	6,965	9,925	22,33
3	1,638	2,353	3,182	4,541	5,841	10,21
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,592	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,443	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
200	1,286	1,652	1,972	2,345	2,601	3,131
500	1,283	1,648	1,965	2,334	2,586	3,107
∞	1,282	1,645	1,960	2,326	2,576	3,090

$n=111$
 $\alpha=0.05$
 Zweiseitiger Test

t-Verteilungs Tabelle



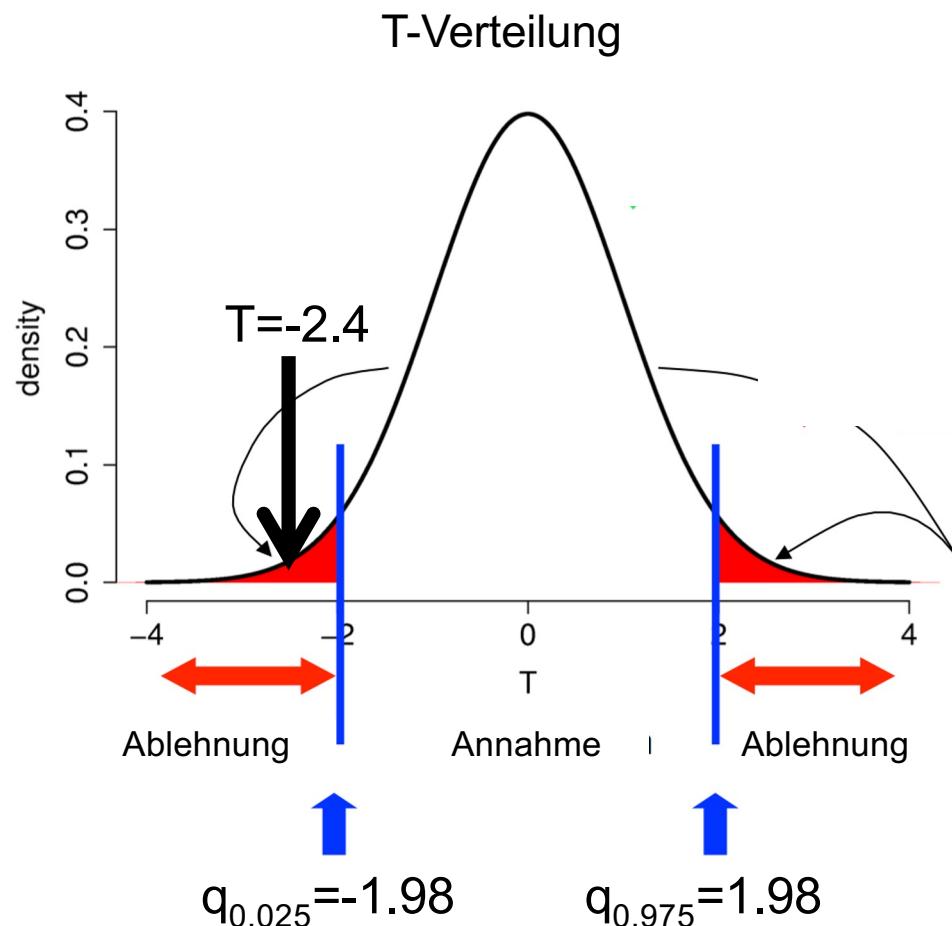
FG

v	Statistische Sicherheit $1-\alpha$					
	0,90	0,95	0,975	0,99	0,995	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3
2	1,886	2,920	4,303	6,965	9,925	22,33
3	1,638	2,353	3,182	4,541	5,841	10,21
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,592	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,443	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
200	1,286	1,652	1,972	2,345	2,601	3,131
500	1,283	1,648	1,965	2,334	2,586	3,107
∞	1,282	1,645	1,960	2,326	2,576	3,090

v	Statistische Sicherheit $1-\alpha$						
	0,80	0,90	0,95	0,98	0,99	0,998	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	1,638	2,353	3,182	4,541	5,841	10,21	12,92
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,992
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,592	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,443	2,704	3,307	3,551
50	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	1,294	1,667	1,994	2,381	2,648	3,211	3,416
80	1,292	1,664	1,990	2,374	2,639	3,195	3,416
90	1,291	1,662	1,987	2,368	2,632	3,183	3,416
100	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	1,286	1,652	1,972	2,345	2,601	3,131	3,340
500	1,283	1,648	1,965	2,334	2,586	3,107	3,310
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Kritische Grenzwerte:
-1.98 und +1.98

T-Test für den Mittelwert



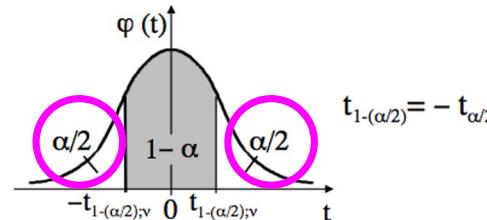
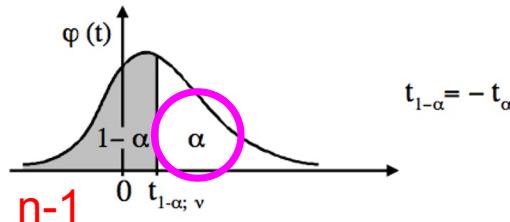
Test

- Wert von $T (= -2.40)$ ist im Ablehnungsbereich
- Stichprobenmittelwert unterscheidet sich von Erwartungswert (9°C) auf dem 5% Signifikanzlevel.
- Test p-Wert von 0.018

p-Wert

- Das kleinste α für das H_0 abgelehnt würde

t-Verteilungs Tabelle



v	Statistische Sicherheit 1- α					
	0,90	0,95	0,975	0,99	0,995	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3
2	1,886	2,920	4,303	6,965	9,925	22,33
3	1,638	2,353	3,182	4,541	5,841	10,21
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,592	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,443	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
200	1,286	1,652	1,972	2,345	2,601	3,131
500	1,283	1,648	1,965	2,334	2,586	3,107
∞	1,282	1,645	1,960	2,326	2,576	3,090

oder entsprechend -1,96

v	Statistische Sicherheit 1- α						
	0,80	0,90	0,95	0,98	0,99	0,998	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	1,638	2,353	3,182	4,541	5,841	10,21	12,92
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,992
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,592	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,443	2,704	3,307	3,551
50	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	1,294	1,667	1,994	2,381	2,648	3,211	3,416
80	1,292	1,664	1,990	2,374	2,639	3,195	3,416
90	1,291	1,662	1,987	2,368	2,632	3,183	3,400
100	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	1,286	1,652	1,972	2,345	2,601	3,131	3,340
500	1,283	1,648	1,965	2,334	2,586	3,107	3,310
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

nur der positive Wert angegeben,
beim 2-seitigen Test ist der negative
kritische Grenzwert entsprechend -1,96

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

STATISTISCHE TESTS TEIL 2

KONFIDENZINTERVALLE

Bahrenberg I: Kap. 5;
Ernste Anh. B;
Ewing I: Kap. 10

T-Test für den Mittelwert

(Vergleich von Stichprobemittel und mit fixem Wert,
z.B. bekanntem Mittel der GG oder Grenzwert)

Hypothesen:

H_0 : Erwartete Sept.-Nov. Mitteltemperatur beträgt 9°C ($x = \mu_0$)

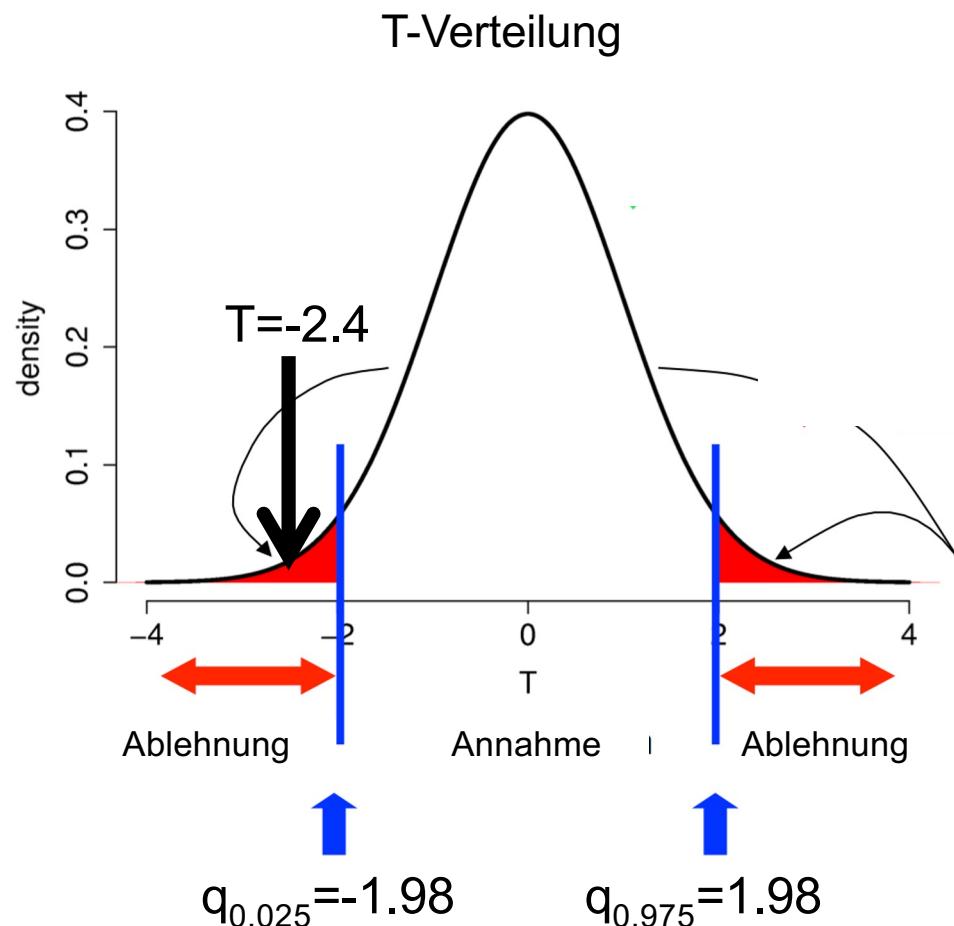
H_A : Die Temperatur weicht signifikant von 9°C ab ($x \neq \mu_0$)

Teststatistik: $T = \frac{\bar{x} - \mu_0}{s_x}$ $x=8.75^\circ\text{C}$, $s_x=0.10^\circ\text{C}$, $n=111$, $T=-2.40$

ähnlich der Standardisierung / Z-Transformation / Standardnormalverteilung

mit dem Standardfehler $s_x = \sqrt{\frac{s^2}{n}}$

T-Test für den Mittelwert



Test

- Wert von $T (= -2.40)$ ist im Ablehnungsbereich
- Stichprobenmittelwert unterscheidet sich von Erwartungswert (9°C) auf dem 5% Signifikanzlevel.
- Test p-Wert von 0.018

p-Wert

- Das kleinste α für das H_0 abgelehnt würde

Testtheorie

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: Beratungen und Überlegungen

- > Angenommen H_0 ist wahr, wie wahrscheinlich ist es zufällig diesen oder einen grösseren Testwert zu erhalten?
- > Der p-Wert ist die numerische Antwort
- > Je kleiner p, desto stärker die Beweise gegen H_0 (oft 0.05 als Grenze)
- > p-Wert sagt **NICHT**: Wie wahrscheinlich ist es, dass die H_0 wahr ist
- > p-Wert sagt: Wie wahrscheinlich die Daten sind, wenn H_0 wahr ist

Testtheorie

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie im Gericht: Das Urteil

Option 1:

wichtig!

- > grosser p-Wert ($p > \alpha$, z.B. mit $\alpha=0.05$ für 95%ige Sicherheit)
- > wir schliessen daraus -> Daten sind konsistent mit H_0

Option 2:

- > kleiner p-Wert ($p < \alpha$)
- > wir schliessen daraus -> es gibt ausreichend Beweise, um H_0 abzulehnen und H_A anzunehmen (statistisch signifikant).

- > Eine Masszahl für die Unsicherheit der Parameterschätzung
- > Wertebereich eines Parameters, für den H_0 nicht abgelehnt wird
- > Eng verknüpft mit dem Signifikanzlevel α !

Beispiel:

$$T = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \Rightarrow \mu \in \left\{ \bar{x} \pm q_{T,0.975} \cdot s_{\bar{x}} \right\}$$



in diesem Beispiel das 0.975 Quantil der T-Verteilung mit $n-1$ Freiheitsgraden.
 n : Stichprobengrösse

Konfidenzintervalle

- > Wie repräsentativ sind unsere Stichprobenergebnisse für die Grundgesamtheit?
- > Wir wollen den Unsicherheitsbereich um den Stichprobenmittelwert bestimmen, sogenanntes Konfidenzintervall:

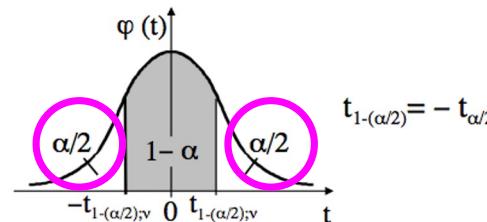
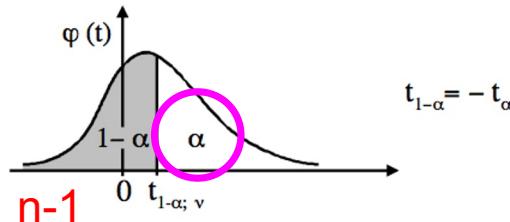
$$\mu_{Obergrenze} = \bar{x} + q_{1-\alpha} \cdot \sqrt{\frac{s^2}{n}}$$

(identisch mit vorheriger Folie,
nur inklusive Berechnung der
Standardfehlers!)

$$\mu_{Untergrenze} = \bar{x} - q_{1-\alpha} \cdot \sqrt{\frac{s^2}{n}}$$

- > q abzulesen aus t-Tabelle normalerweise für den zweiseitigen Test (wenn Ober- und Untergrenze des Konfidenzintervals bestimmt werden), n-1 Freiheitsgrade und α oder aus
- > R: `t.test(x, alternative = ("two.sided" ODER "less" ODER "greater"), conf.level = 0.95, ...)`
- > Excel Funktion T.INV bzw. T.INV.2T, Excel Funktion T.TEST (zur Bestimmung des p-Wertes).

t-Verteilungs Tabelle



FG

v	Statistische Sicherheit $1-\alpha$					
	0,90	0,95	0,975	0,99	0,995	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3
2	1,886	2,920	4,303	6,965	9,925	22,33
3	1,638	2,353	3,182	4,541	5,841	10,21
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,592	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,443	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
200	1,286	1,652	1,972	2,345	2,601	3,131
500	1,283	1,648	1,965	2,334	2,586	3,107
∞	1,282	1,645	1,960	2,326	2,576	3,090

oder entsprechend -1.98

v	Statistische Sicherheit $1-\alpha$						
	0,80	0,90	0,95	0,98	0,99	0,998	0,999
1	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	1,638	2,353	3,182	4,541	5,841	10,21	12,92
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,992
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,592	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,443	2,704	3,307	3,551
50	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	1,294	1,667	1,994	2,381	2,648	3,211	3,436
80	1,292	1,664	1,990	2,374	2,639	3,195	3,416
90	1,291	1,662	1,987	2,368	2,632	3,183	3,400
100	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	1,286	1,652	1,972	2,345	2,601	3,131	3,340
500	1,283	1,648	1,965	2,334	2,586	3,107	3,310
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

nur der positive Wert angegeben,
beim 2-seitigen Test ist der negative
kritische Grenzwert entsprechend -1.98

95% Konfidenzintervall der erwarteten Temperatur aus dem vorherigen Beispiel:

- > alle Werte, für die T kleiner ist als der kritische Wert mit:
 $x = 8.75^\circ\text{C}$, $s_x = 0.10^\circ\text{C}$, $n = 111$, $q_{T,0.95(\text{zweiseitig})} = 1.98$

Konfidenzintervall:

- > Untergrenze: $8.75 - 1.98 \cdot 0.1 = 8.55^\circ\text{C}$
- > Obergrenze: $8.75 + 1.98 \cdot 0.1 = 8.95^\circ\text{C}$

Anmerkung:

- > 9°C liegt nicht im Konfidenzintervall, was im Einvernehmen mit der vorherigen Ablehnung von H_0 ($\mu = 9$) steht

Daumenregel:

- > Konfidenzintervall = Stichprobenergebnis \pm 2 Standardfehler

Was sind die Voraussetzungen, damit der t-Test angewendet werden darf?

Was sind die Voraussetzungen, damit der t-Test angewendet werden darf?

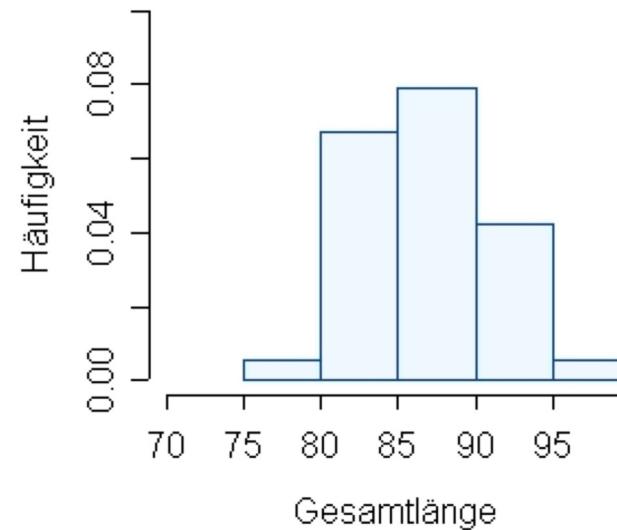
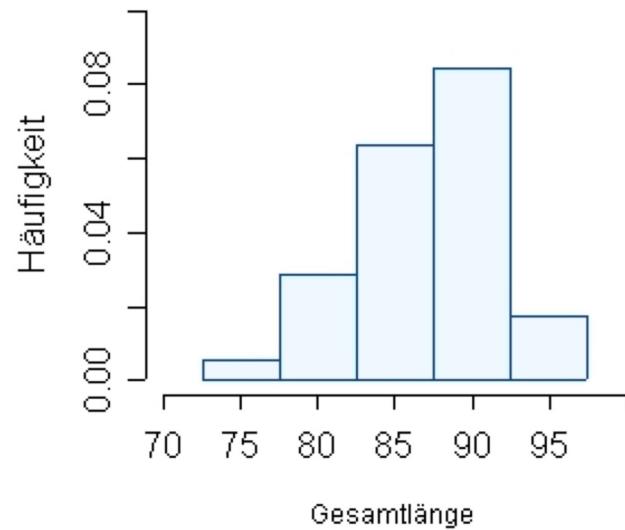
Für den T-Test alles, was Voraussetzung für das arithmetische Mittel ist, d.h. metrische Daten, keine Ausreisser in den Daten, und symmetrische Verteilung

Überprüfung auf Normalverteilung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



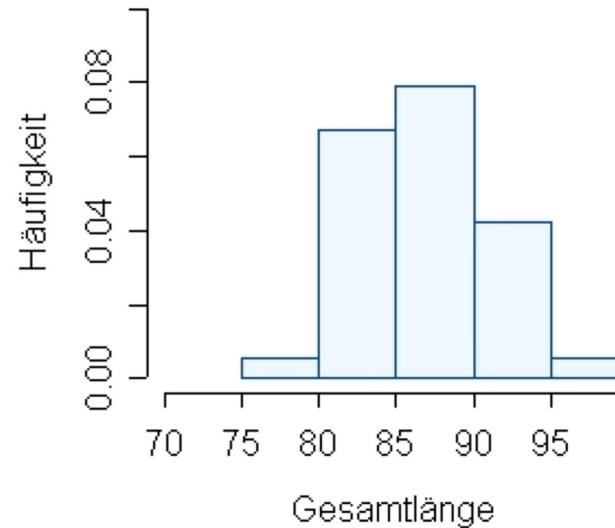
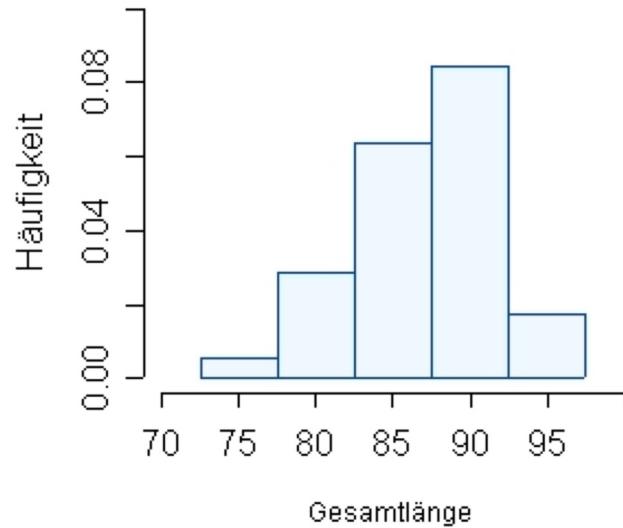
Sind diese Daten normalverteilt?

Überprüfung auf Normalverteilung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Gleiche Daten, aber
visuell unterschiedliche Histogramme

Hypothesen aufstellen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Wie lautet H_0 und H_A beim Test, ob die Daten normalverteilt sind?

Hypothesen aufstellen



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

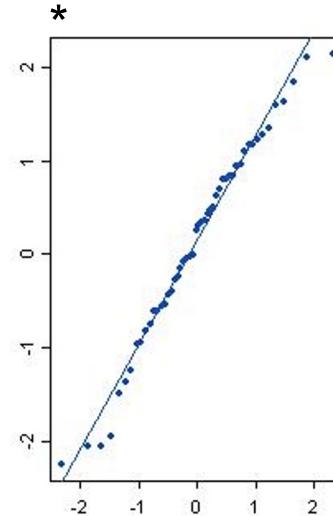
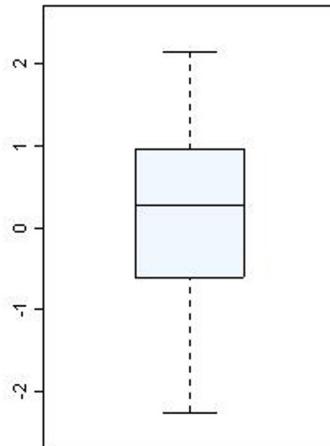
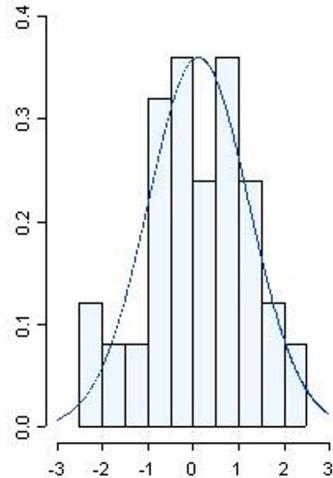
Warum ist Normalverteilung Nullhypothese, wenn man doch gerade beweisen will, dass Daten normalverteilt sind?

Ein Test sagt ja nur gleich oder ungleich bzw. es gibt (nicht) ausreichend Beweise gegen H_0 . Wenn wir nun testen, ob H_0 nicht z.B. gleichverteilt,... ist. Wenn H_0 verworfen werden kann, heisst das aber noch nicht, dass die Daten normalverteilt sind, könnten auch irgendwie anders verteilt sein.

Der einzige Weg, um herauszubekommen, ob die Daten normalverteilt sind, ist also zu prüfen ob man die Hypothese der Normalverteilung widerlegen kann!

Daten sind gleich der Normalverteilung
Das = steht also wieder in H_0

Überprüfung auf Normalverteilung



Kenngroße	Messreihe A	Normalverteilung
(Mittelwert - Median) / s	-0,14	0
IQR / s	1,36	1,34
# 1s Intervall / n	72 %	68 %
# 2s Intervall / n	98 %	95 %
# 3s Intervall / n	100 %	99,73 %

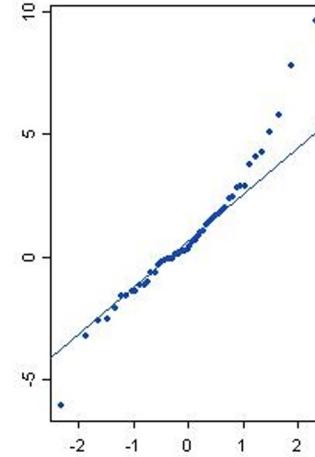
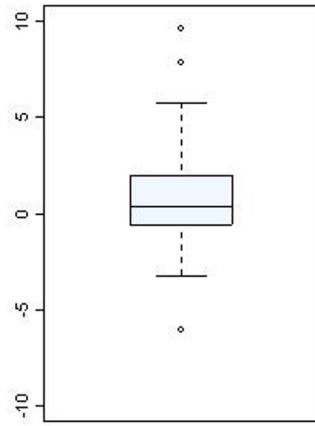
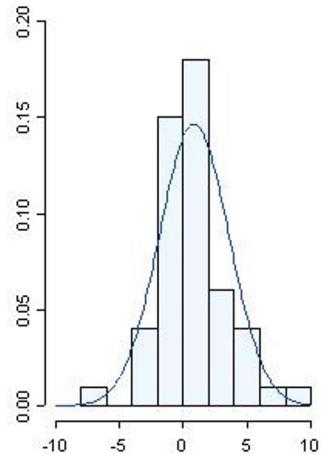
Test	p-Wert
Shapiro-Wilks	0,4544
Anderson-Darling	0,6796
Cramér-von Mises	0,7521

Die p-Werte der Tests auf Normalverteilung sind alle deutlich größer als 0.05, somit wird die Nullhypothese der Normalverteilung beibehalten.

Beim Test auf Normalverteilung wird immer in der Nullhypothese angenommen, dass die Messreihe aus einer Normalverteilung stammt.

*QQ-Plot: Quantile zweier statistischer Variablen gegeneinander abgetragen werden, um ihre Verteilungen zu vergleichen

Überprüfung auf Normalverteilung



Kenngröße	Messreihe B	Normalverteilung
(Mittelwert - Median) / s	0,18	0
IQR / s	0,94	1,34
# 1s Intervall / n	76 %	68 %
# 2s Intervall / n	94 %	95 %
# 3s Intervall / n	98 %	99,73 %

Test	p-Wert
Shapiro-Wilks	0,0303
Anderson-Darling	0,0285
Cramér-von Mises	0,0331

Die p-Werte der Tests auf Normalverteilung lehnen in jedem Fall die Nullhypothese zum Niveau $\alpha=0,05=5\%$ ab

p-Werte sind deutlich kleiner als 0,05, d. h. die Nullhypothese "Messreihe ist normalverteilt" wird verworfen

Überprüfung auf Normalverteilung in R

U^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Test	Vorteile	Nachteile
Chi-Quadrat-Test	- geeignet für beliebig skalierte Variablen	- Gruppierung der Beobachtungen notwendig - ungeeignet für kleine Stichproben - quadratische Testgrösse, d.h. sensibel auf Ausreisser
Kolmogorov-Smirnov Test	- geeignet für kleine Stichproben - wie Chi-Quadrat auch zum Vergleich anderer Verteilungen geeignet - nicht-parametrischer Test, d.h. NICHT sensibel auf Ausreisser	- geringe Teststärke im Vergleich zu den folgenden Tests
Cramér-von-Mises-Test	- höhere Güte als KS-Test	- quadratische Testgrösse
Lilliefors-Test	- bessere Trennschärfe als KS-Test - nicht-parametrischer Test, d.h. NICHT sensibel auf Ausreisser	- nur zum Test auf Normalverteilung
Anderson-Darling-Test	- sehr hohe Güte bei Test auf Normalverteilung	- keine kategorialen Daten - quadratische Testgrösse, d.h. sensibel auf Ausreisser
Shapiro-Wilk-Test	- Test mit höchster Güte	- ausschließlich Test auf Normalverteilung - manuell schlecht durchführbar - sensibel auf Ausreisser und viele identische Werte

Überprüfung auf Normalverteilung in R

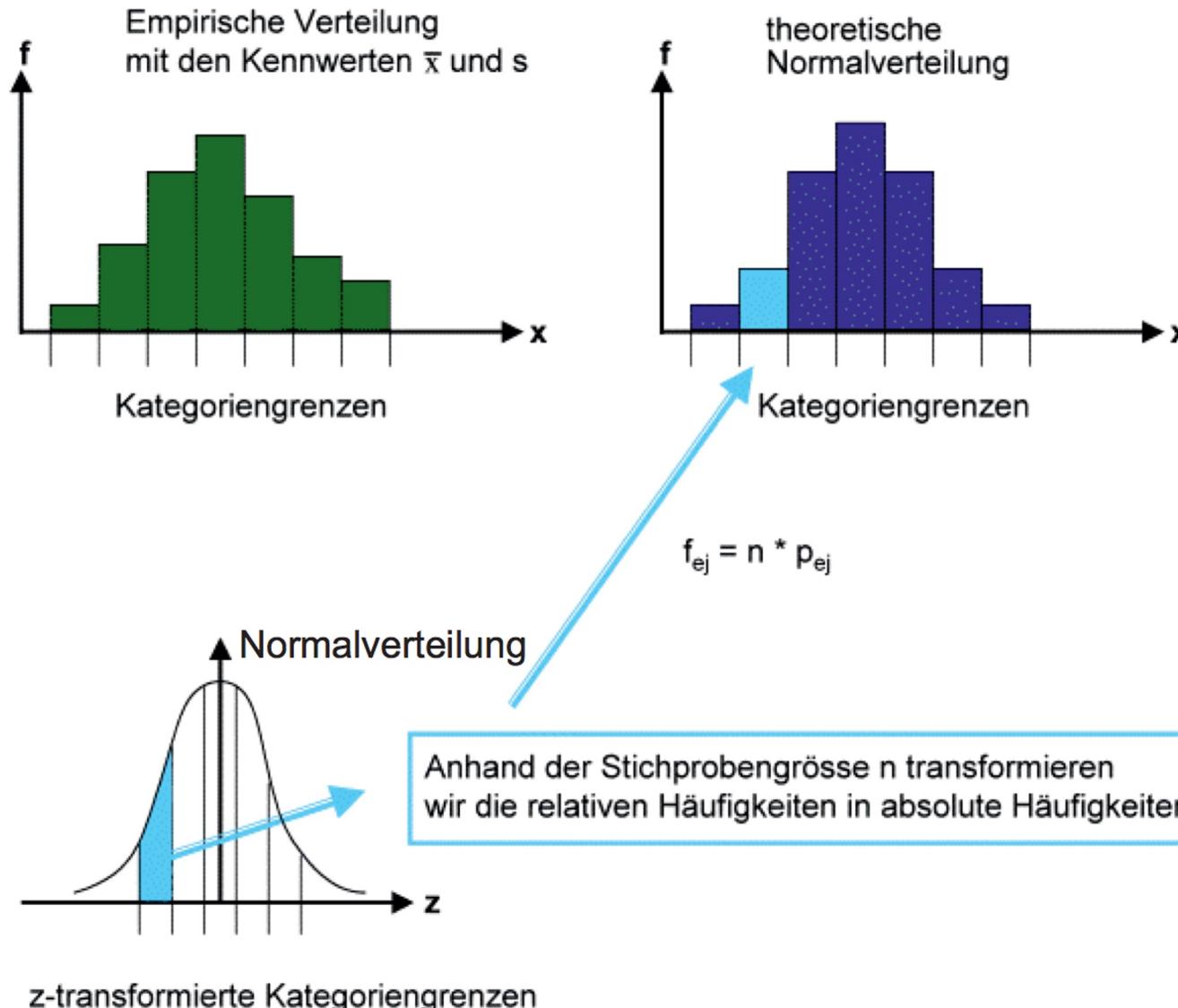


b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

```
> data <- rnorm(500) # 500 normalverteilte Zufallszahlen
> data2 <- data^2      # quadrierte Zufallszahlen
> hist(data)           # Histogramm
> plot(density(data)) # Dichtefunktion
> boxplot(data)        # Boxplot
> qqnorm(data)         # Quantil-Quantil Plot
> qqline(data, col = 2) # Erwartung bei Normalverteilung
> shapiro.test(data)  # Shapiro Normalverteilungstest
```

χ^2 Test auf Normalverteilung



Allgemeiner Test für Verteilungen

χ^2 -Verteilungstest

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

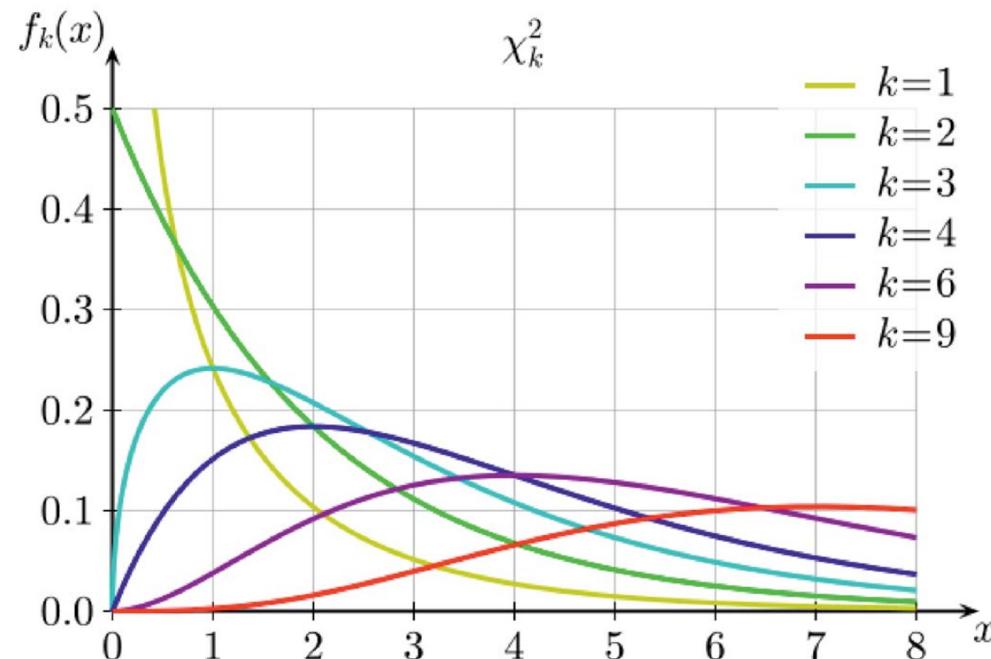
- > Summe der normierten quadrierten Abweichungen:

$$\chi^2 = \sum_{i=1}^k \frac{(N_i - n_i)^2}{n_i}$$

- > N beobachtete Häufigkeit
- > n erwartete Häufigkeit
- > i Klassen
- > k Anzahl an Klassen

χ^2 -Verteilung

- > Stetige Wahrscheinlichkeitsverteilung mit der Anzahl Freiheitsgrade k als einzigem Parameter.
- > Verteilung der Summe der Quadrate von k unabhängigen und standardnormalverteilten Zufallsvariablen.



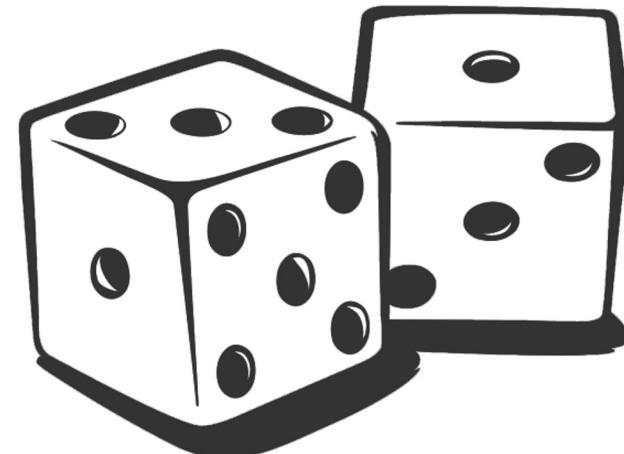
Würfelergebnisse

u^b

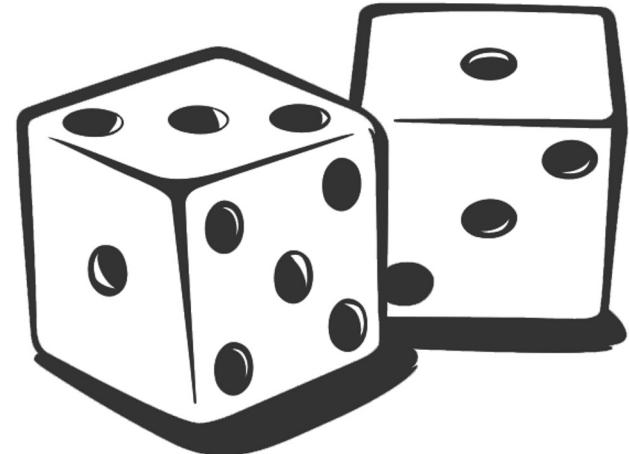
b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Was sind die Null-, und Alternativhypothese unseres Würfelexperiments?



Würfelergebnisse



- > Was sind die Null-, und Alternativhypothese unseres Würfelexperiments?
- > H_0 : Die Verteilung der gewürfelten Zahlen entspricht den theoretischen Wahrscheinlichkeiten
- > H_A : Die Verteilung der gewürfelten Zahlen entspricht NICHT den theoretischen Wahrscheinlichkeiten

Richtigen Test und statistische Verfahren finden

u^b

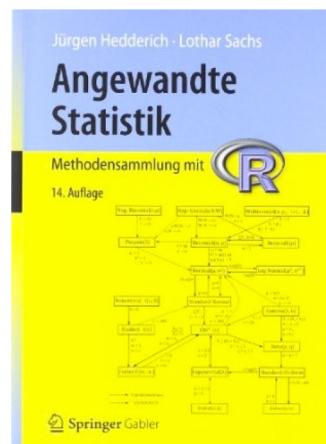
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Chatbot?

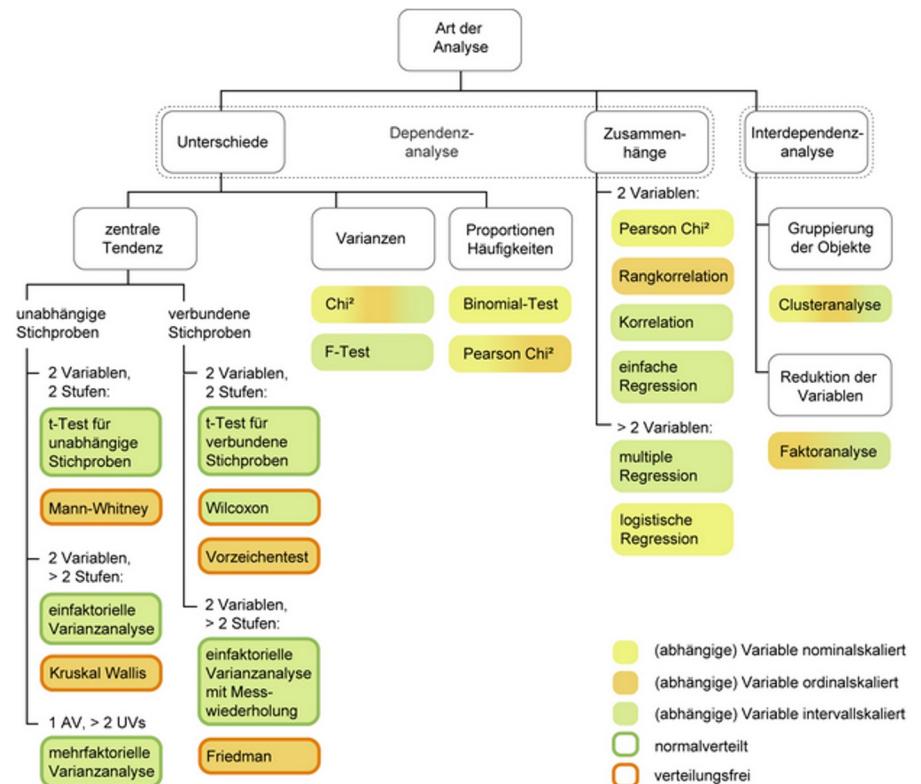
Bücher:

- > Ernste S. 21 ff.
- > Jürgen Hedderich, Lothar Sachs,
2012: Angewandte Statistik:
Methodensammlung mit R. Springer

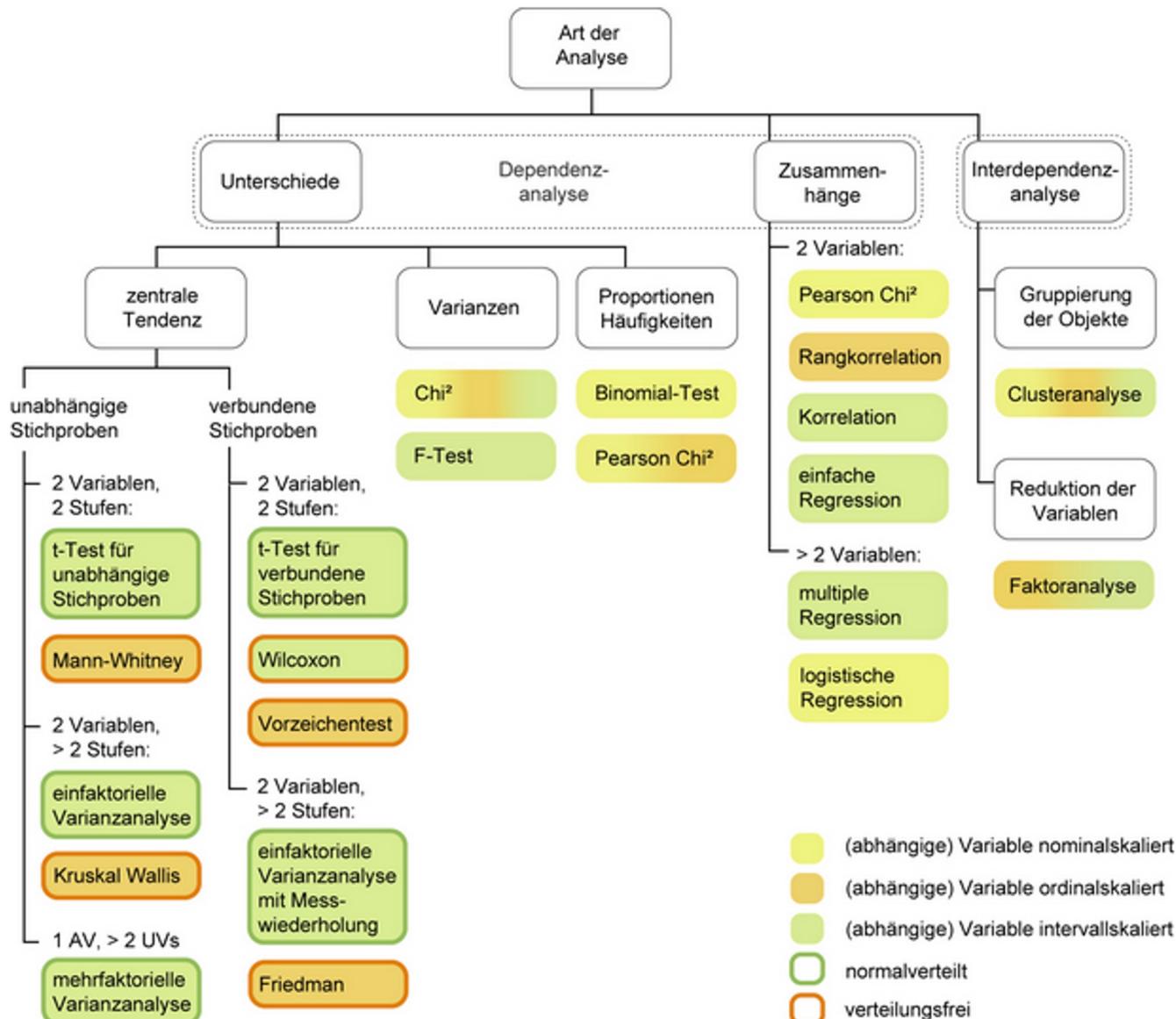


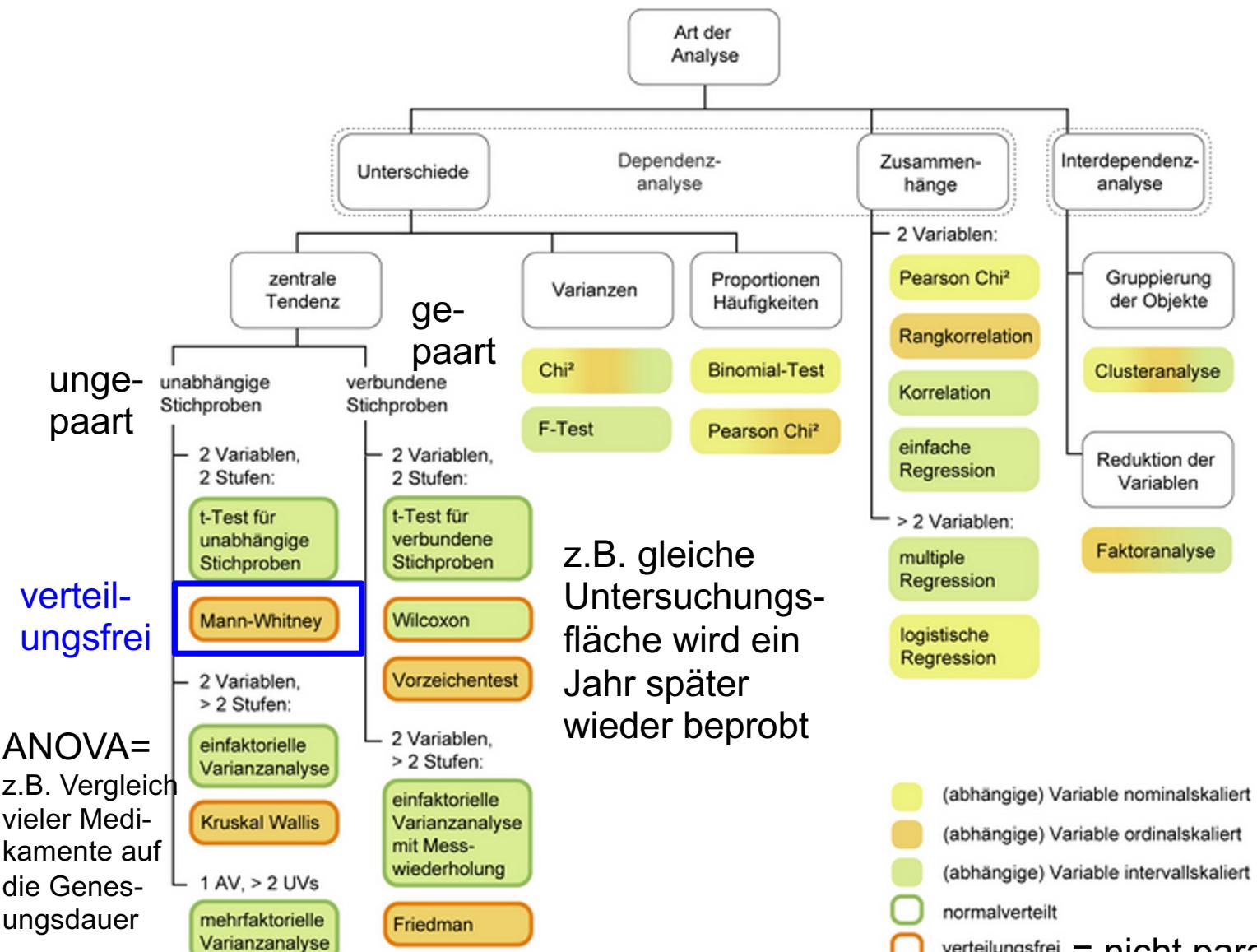
Entscheidungsbäume

- > <http://www.methodenberatung.uzh.ch/index.html>
- > <http://etools.fernuni.ch/entscheidungsbaum/>



Angenommen ihr wollt Winter- und Sommermittelwerte von nicht-normalverteilten Niederschlagssummen vergleichen.
Welchen Test müsst ihr nehmen?





AV=abhängige Variable
UV=unabhängige Variable

Verfahren kann oft auch für metrische Daten verwendet werden, die nicht normalverteilt sind

Take-home Message

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Standardvorgehen bei statistischen Tests:

1. Hypothesen aufstellen
2. Schwellenwert festlegen (z.B. $\alpha = 0.05$, d.h. bei jedem 20. Test eine falsche Entscheidung)
3. Ziehung der Stichprobe
- 4.1. Traditionelle und manuelle Zusammenfassung der Daten in einem Wert (einer Teststatistik) und Vergleich ob im Annahme- oder Ablehnungsbereich mithilfe einer Tabelle **oder**
- 4.2. p(robability)-Wert durch Software ermitteln. Wenn der p-Wert < als z.B. 0.05 ist bedeutet dies, dass es eine sehr kleine (<5%) Wahrscheinlichkeit gibt, die Stichprobe zu erhalten, wenn die Nullhypothese wahr ist **oder**
- 4.3. Konfidenzintervalle bestimmen. Diese beschreiben (z.B. für den Mittelwert) einer Stichprobe, den Bereich in dem z.B. mit 95% Sicherheit der entsprechende Wert (hier Mittelwert) Grundgesamtheit liegt.

Schliessende Statistik beruht meist auf Wahrscheinlichkeiten und liefert uns KEINE 100% sicheren Ergebnisse!

Enger mathematischer Zusammenhang zwischen Tests und Konfidenzintervallen

Alle Tests funktionieren nach ähnlichem Muster: Wenn ihr einen Verstanden habt, könnt ihr jeden beliebigen Test in eurer eigenen Arbeit anwenden. Webseiten und Bücher helfen den richtigen Test zu finden, enthalten Informationen über Freiheitsgrade, etc.

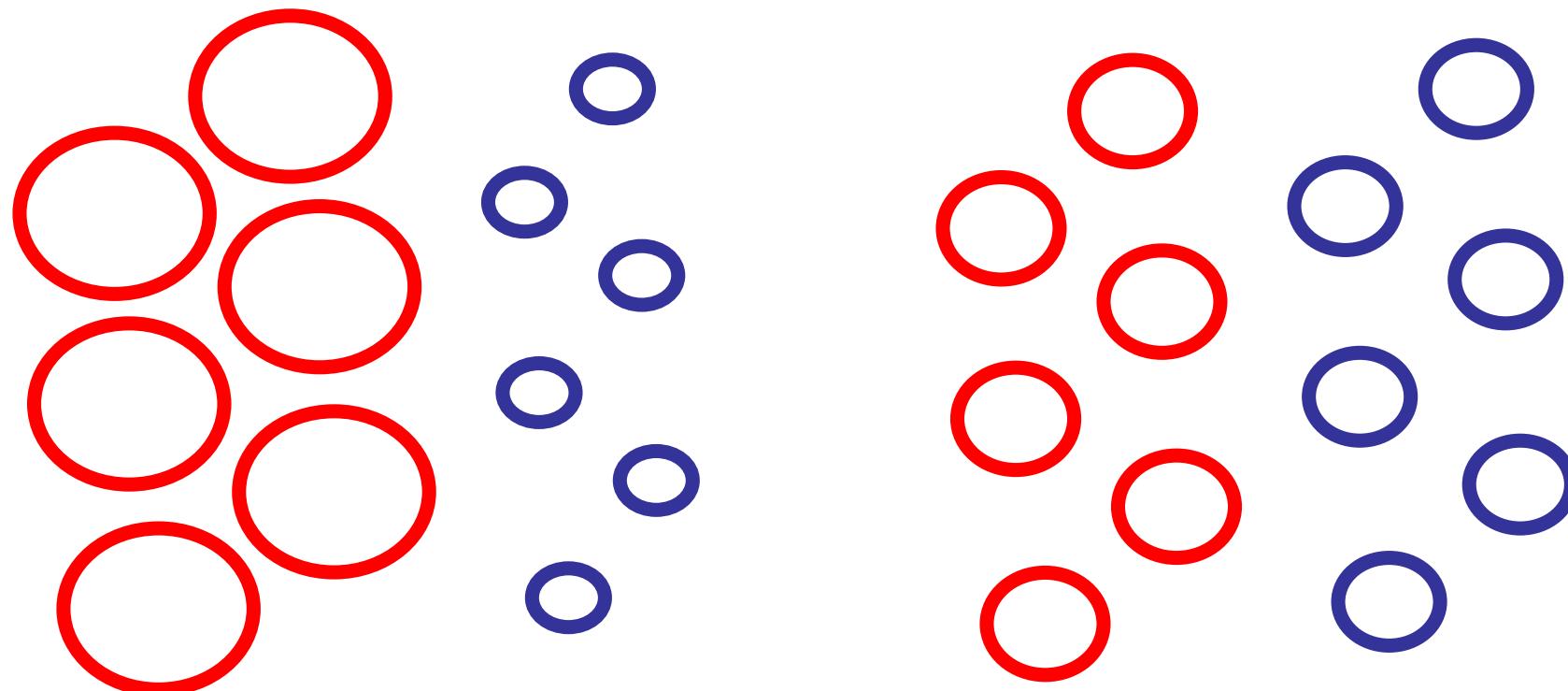
Statistisch signifikant

- > Eine Studie mit 10000 Personen zeigt, dass ein neues Medikament die Dauer einer Grippe statistisch signifikant ($\alpha < 0.05$) um eine Stunde verkürzt.
- > Was haltet ihr von dem Medikament?

ACHTUNG:

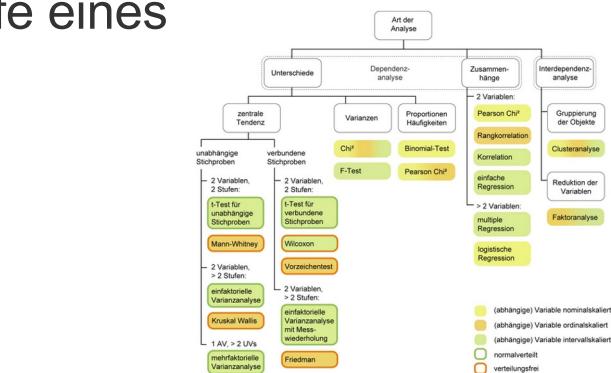
Bei einer grossen Stichprobe kann auch ein kleiner Unterschied statistisch signifikant sein

- > Ob die Grösse des Unterschieds eine Bedeutung hat, müsst ihr interpretieren:
- > **Grösse des Effekts**



Beispiel Prüfungsfrage

- > Was versteht man in der Statistik unter dem “Standardfehler”?
 - Abweichungen der Stichprobenwerte vom Stichprobenmittelwert
 - Abweichungen der Stichprobenmittelwerte vom Mittelwert der Grundgesamtheit
 - Die Varianz der Mittelwerte von mehreren Stichproben der gleichen Grundgesamtheit
 - Die Varianz der Stichprobenwerte von Stichproben aus verschiedenen Grundgesamtheiten
- > Interpretiert die R Ausgabe mit dem Ergebnis eines t-Tests (siehe Übung aus Teil 1)
- > Welcher statistische Test eignet, um die Mittelwerte nicht normalverteilter, gepaarter Stichproben zu vergleichen (mit Hilfe eines Entscheidungsdiagramms)?



u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

KORRELATION

Ernste Kap. 2, Anh. A5.1-A5.2;
Ewing I: Kap. 9

Statistische Datenanalyse

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen
Fallen der Statistik			
weiterführende Methoden	Daten zusammenfassen	Extremwertstatistik	
	Hauptkomponenten-analyse	Zeitreihenanal. etc.	Clusteranalyse

Beispiele von Beziehungen

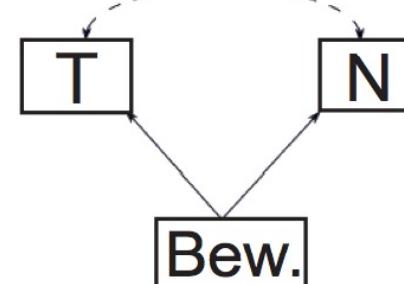
Einfache Kausalbeziehung

Ursache

Folge



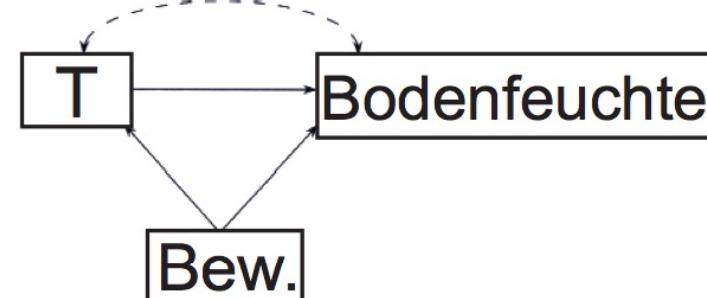
Scheinrelation



Bodenfeuchte

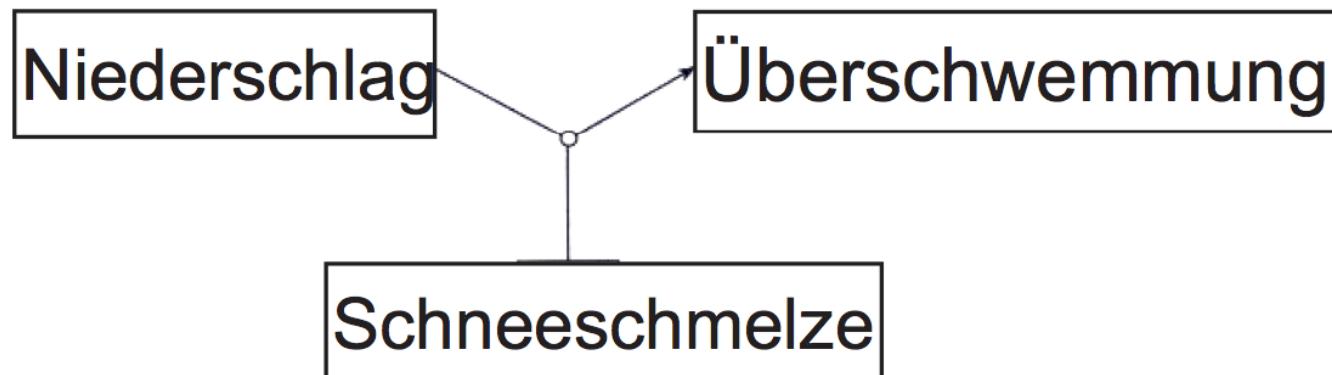
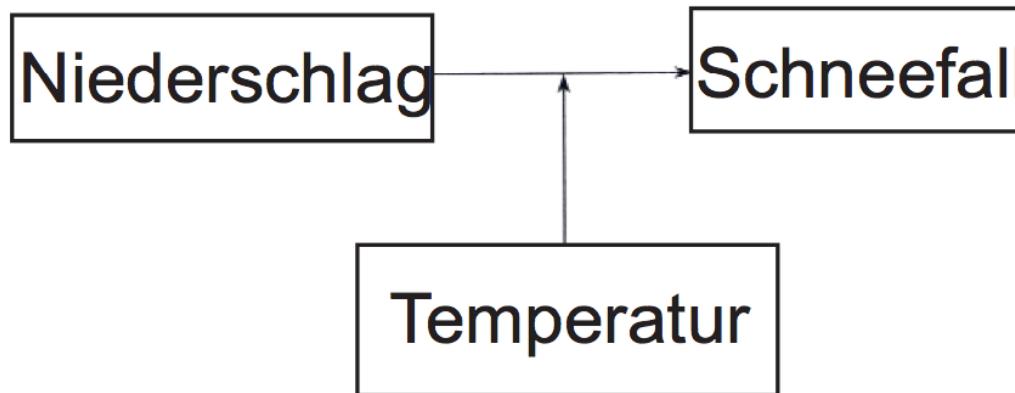
Niederschlag

Kovariation



Partielle Scheinrelation

Beispiel für konditionale Variablen



Korrelationsanalyse



Abbildung 1.3: Pfaddiagramm einer einfachen kausalen Beziehung



Abbildung 1.4: Pfaddiagramm einer Kovariation

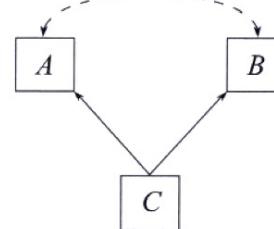


Abbildung 1.5: Pfaddiagramm einer Scheinrelation

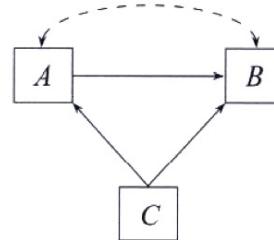


Abbildung 1.6: Pfaddiagramm einer partiellen Scheinrelation

- > Masszahl für die Stärke des **linearen** Zusammenhangs **metrischer** Variablen
- > sagt nichts über die Kausalität und Richtung des Zusammenhangs aus

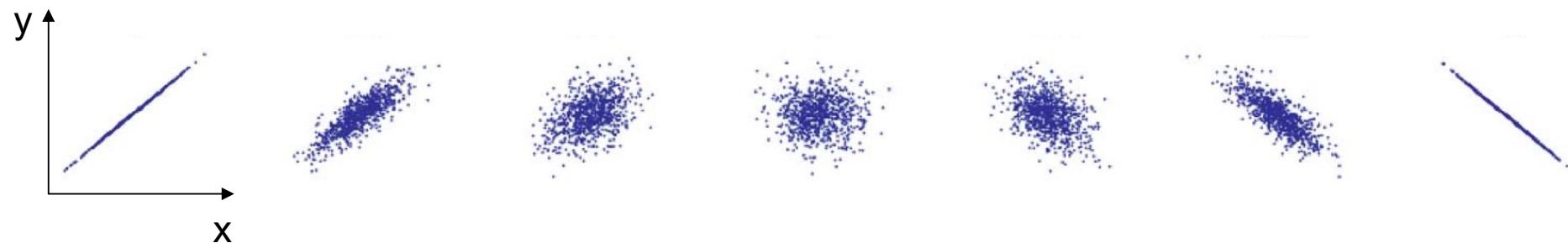
Streudiagramme oder Scatterplots

u^b

Zusammenhänge zwischen zwei Variablen

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Korrelationskoeffizient?

Pearson-Korrelation

Wie messen wir den Zusammenhang zweier Variablen?

1. Idee: Das Produkt der Anomalien

$$X_i^d = X_i - \mu_x ; Y_i^d = Y_i - \mu_y$$

$$\sum_{i=1}^N (X_i - \mu_x)(Y_i - \mu_y) = \sum_{i=1}^N X_i^d Y_i^d$$

Ist von Stichprobengrösse (n)
abhängig

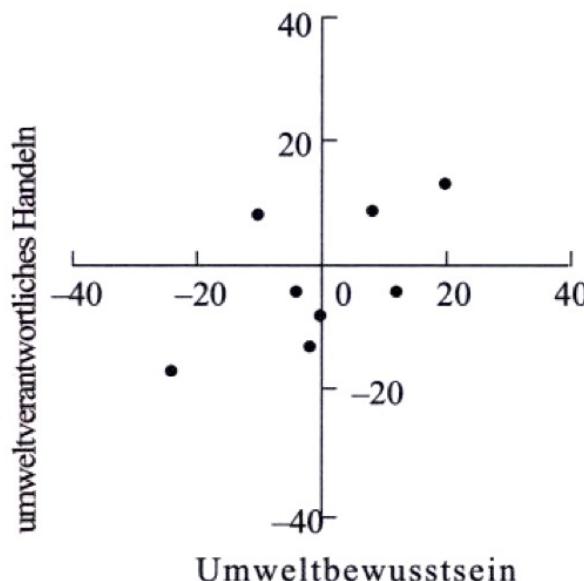


Abbildung 2.9: Transformation der ursprünglichen Werte zu Abweichungen vom Mittelwert

Pearson-Korrelation

2. Division durch Stichprobengrösse

Kovarianz zwischen Y und X

$$\sigma_{xy} = \frac{\sum_{i=1}^N (X_i - \mu_x)(Y_i - \mu_y)}{N}$$

Ist von den Einheiten abhängig

Pearson-Korrelation

3. Standardisierung

mit Standardabweichungen von X und Y

$$\rho_{xy} = \frac{\sum_{i=1}^N \frac{(X_i - \mu_x)}{\sigma_x} \frac{(Y_i - \mu_y)}{\sigma_y}}{N}$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_x)^2}{N}} \text{ und}$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^N (Y_i - \mu_y)^2}{N}}$$

Korrelationskoeffizient (Pearson)

$$\rho_{xy} = \frac{\sum_{i=1}^N (X_i - \mu_x) (Y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (X_i - \mu_x)^2 \sum_{i=1}^N (Y_i - \mu_y)^2}}$$

Prüfung!

Pearson-Korrelation

- > Der Korrelationskoeffizient drückt das Verhältnis der beobachteten Kovarianz zur maximalen Kovarianz aus bzw.
- > Der Korrelationskoeffizient ist die bezüglich der Standardabweichung von X und Y normierte Kovarianz.

Korrelationskoeffizient (Pearson)

$$\rho_{xy} = \frac{\sum_{i=1}^N (X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (X_i - \mu_x)^2 \sum_{i=1}^N (Y_i - \mu_y)^2}}$$

- > ρ (rho) ist der Standardbuchstabe für den Korrelationskoeffizienten der Grundgesamtheit
- > r für den Korrelationskoeffizienten der Stichprobe

Ausreisser

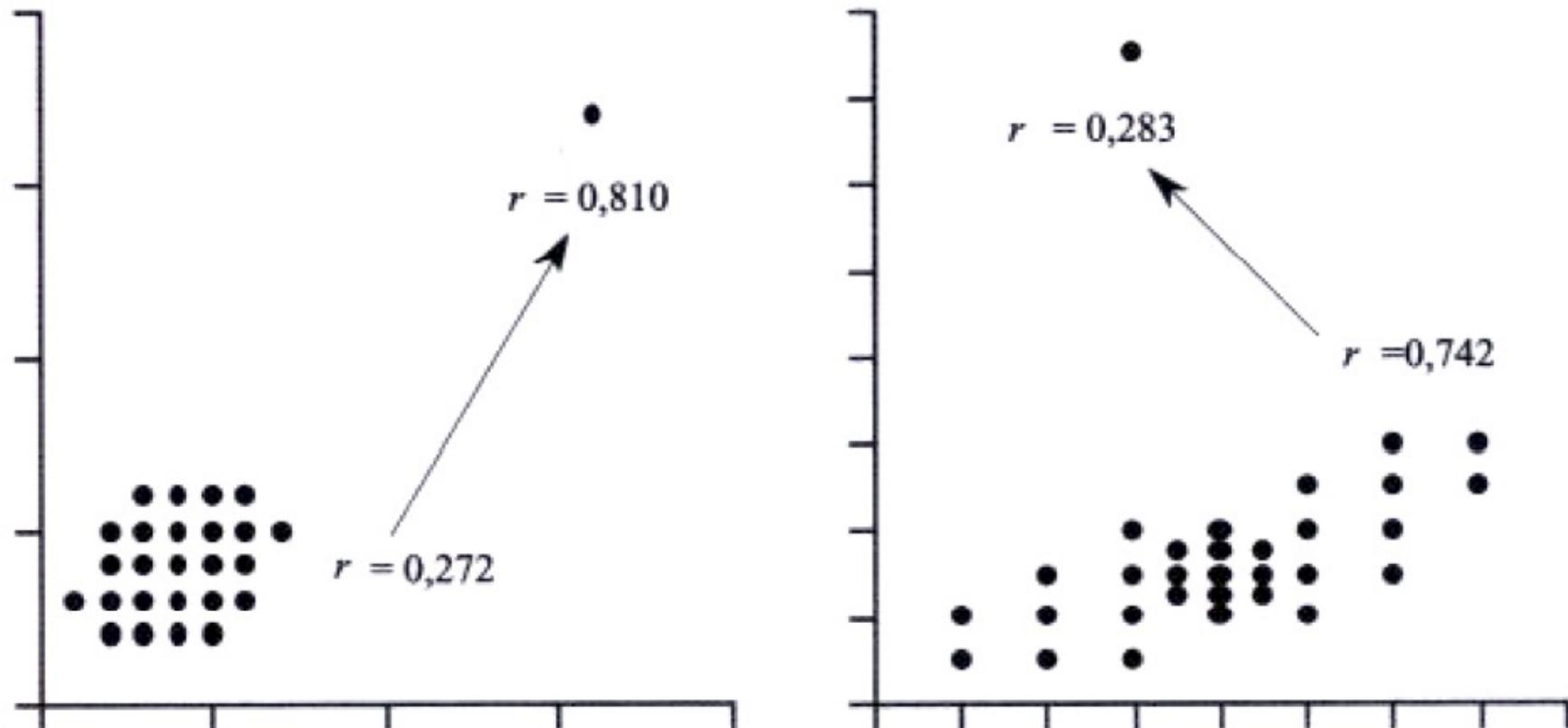


Abbildung 2.11: Einfluss eines Ausreissers auf den Korrelationskoeffizienten

Spearman Rangkorrelation

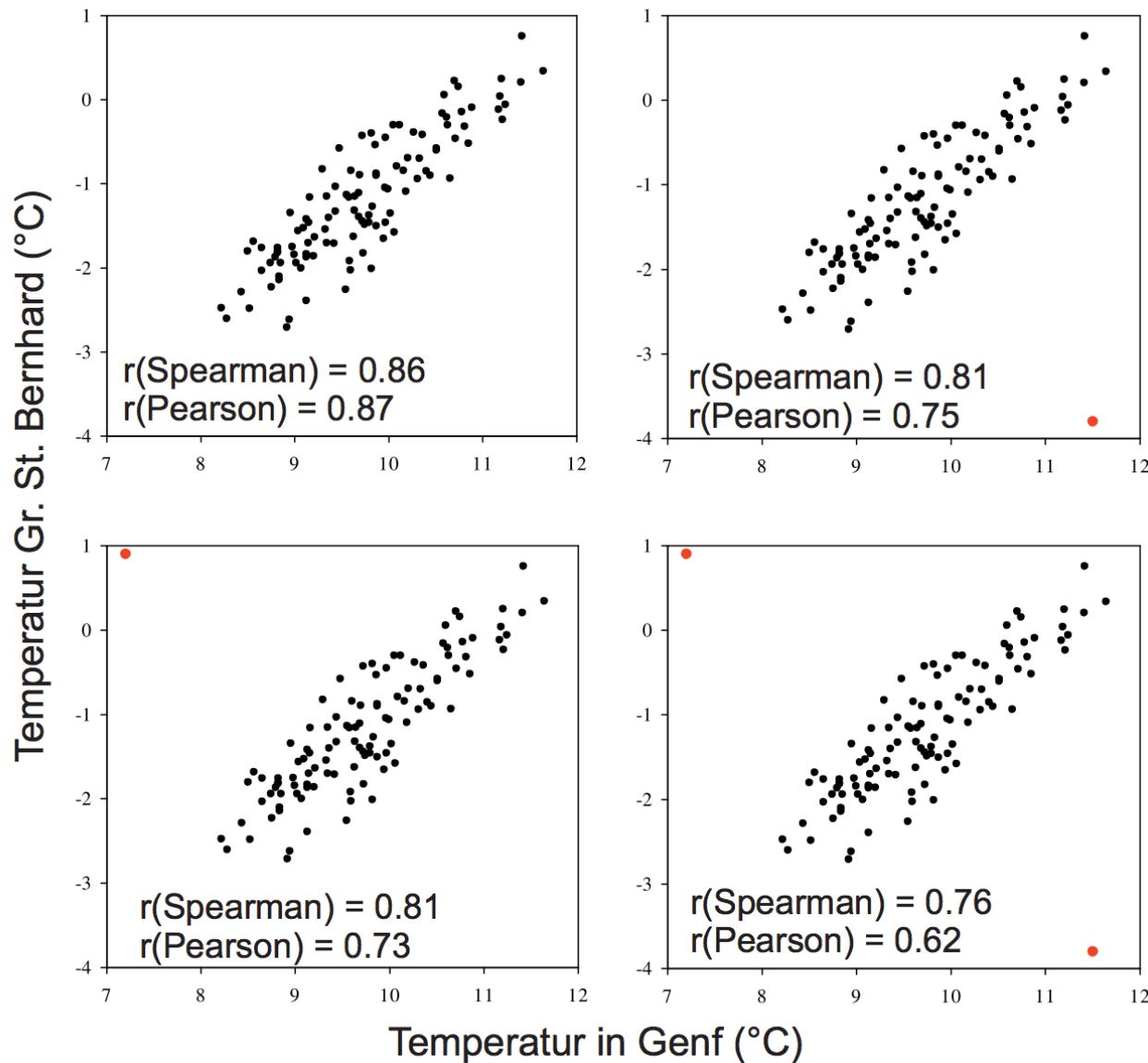
- > Für ordinal skalierte Variablen kann der Rang eines Objekts (Beobachtung) in zwei Variablen verwendet werden:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n (r_i - s_i)^2}{n^3 - n}$$

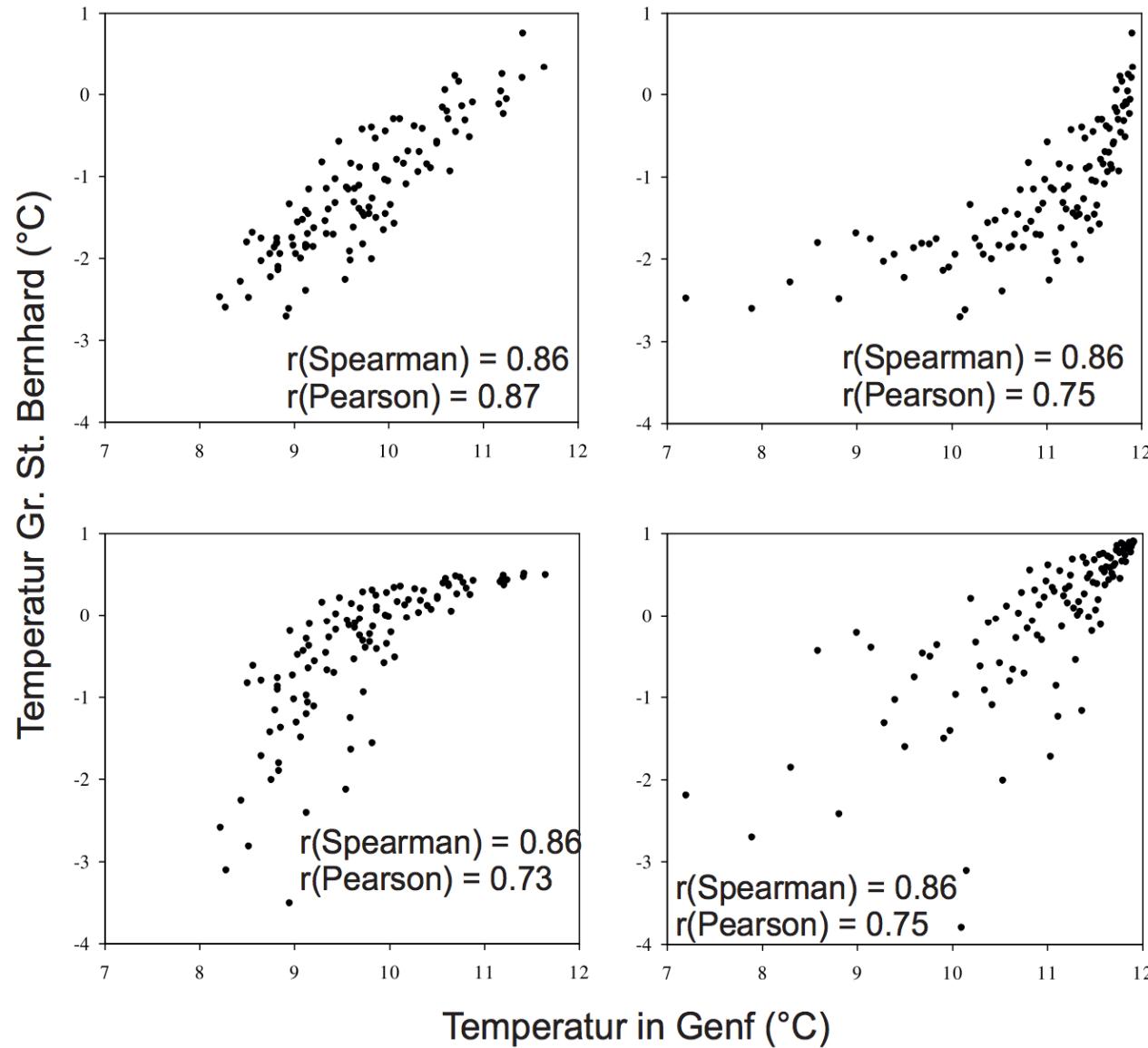
(vereinfachte Formel, wenn jeder Rang nur einmal vergeben ist)

- > r_i = Rang von Objekt i in der Variablen R, s_i = Rang von Objekt i in der Variablen S.
- > Der Spearman Rangkorrelationskoeffizient wird sehr oft auch für metrische Variablen verwendet, da er robust ist gegenüber Ausreisern. Im Zweifelsfall ist der Spearman-Koeffizient dem Pearson-Koeffizient vorzuziehen.

Spearman Rangkorrelation: Ausreisser



Spearman Rangkorrelation: Nicht-Linearität



Transformation der Achsen

- > Pearson-Korrelation: Misst Stärke des **linearen** Zusammenhangs, ist nicht sensitiv gegenüber linearen Transformationen (z.B. $y = a + bx$). Bei klarer (oder theoretisch begründeter) Nicht-linearität ist eine Transformation möglich.
- > Spearman: Misst Stärke eines **monotonen** Zusammenhangs, ist nicht sensitiv gegenüber monotonen Transformationen (z.B. $y = \ln(x)$)

Wann machen Transformationen Sinn?

- > Monotone Transformation nicht unbedingt (einfach Spearman statt Pearson verwenden)
- > Nicht-monotone Transformationen (z.B. $y = x^2$) können Sinn machen.
- > Alle Transformationen machen Sinn bei der Regression

Voraussetzungen für die Pearson-Korrelation

- > metrische Daten
- > beide Variablen annähernd normalverteilt
- > linearer Zusammenhang zwischen x und y
- > statistischer Zusammenhang zwischen x und y nur, wenn der ermittelte Korrelationskoeffizient signifikant von Null abweicht (t-Test)

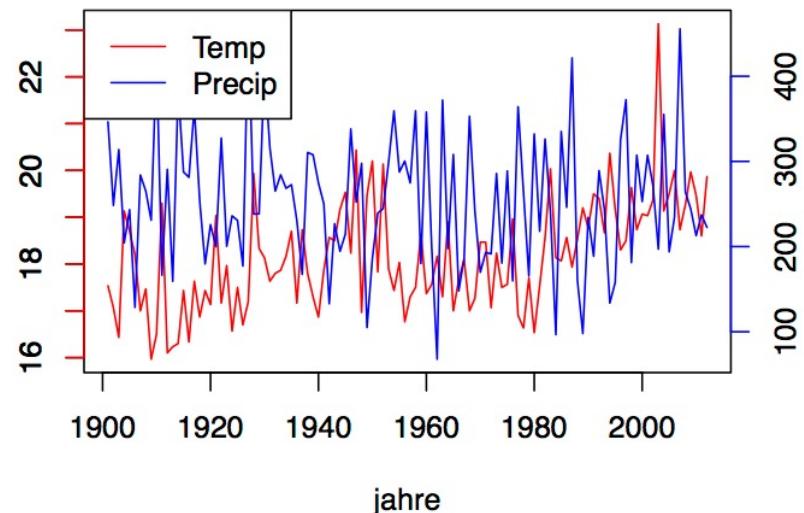
Rückschlüsse auf die Grundgesamtheit

Hypothesen:

- > H_0 : x und y sind unkorreliert $\rho_{xy} = 0$
- > H_A : x und y sind korreliert $\rho_{xy} \neq 0$
- > Annahme: Verteilung von x und y in der Grundgesamtheit ist eine bivariate Normalverteilung
- > Teststatistik: t-Test mit $n-2$ Freiheitsgraden

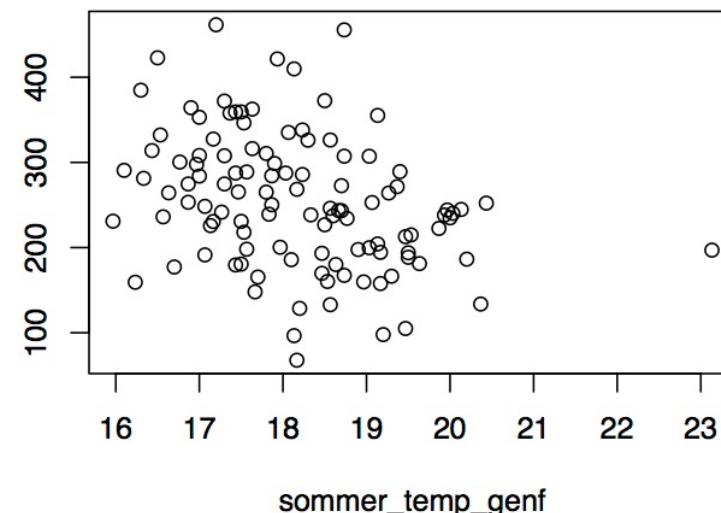
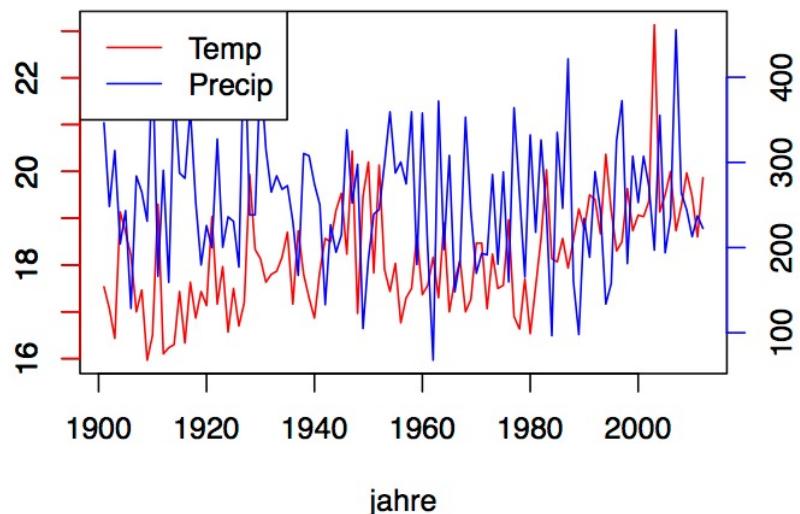
$$t = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

Korrelation von Zeitreihen



Korrelationskoeffizient?

Korrelation von Zeitreihen



```
cor.test(sommer_temp_genf, sommer_precip_genf)
```

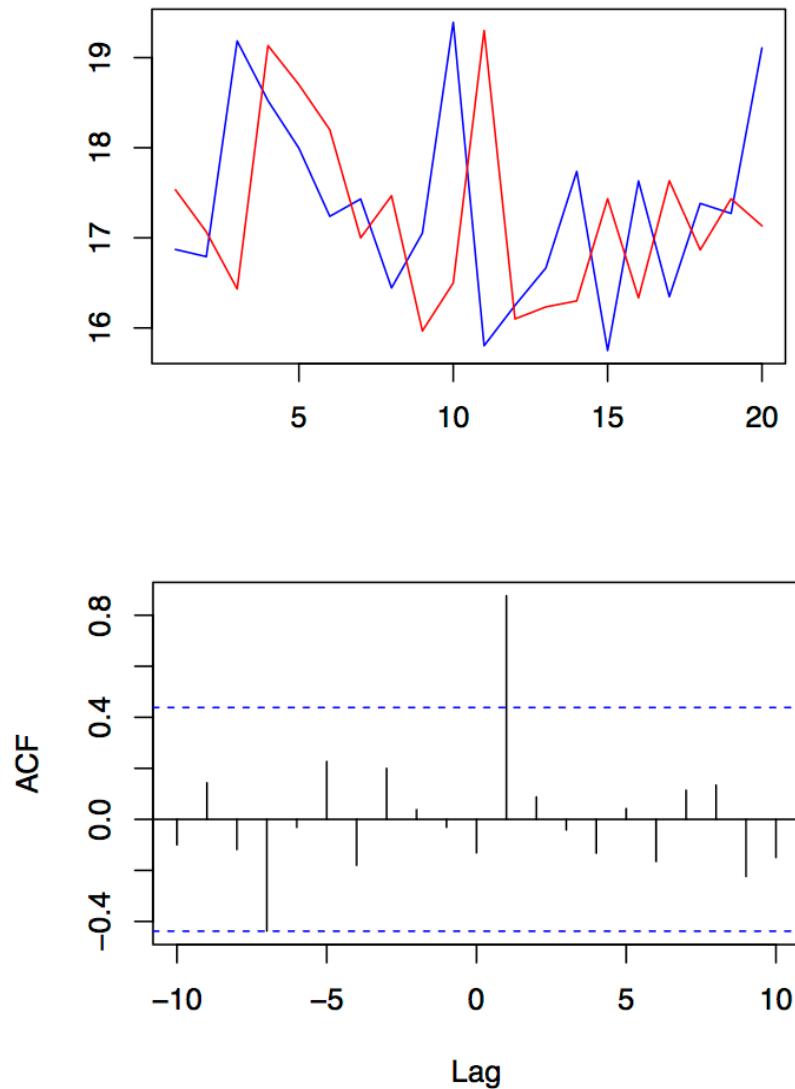
```
Pearson's product-moment correlation
data: sommer_temp_genf and sommer_precip_genf
t = -3.7533, df = 110, p-value = 0.0002805
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: -0.4917509 -0.1614801
sample estimates: cor -0.33694
```

Autokorrelation in Zeit und Raum

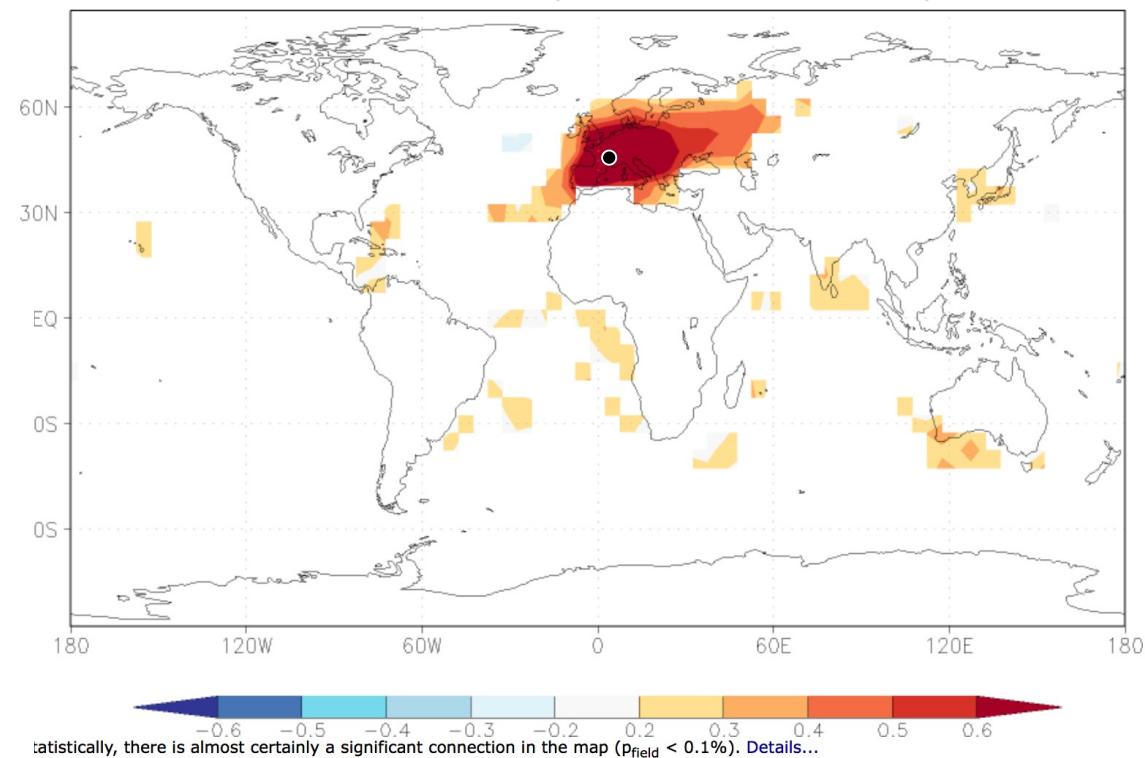
u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



corr Jan GENEVE-COINTR ghcn_v3_mean_temperature with Jan HadCRUT4.5 SST/T2m anom 1850:2016 p<10% (eps, pdf)
corr Jan GENEVE-COINTR ghcn_v3_mean_temperature
with Jan HadCRUT4.5 SST/T2m anom 1850:2016 p<10%



Varianz-Kovarianzmatrix

- > Symmetrische Matrix mit den Varianzen in der Diagonalen und den Kovarianzen abseits der Diagonalen

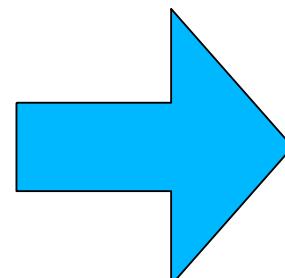
$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix};$$

R Beispiel:

```
A <- matrix(c(2,3,1,-1,1,1,0,4,2,-1,0,0),nrow=4,ncol=3, byrow=T)
cor(A) # oder ,method="pearson", Pearson ist Standardeinstellung
cor(A,method="spearman")
```

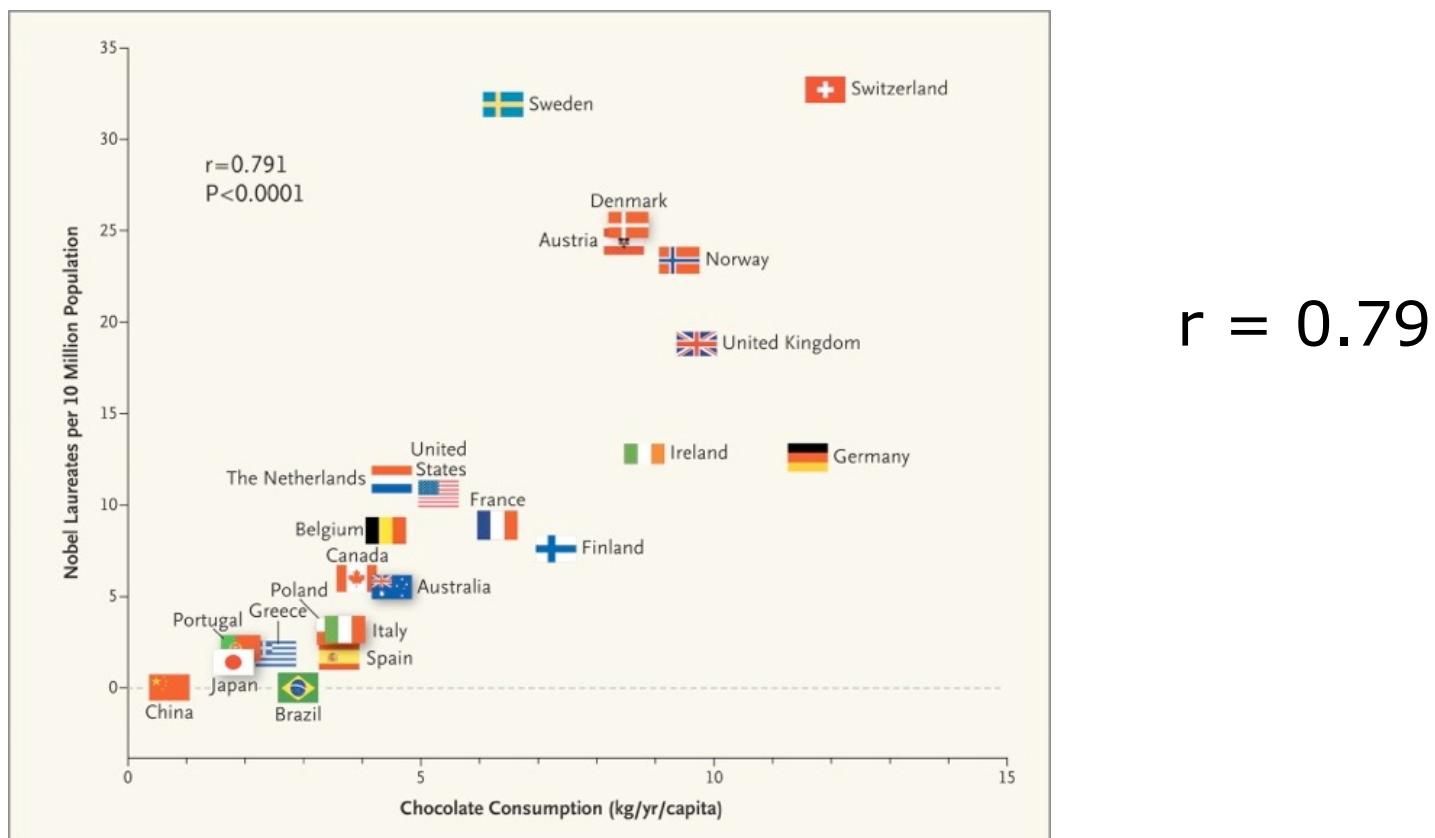
Korrelation

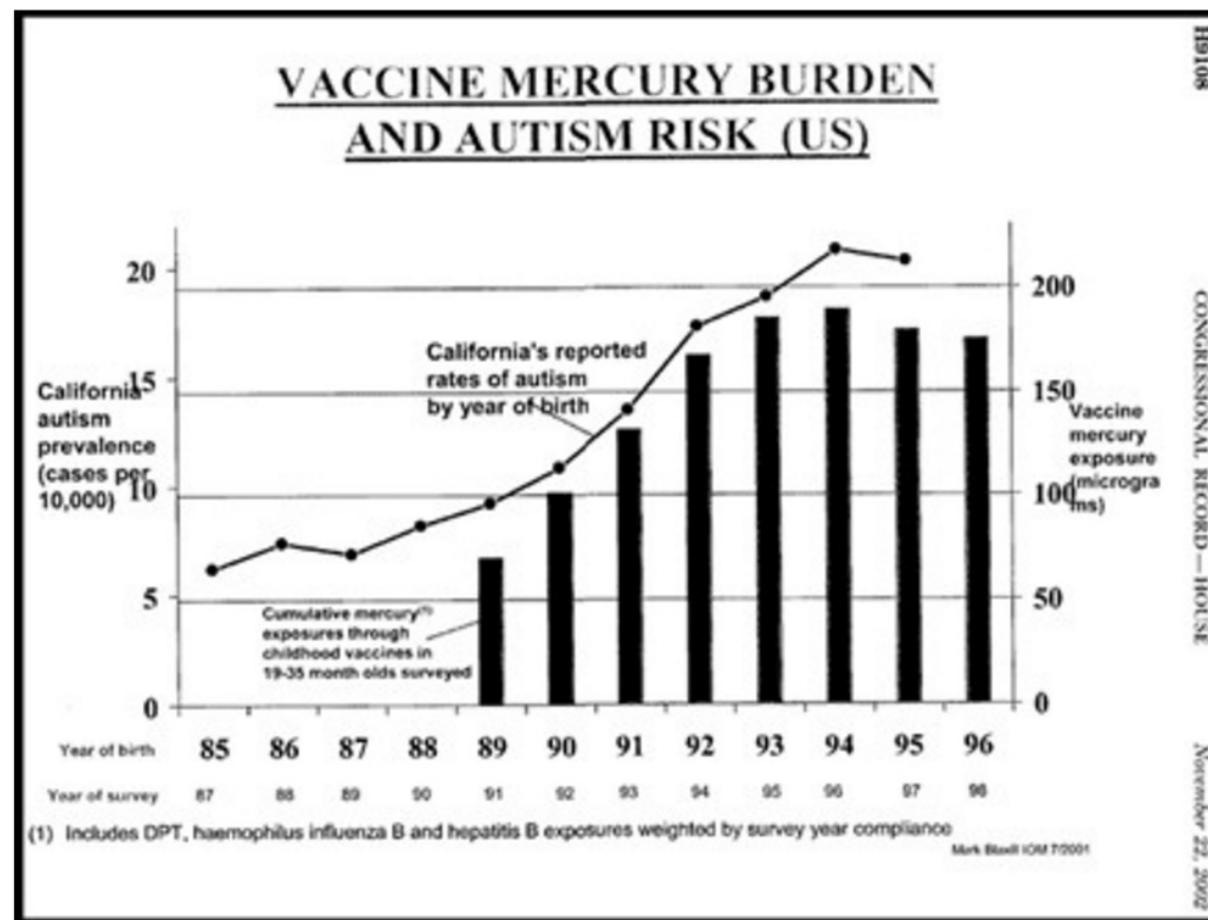
- > "Dietary flavonoids ... which are widely present in cocoa ... have been shown to improve cognitive function. ... I wondered if there is a correlation between a country's chocolate consumption and its population's cognitive function ..." Messerli (2012)



Korrelation

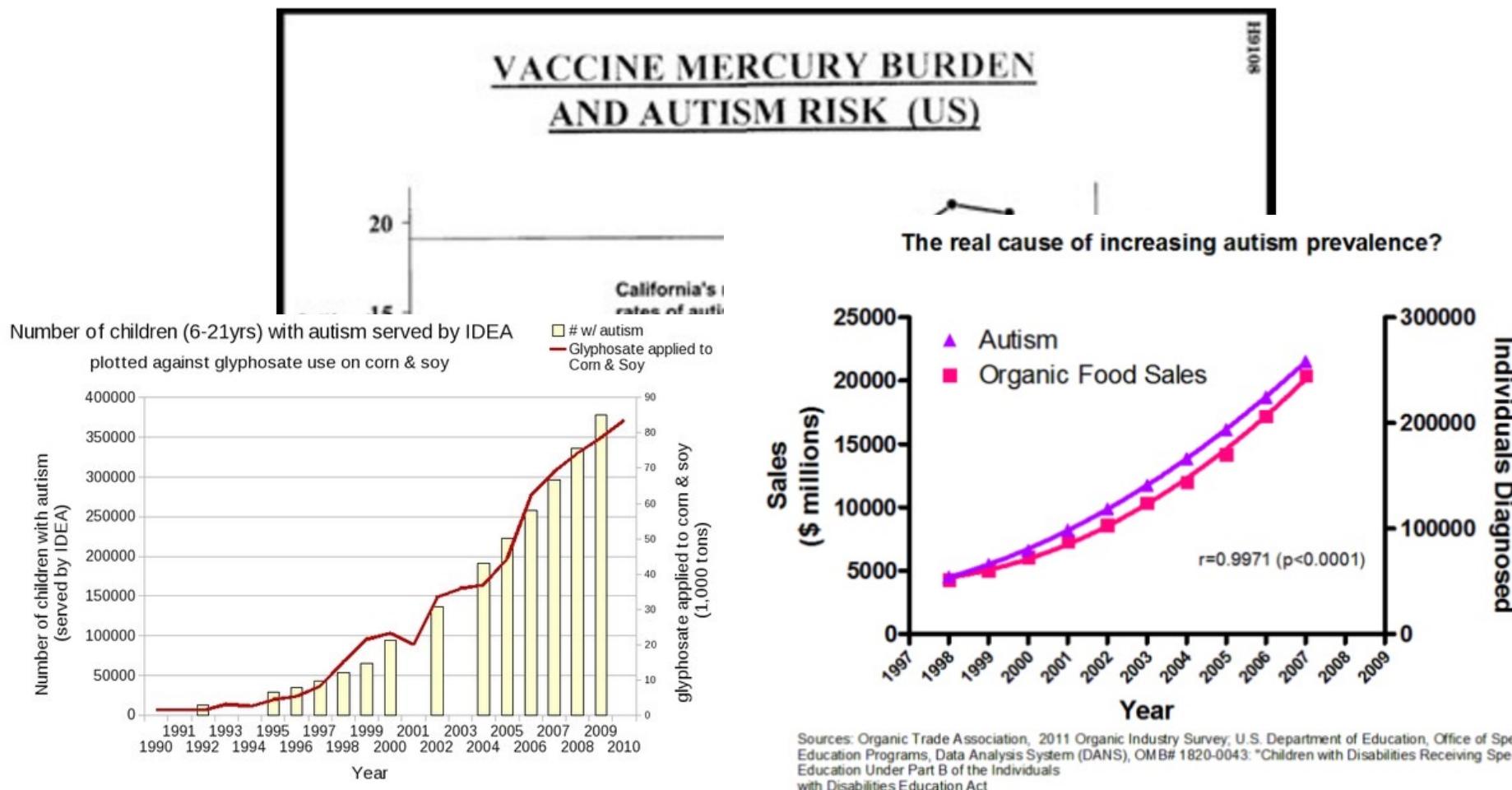
- > “Dietary flavonoids ... which are widely present in cocoa ... have been shown to improve cognitive function. ... I wondered if there is a correlation between a country’s chocolate consumption and its population’s cognitive function ...” Messerli (2012)





Korrelation \neq Kausalzusammenhang

- > Korrelation beschreibt nur die Stärke des Zusammenhangs, KEINE Information über Kausalzusammenhänge



Take-home Message Korrelation

- > Korrelation beschreibt nur die Stärke des Zusammenhangs, KEINE Information über Kausalzusammenhänge
- > Pearson Korrelation nur für metrische Daten ohne Ausreisser und linearen Zusammenhang zwischen den beiden Variablen
- > Spearman Rangkorrelation für metrische und kategoriale Daten, robuster gegenüber Ausreissern und Nicht-Linearität
- > +1 heisst perfekter positiver Zusammenhang; 0 heisst kein Zusammenhang; -1 perfekter negativer Zusammenhang
- > zyklische Schwankungen und Trends können Korrelationskoeffizienten stark beeinflussen

REGRESSION

Bahrenberg I: Kap. 6;
Ernste Kap. 3;
Ewing I: Kap. 12

Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen
Fallen der Statistik			
weiterführende Methoden	Daten zusammenfassen		Extremwertstatistik
	Hauptkomponenten-analyse	Clusteranalyse	Zeitreihenanal. etc.

Regression

Ziel

- > Prognosen oder lineare Inter-, Extrapolationen machen.

Voraussetzung

- > kausaler Zusammenhang zwischen abhängiger und unabhängigen Variablen
- > ein **Modell**, das die Zusammenhänge erklärt:

$$y = a + bx \quad \text{Geradengleichung}$$

Bei der Regressionsanalyse gibt es eine **abhängige Variable (y)**, welche erklärt werden soll und **eine oder mehrere unabhängige Variablen (x)**, die mit der zu erklärenden Variablen in Verbindung stehen.

Regression

$$y = a + bx$$

y ist die abhängige Variable

x ist die unabhängige Variable

a ist eine Konstante (**Regressionskonstante**)

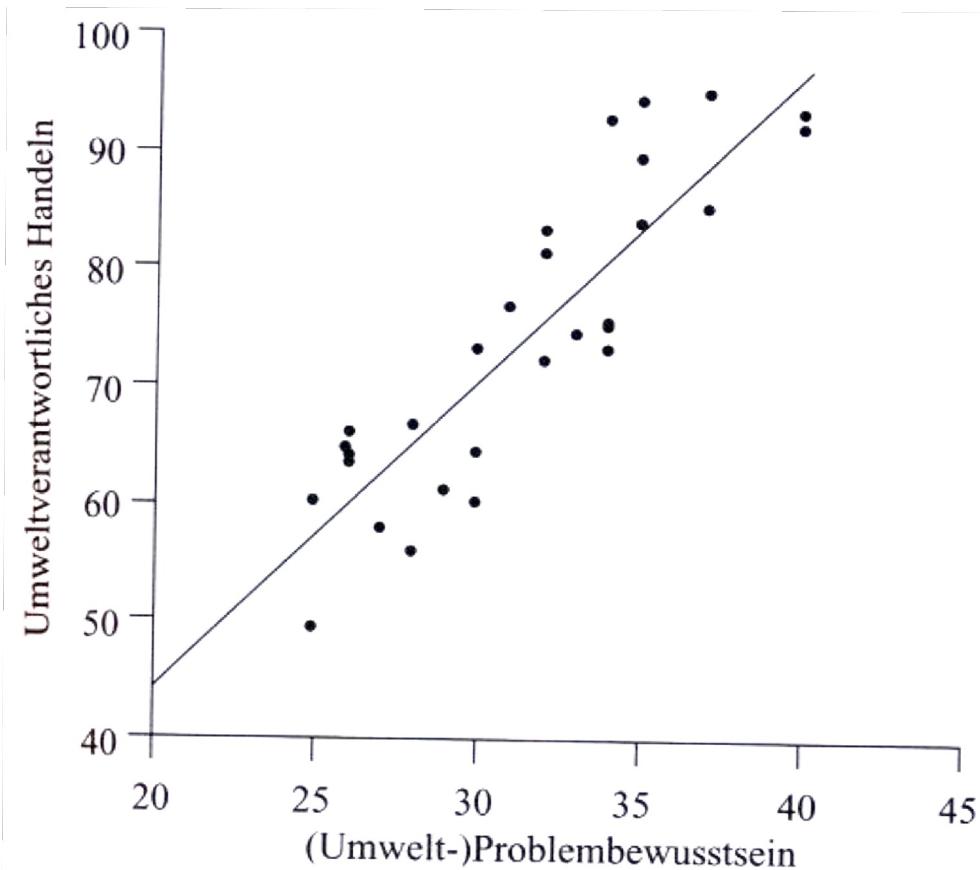
b ist die Steigung der Geraden (**Regressionskoeffizient**)

Für jede Zunahme von x um 1, ändert sich y in Abhängigkeit von b

Es gibt perfekt lineare Zusammenhänge, z.B. Telefonrechnungen:

Gesamtbetrag = Grundgebühr + 0.1 * Gesprächsminuten

Mit Grundgebühr und Anzahl an Gesprächsminuten können die Gesamtkosten der Rechnung exakt vorhergesagt werden.



In Stichproben sind die Zusammenhänge meist nicht perfekt linear, z.B. aufgrund von Messfehlern, weiteren Einflussfaktoren, etc.

Distanzmasse

- > Welches Mass eignet sich am besten, um die Abweichung einer Beobachtung von der Regressionsgeraden zu messen?

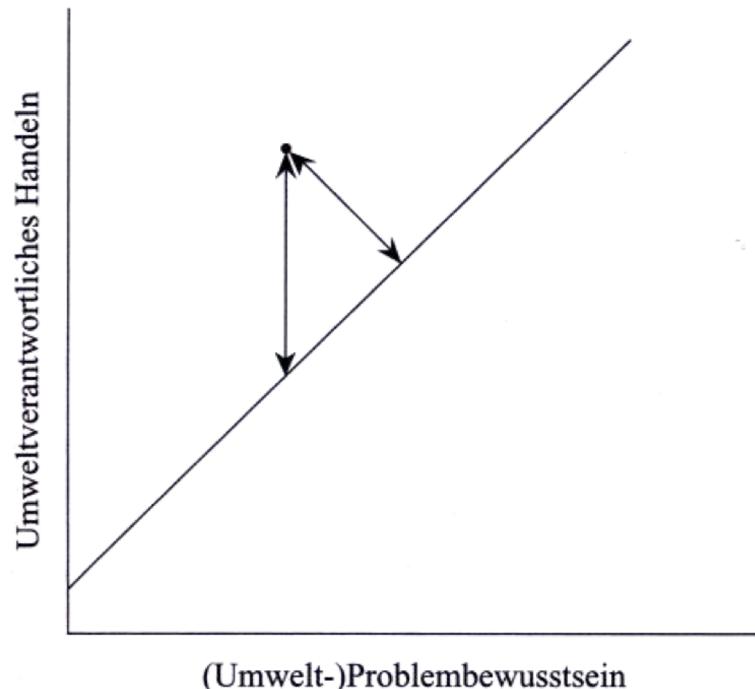


Abbildung 3.3: Verschiedene Entfernungsmethoden zur Regressionsgeraden

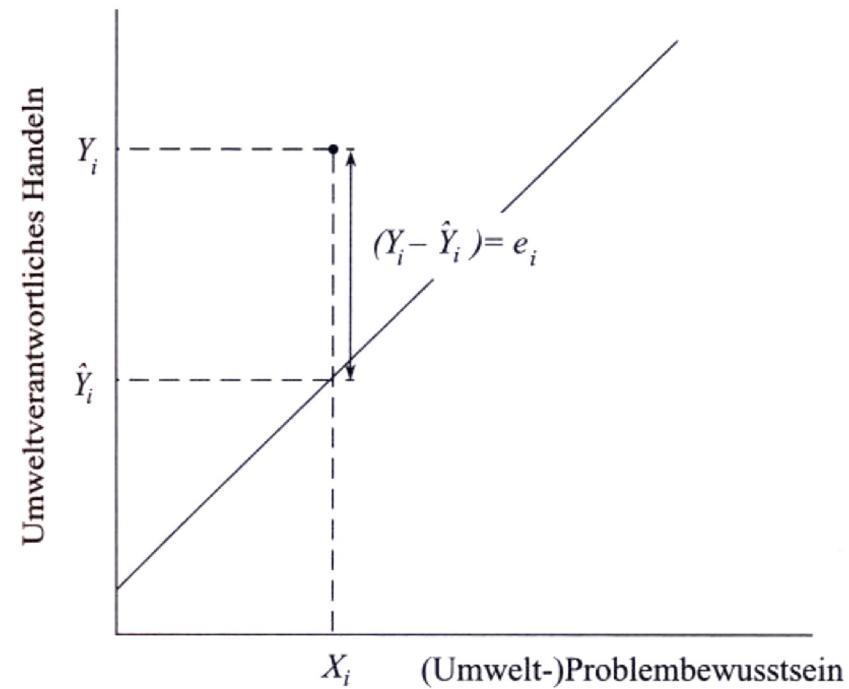


Abbildung 3.4: Vertikale Abweichung oder Residualwert

Optimierung

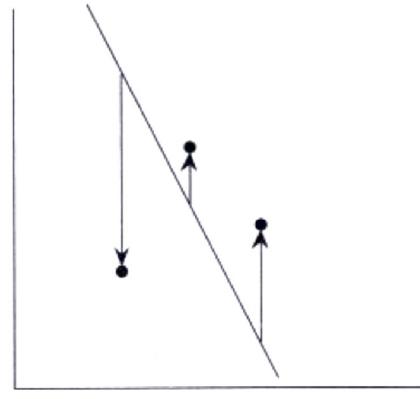
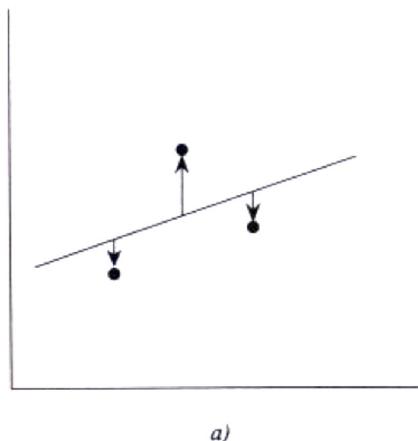
u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Nachdem eine Distanz definiert wurde stellt sich als nächste Frage, welche Funktion minimiert werden soll.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \rightarrow \text{minimal}$$



$$\sum_{i=1}^n |Y_i - \hat{Y}_i| \rightarrow \text{minimal}$$

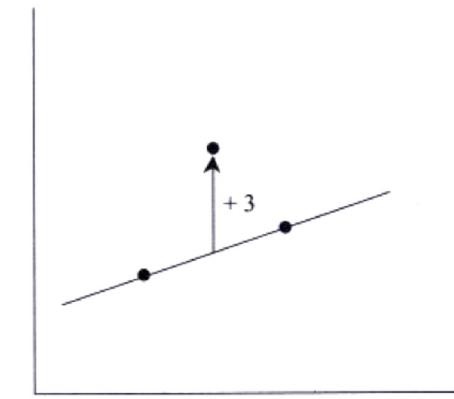
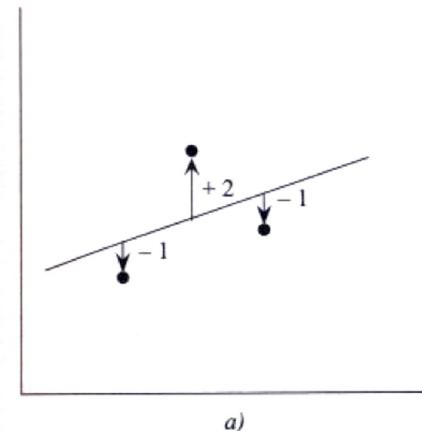


Abbildung 3.5: Zwei Regressionsgeraden, die beide gleich gut zu den beobachteten Punkten passen

Abbildung 3.6: Beispiel zweier Regressionslinien

Methode der kleinsten Quadrate

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \rightarrow \text{minimal}$$

Am häufigsten angewandte Methode in der Statistik (Ordinary Least Squares, OLS)

- > Hat "günstige" statistische Eigenschaften
- > Hat analytische Lösung
- > Wird allerdings durch Ausreisser beeinflusst

Die Regressionsschätzung

u^b

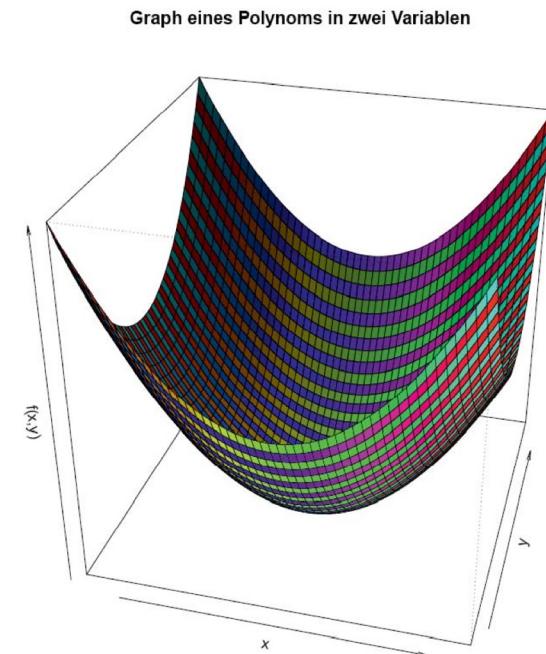
b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Die Y-Werte lassen sich beschreiben durch:
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
 (mit zufälligen Fehlern ε)
- > Zur Bestimmung der Regressionsgerade wollen wir die Regressionskonstante β_0 (Intercept, Schnittpunkt mit der y-Achse) und den Regressionskoeffizienten β_1 (Slope, Steigung) schätzen (das Dach steht für einen Schätzwert), so dass folgende Funktion minimiert wird:

$$\sum_{i=1}^n (Y_i - \hat{\beta}'_0 - \hat{\beta}_1 X)^2$$

- > Minimierung durch Nullsetzen der partiellen Ableitungen



Die Regressionsschätzung

u^b

b
UNIVERSITÄT
BERN

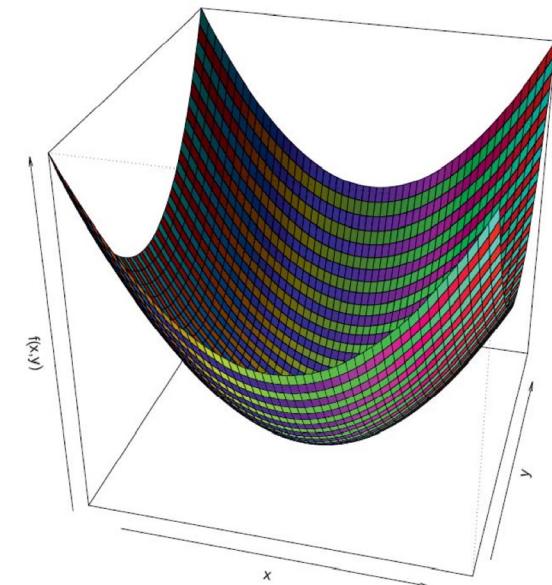
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Minimierung durch Nullsetzen der partiellen Ableitungen lässt sich wie folgt darstellen:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

Graph eines Polynoms in zwei Variablen



Die Regressionsschätzung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

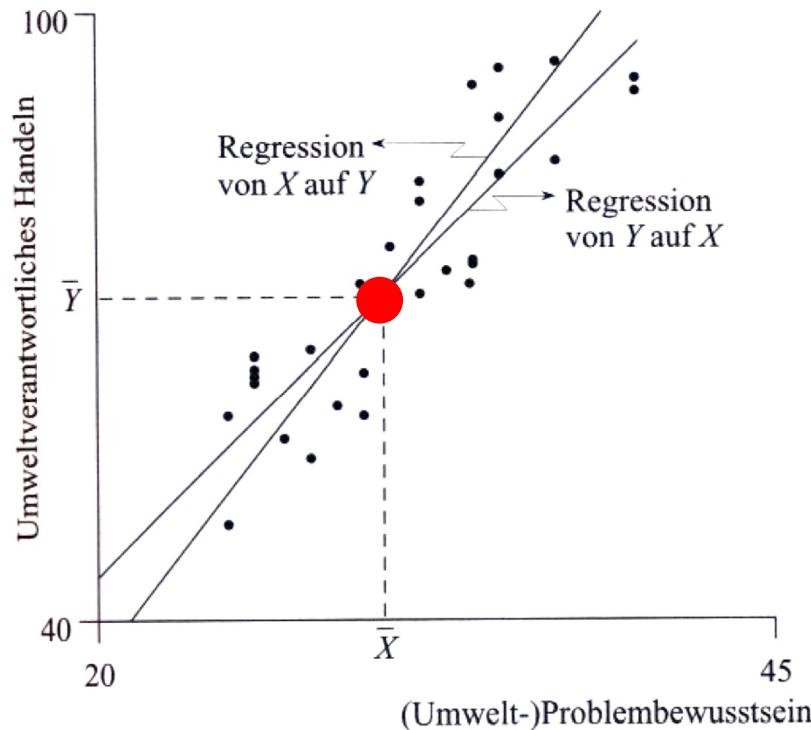


Abbildung 3.7: Regressionsschere

⁸ Der Korrelationskoeffizient r verbindet die beiden Regressionskoeffizienten miteinander, da er den geometrischen Mittelwert aus den beiden Regressionskoeffizienten darstellt: $r = \sqrt{b_{yx} b_{xy}}$.

Regressionsschere:

- > Durch die Minimierung der vertikalen Distanzen ist der Regressionskoeffizient davon abhängig, ob X von Y oder Y von X abhängt. Beide Gerade scheiden sich im Mittelwert von X und Y
- > Normalerweise macht nur eine Variante Sinn, weil ein Kausalzusammenhang besteht und damit klar sein sollte welche Variable (un)abhängig ist!

Güte des Regressionsmodells Varianzaufteilung

U^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) ,$$

a

b

c

totale Abweichung
vom Mittelwert

durch Einfluss von
 X auf Y 'erklärter'
Teil

'unerklärter' Teil

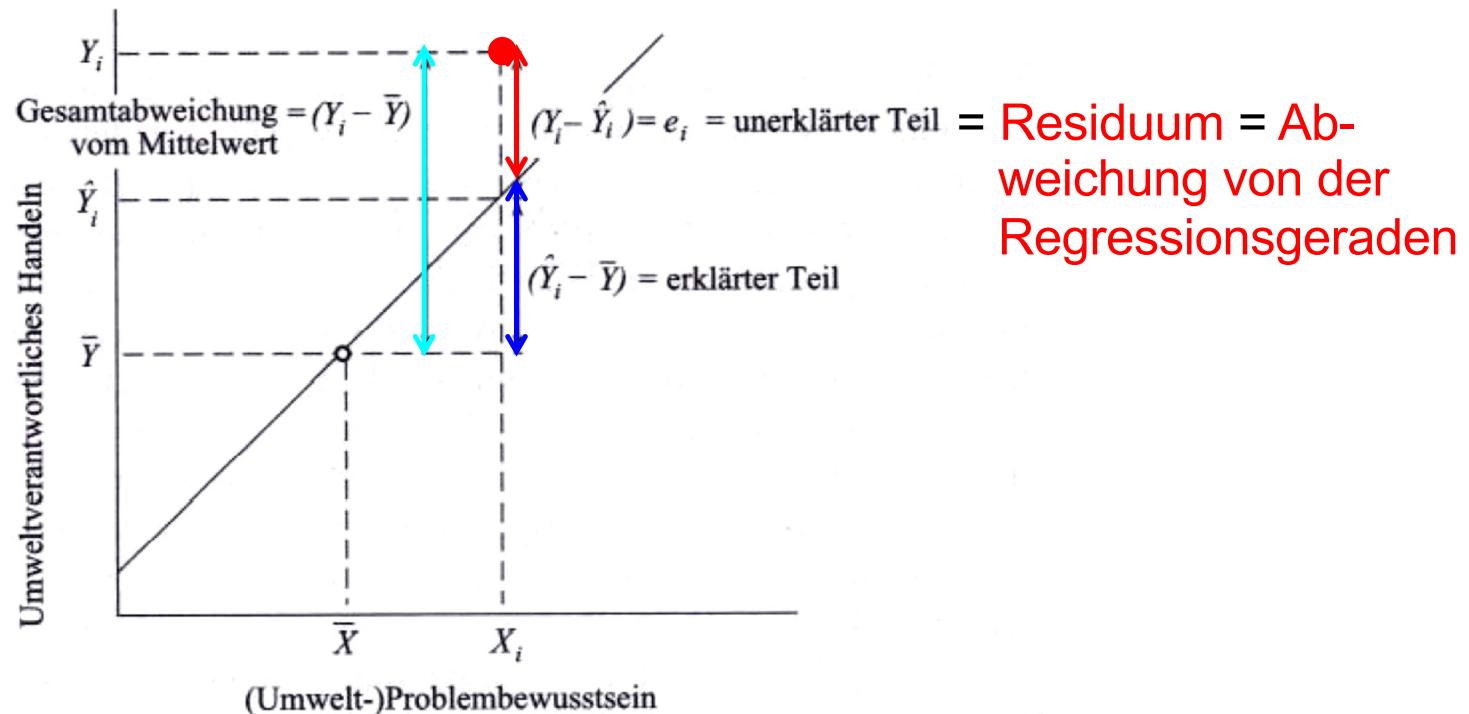


Abbildung 3.8: Variationszerlegung graphisch dargestellt

Bestimmtheitsmass R²

- > Anteil der Variation der abhängigen Variablen Y, der durch die unabhängige Variablen X “erklärt” werden kann

$$R^2 = \frac{\text{erklärte Variation}}{\text{gesamte Variation}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\text{Restvariation}}{\text{Gesamtvariation}}$$

$$R^2 = r^2$$

- > $R^2 = 1$, wenn alle Punkte auf der Regressionsgraden liegen
- > $R^2 = 0$, Modell liefert keinerlei Erklärung für die Variation von Y

Zusammenhang zwischen Stichproben-Korrelation und Regression

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Korrelation

$$r_{xy} = \frac{\sum_{i=1}^N (X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{x})^2 \sum_{i=1}^N (Y_i - \bar{y})^2}}$$

Regression

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

VORSICHT:

Bestimmtheitsmass R^2 ist von vielen Faktoren abhängig, z.B.
Messfehler:

- > Mit neuer Messtechnik verringert sich die Restvariation und R^2 steigt, ohne dass das Modell besser geworden ist
- > Bei multipler Regression (nächste Einheit) nimmt R^2 mit der Anzahl der unabhängigen Variablen zu

Tests für den Regressionskoeffizienten b_1

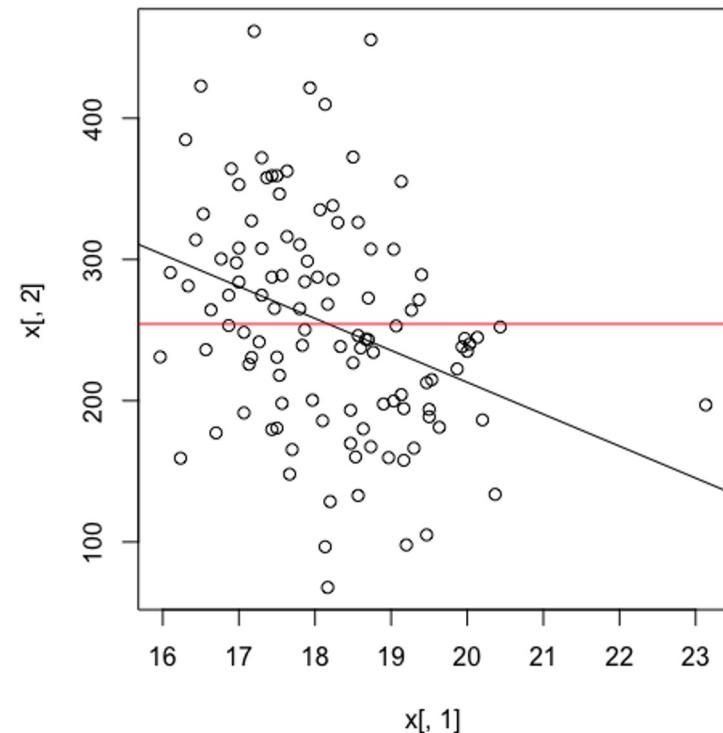
U^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Wie lauten Null- und Alternativhypothese?

- > $H_0:$
- > $H_1:$



Tests für den Regressionskoeffizienten b_1

U^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

$H_0: \beta_j = 0$ (KEINE lineare Abhängigkeit)

$H_1: \beta_j \neq 0$ (linearere Abhängigkeit im Falle eines zweiseitigen Tests)

- > β_1 der Stichprobe variiert um β_1 der Grundgesamtheit mit bestimmter Wahrscheinlichkeitsverteilung und zwar
 - > einer Normalverteilung mit der Standardabweichung σ_{β_1}
 - > σ_{β_1} wird aus s_{β_1} geschätzt
 - > um die Schätzunsicherheiten bei kleinen Stichproben zu berücksichtigen, wird die T-Verteilung benutzt.

$$T = \frac{\hat{\beta}}{s_{\beta_1}} \quad \text{mit } s_{\beta_1} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y})^2 / (n - 2)}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

n-2 Freiheitsgrade, weil wir den Regressionskoeffizienten β_j und die Streuung σ_{β} aus der Stichprobe schätzen müssen.

Konfidenzintervall des Regressionskoeffizienten

U^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Konfidenzintervall für den Regressionskoeffizienten:

$$\beta_1 = \widehat{\beta}_1 \pm q_t s_{\widehat{\beta}_1} \quad \text{mit } q_t \text{ aus T - Tabelle}$$

$\widehat{s}_{\beta} = s_{\widehat{\beta}_1}$ Standardfehler der Steigung

- > Erinnert euch an die **Daumenregel für Konfidenzintervalle**:
- > **Konfidenzintervall = Stichprobenergebnis ± 2 Standardfehler**

p-Wert bei der Regression

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > $H_0: \beta_1 = 0$ (KEINE lineare Abhangigkeit)
- > $H_1: \beta_1 \neq 0$ (linearere Abhangigkeit im Falle eines zweiseitigen Tests)

- > Was bedeutet ein p-Wert von 0.001?

p-Wert bei der Regression

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

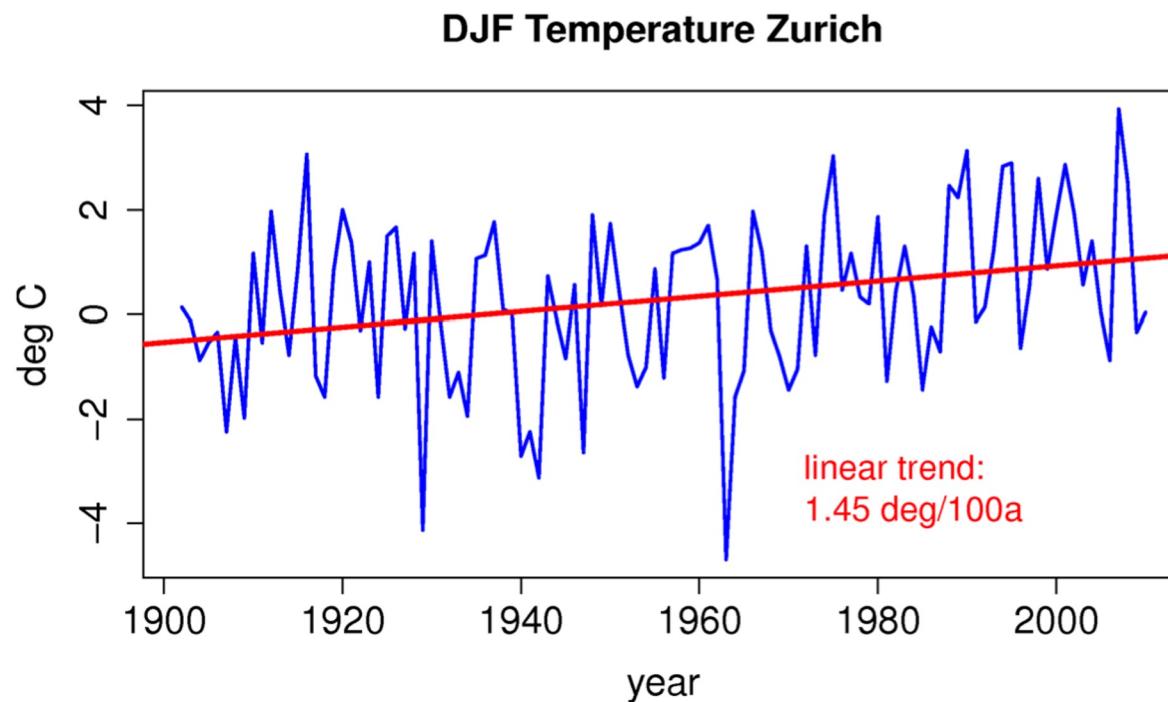
- > $H_0: \beta_1 = 0$ (KEINE lineare Abhängigkeit)
 - > $H_1: \beta_1 \neq 0$ (linearere Abhängigkeit im Falle eines zweiseitigen Tests)
-
-
-
-
-
-
- > Der p-Wert drückt die Wahrscheinlichkeit aus, dass der Regressionskoeffizient für die unabhängige Variable Zufall ist und keinen echten Zusammenhang beschreibt.
 - > Ein p-Wert von 0.001 bedeutet beispielsweise, dass es eine 0.1%ige Chance gibt diese Stichprobe zu erhalten, wenn der Zusammenhang zufällig ist.

Ein Trend über die Zeit ist eine lineare Regression mit der Zeit als unabhängiger Variable

u^b

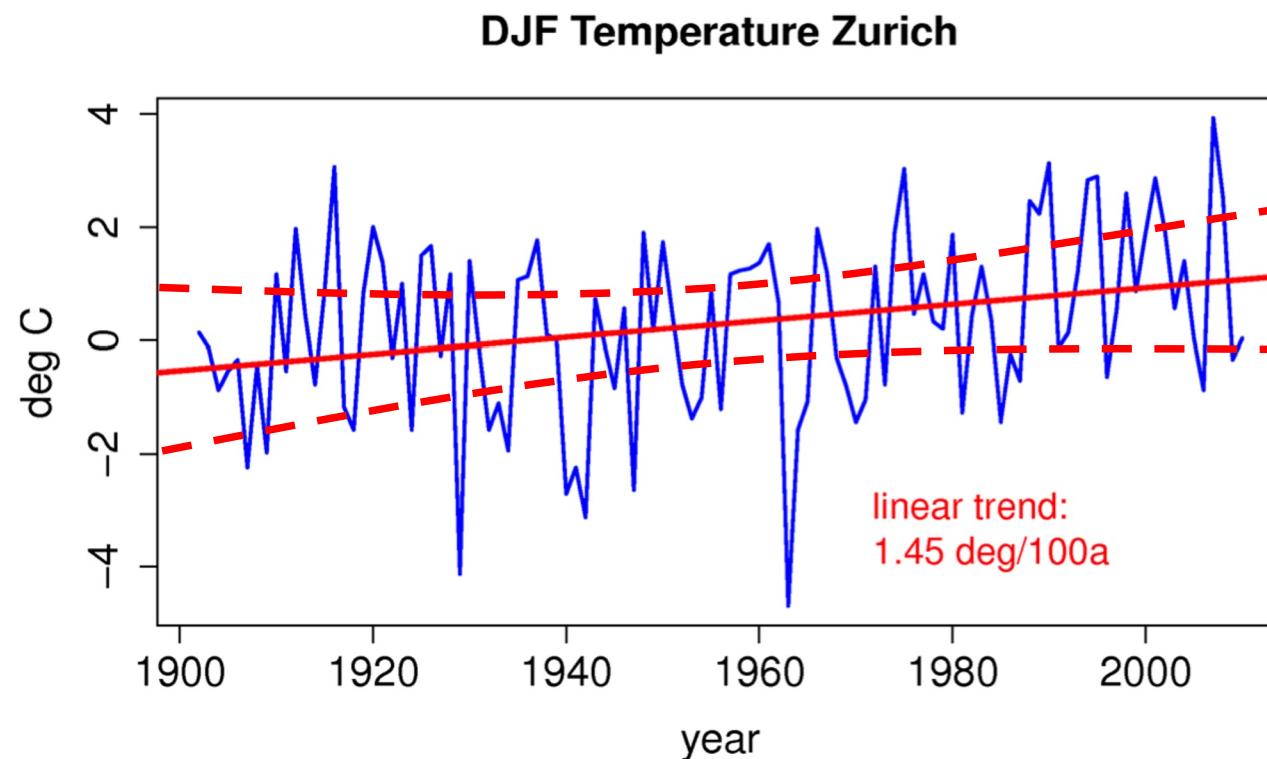
^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Wie würdet ihr das Konfidenzintervall des Regressionskoeffizienten einzeichnen?

Konfidenzintervall des Regressionskoeffizienten

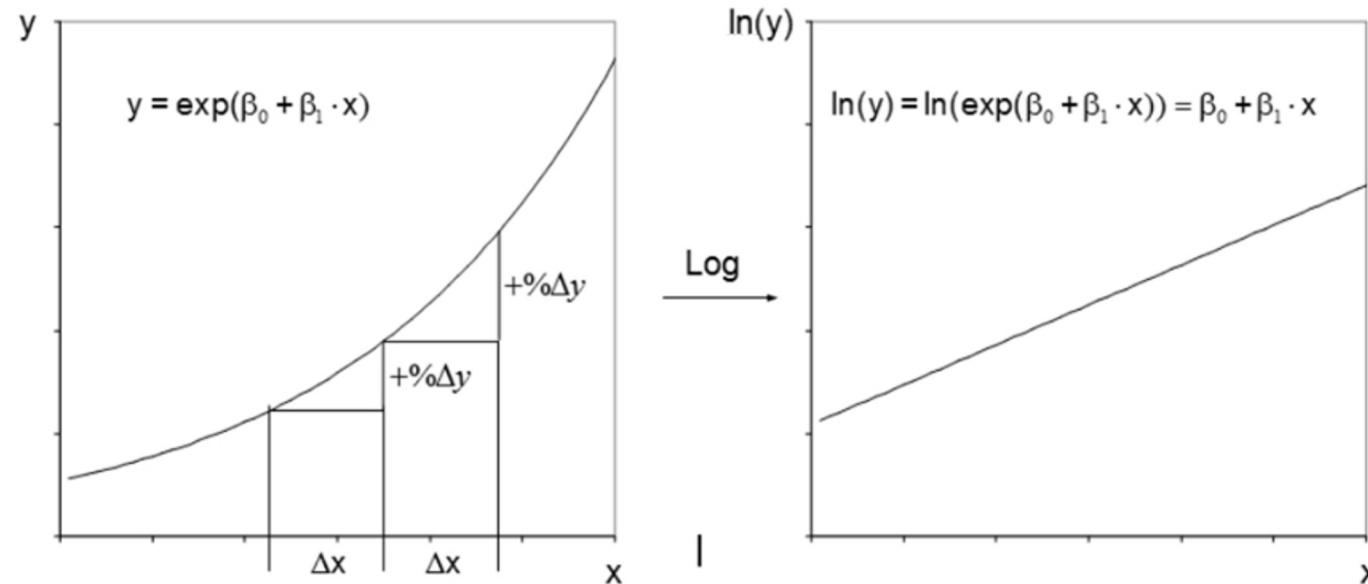


NICHT-lineare Zusammenhänge

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



können oft nach LINEARISIERUNG der Daten modelliert werden!

Beispiele für Prüfungsfragen

- > Schätze die Regressionskonstante b_0 und den Regressionskoeffizienten b_1 aus dem Diagramm.
- > Ausgabe der R lm() Funktion interpretieren wie in Übungsaufgaben
- > Was ist das Bestimmtheitsmaß bei der linearen Regression?
 - Es beschreibt das Verhältnis von erklärter zu unerklärter Variabilität des Regressionsmodells
 - Es beschreibt das Verhältnis von erklärter Variabilität des Regressionsmodells zur Gesamtvariabilität der abhängigen Daten
 - Es beschreibt das Verhältnis von erklärter Variabilität des Regressionsmodells zur Gesamtvariabilität der unabhängigen Daten
 - Es entspricht dem quadrierten Korrelationskoeffizienten

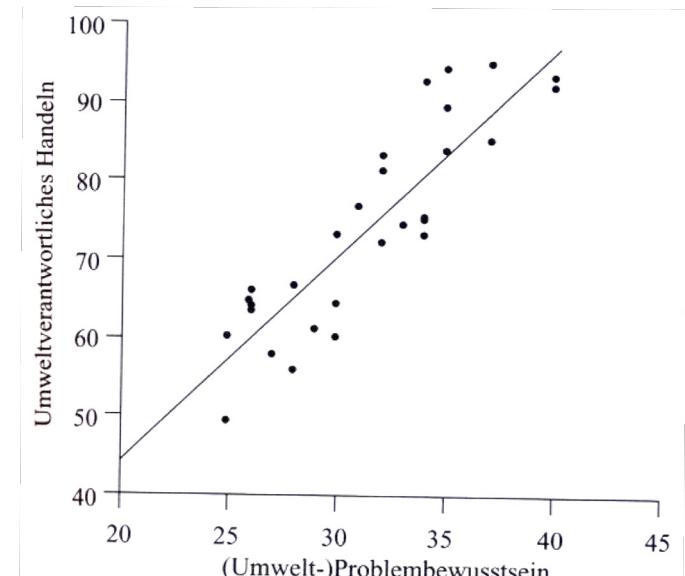


Abbildung 3.2: Streudiagramm umweltverantwortliches Handeln und Problembewusstsein

MULTIPLE REGRESSION

Bahrenberg I: Kap. 2;
Ernste Kap. 4;
Ewing I: Kap. 12

Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen
Fallen der Statistik			
weiterführende Methoden	Daten zusammenfassen		Extremwertstatistik
	Hauptkomponenten-analyse	Clusteranalyse	Zeitreihenanal. etc.

Einfache lineare Regression

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Bei der einfachen linearen Regression haben wir den Störfaktor ε , d.h. die Residuen, als zufälliges Hintergrundrauschen (white noise) betrachtet.

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

- > Oft beinhaltet dieser aber den Einfluss weiterer verursachender Variablen.
- > Ein Hinweis ist oft die grosse Varianz der Residuen.
- > Sollten weitere erklärende Variablen in Modell aufgenommen werden, um eine präzisere Vorhersage zu erhalten?

Streudiagramm mit zusätzlicher Variablen

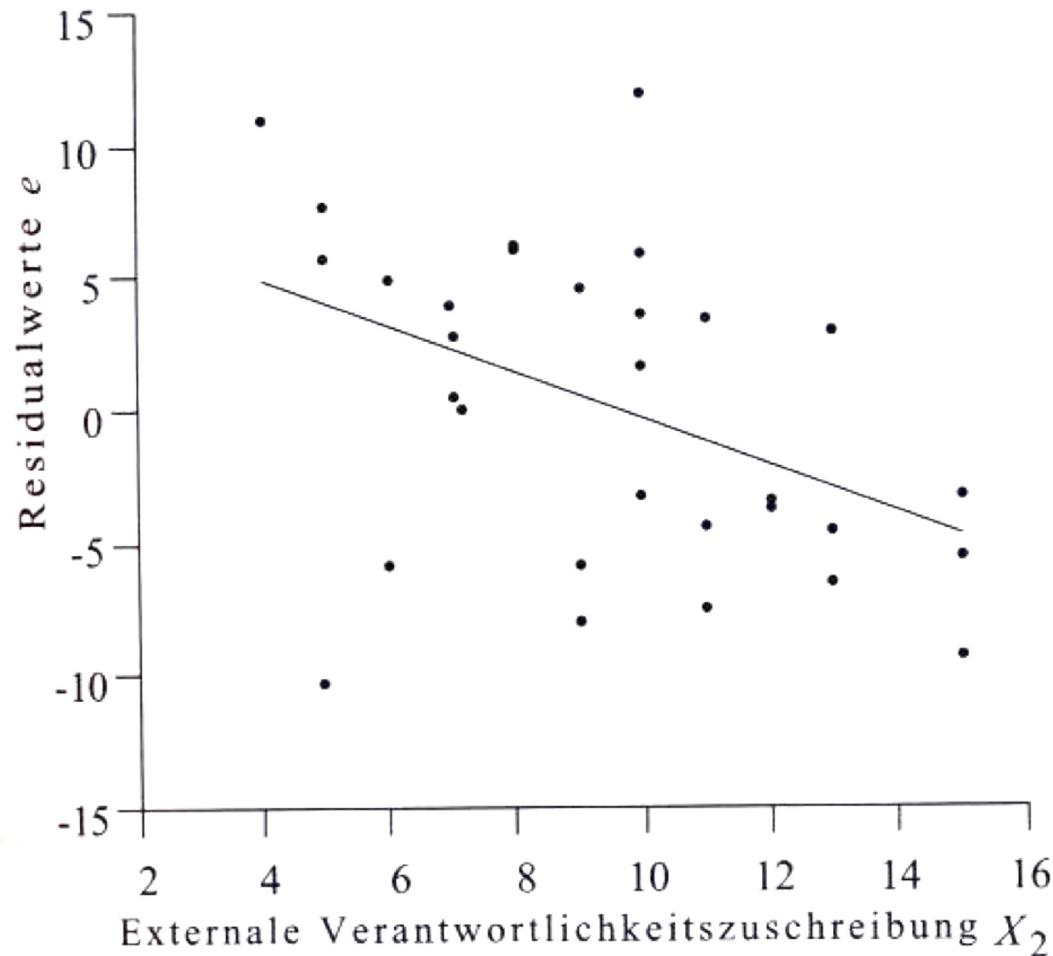


Abbildung 4.1: Streudiagramm der Residualwerte e und der potenziellen zusätzlichen unabhängigen Variablen X_2

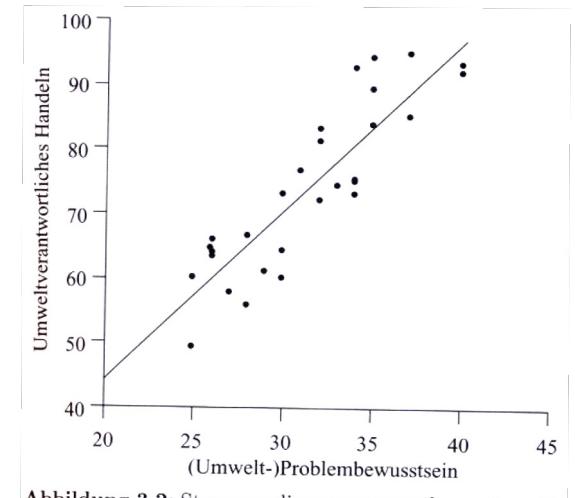


Abbildung 3.2: Streudiagramm umweltverantwortliches Handeln und Problembewusstsein

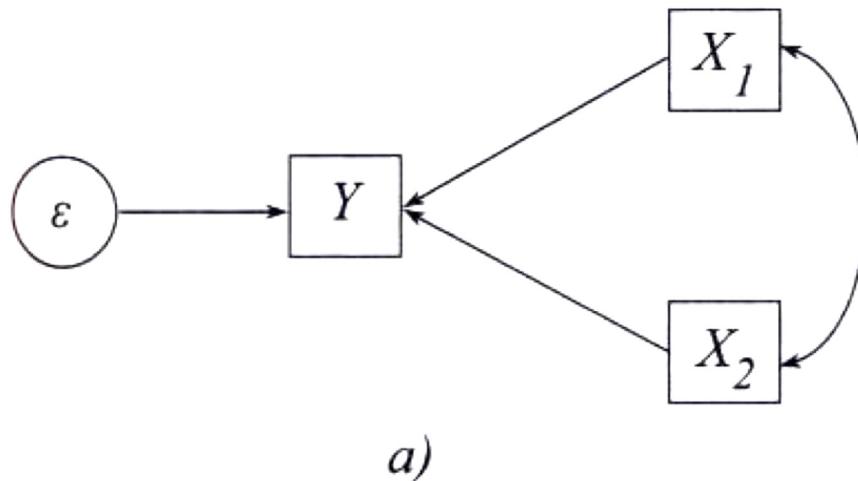
Auswahl sollte
theoriegestützt sein,
um Schein-
korrelationen zu
vermeiden.

Pfaddiagramm mit zusätzlicher Variablen

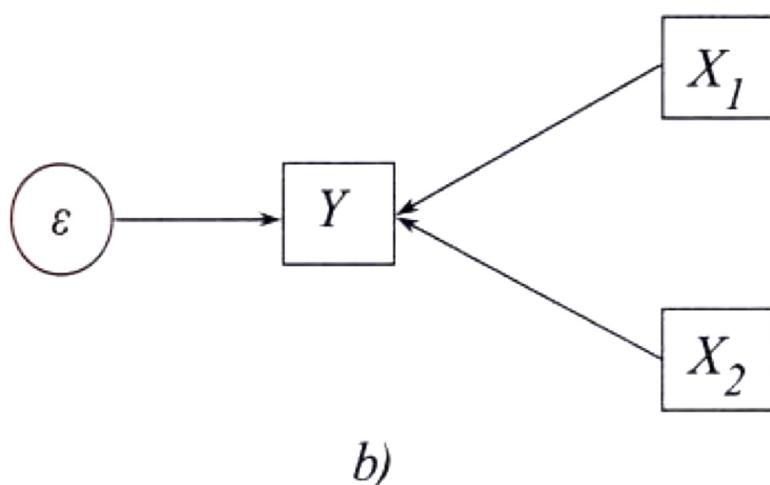
u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



korrelierte unabhängige Variablen,
multiple Regression notwendig



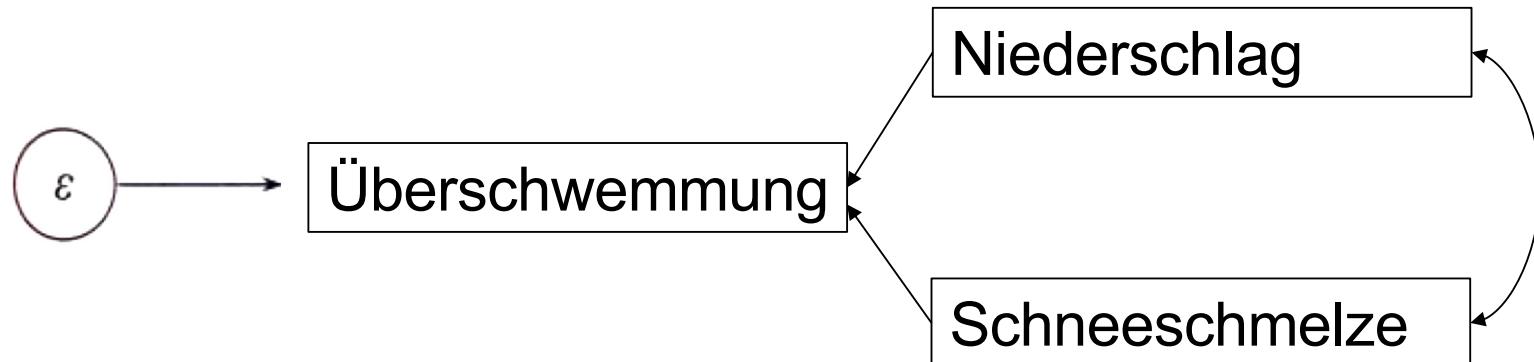
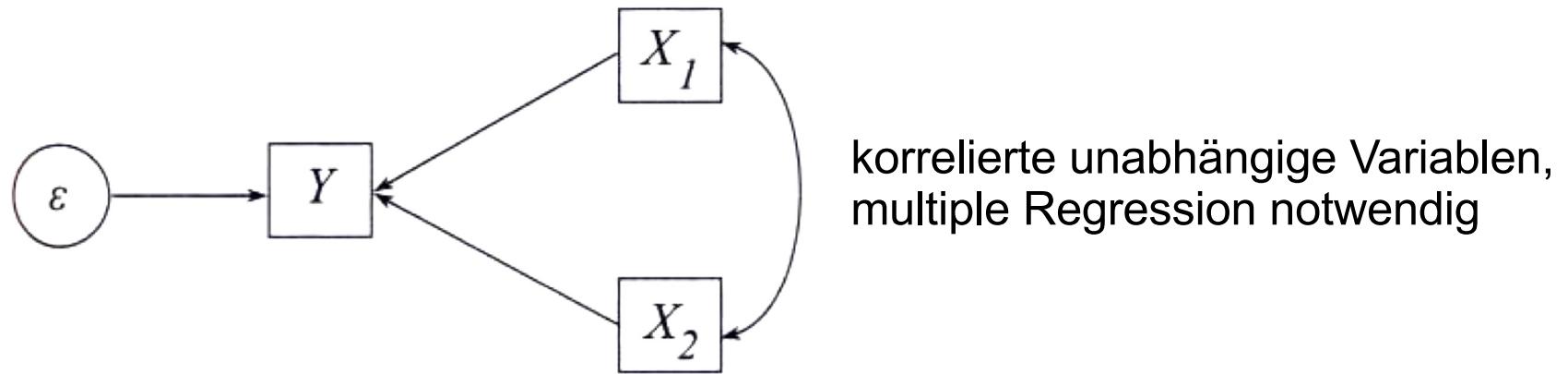
unkorrelierte unabhängige Variablen,
würde Regression der Residuen mit
weiterer erklärender Variable
erlauben

Pfaddiagramm mit zusätzlicher Variablen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Ziel der multiplen Regression

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

1. Etwas modellieren, das von mehreren Variablen beeinflusst wird
2. Den Effekt von einer Variablen untersuchen, unter Berücksichtigung der Effekte weiterer Variablen (explorative Studien)

Beispiel

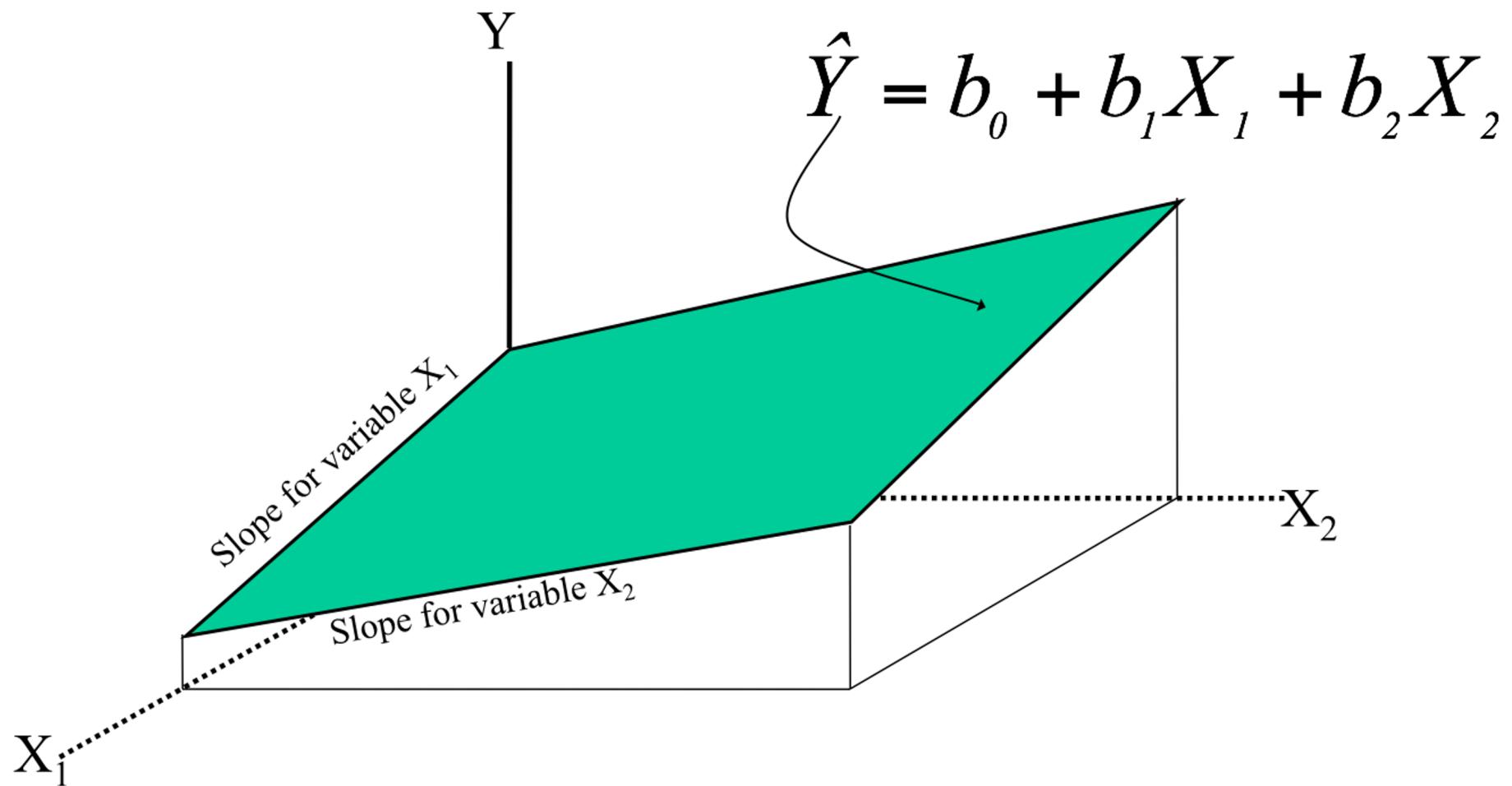
- > Ihr wollt untersuchen, ob Bewegung einen positiven Einfluss auf das Herzinfarktrisiko hat. Die Personen in eurer Untersuchung unterscheiden sich jedoch in vielen anderen Dingen wie Ernährung, Genetik, etc., nicht nur in der Bewegung. Wenn solche Faktoren ebenfalls einen Einfluss auf das Herzinfarktrisiko haben, solltet ihr diese ausschliessen. Ein solcher Faktor könnte z.B. das Alter sein. Hier würde man vom „positiven Einfluss von Bewegung auf das Infarktrisiko **korrigiert für das Alter**“ sprechen.
- > Dies ist insbesondere dann notwendig, wenn man vorhandene Daten nutzt und keine eigenen Daten erheben kann.

Die Regressions-(hyper-)ebene

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Schätzung der Koeffizienten

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Minimierung der quadrierten Abweichungen

$$\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \dots)^2 \rightarrow \text{minimal}$$

- > Dafür müssen die partiellen Ableitungen nach den Koeffizienten gleich Null sein. (Herleitung z.B. in Ernste S. 86ff)
- > Koeffizienten der multiplen Regression werden auch "partielle Regressionskoeffizienten" genannt.

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \dots + \widehat{\beta}_m X_m$$

Schätzung der Regressionskonstante

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_m X_m$$

- > oder in matrixschreibweise:

$$\hat{Y} = \mathbf{X} \hat{\boldsymbol{\beta}}$$

- > mit

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1m} \\ 1 & X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & & & & \\ 1 & X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix}$$

Multiple Regression

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Kind.	2	5	1	9	6	3	0	3	7	7	2	5	1	9	6	3	0	3	7	14
Ausb.	12	16	20	12	9	18	16	14	9	12	12	10	20	11	9	18	16	14	9	8
Einkommen (*1000)	30	40	90	50	40	120	100	10	40	30	100	40	90	40	40	120	100	60	40	10

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \dots + \widehat{\beta}_m X_m$$

$$\widehat{Y}_i = 12 - 0.3X_1 - 0.04X_2$$

$$\text{Geschätzte Anzahl Kinder} = 12 - 0.3 \text{ Ausbildung} - 0.04 \text{ Einkommen}$$

- > Die Steigung ist der Effekt einer unabhängigen Variablen, wenn alles andere Konstant bleibt, also der Einfluss aller anderen Variablen ausgeschaltet wird.
- > Die Koeffizienten kann man nicht vergleichen, da sie von den Einheiten abhängen!

Multiple Regression

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Was sagt uns diese Gleichung?

$$\hat{Y}_i = 12 - 0.3X_1 - 0.04X_2$$

Geschätzte Anzahl Kinder = 12 - 0.3 Ausbildung - 0.04 Einkommen

Bei Ausbildung & Einkommen → Anzahl Kinder

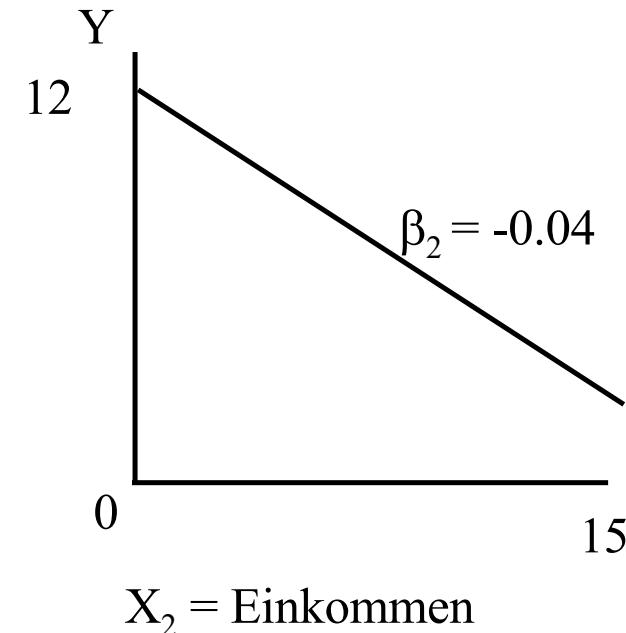
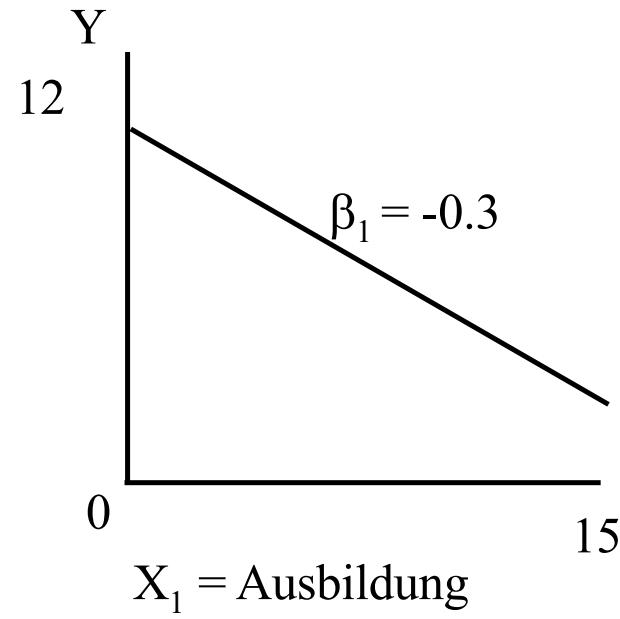
0	0	12
10	0	9
10	100	5
20	200	-2

Multiple Regression

Was sagt uns diese Gleichung?

$$\hat{Y}_i = 12 - 0.3X_1 - 0.04X_2$$

Geschätzte Anzahl Kinder = 12 – 0.3 Ausbildung – 0.04 Einkommen



Wenn eine Variable konstant gehalten wird, kann man es wieder 2D ausdrücken

Güte des Regressionsmodells

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) ,$$

a

b

c

totale Abweichung
vom Mittelwert

durch Einfluss von
X auf *Y* 'erklärter'
Teil

'unerklärter' Teil

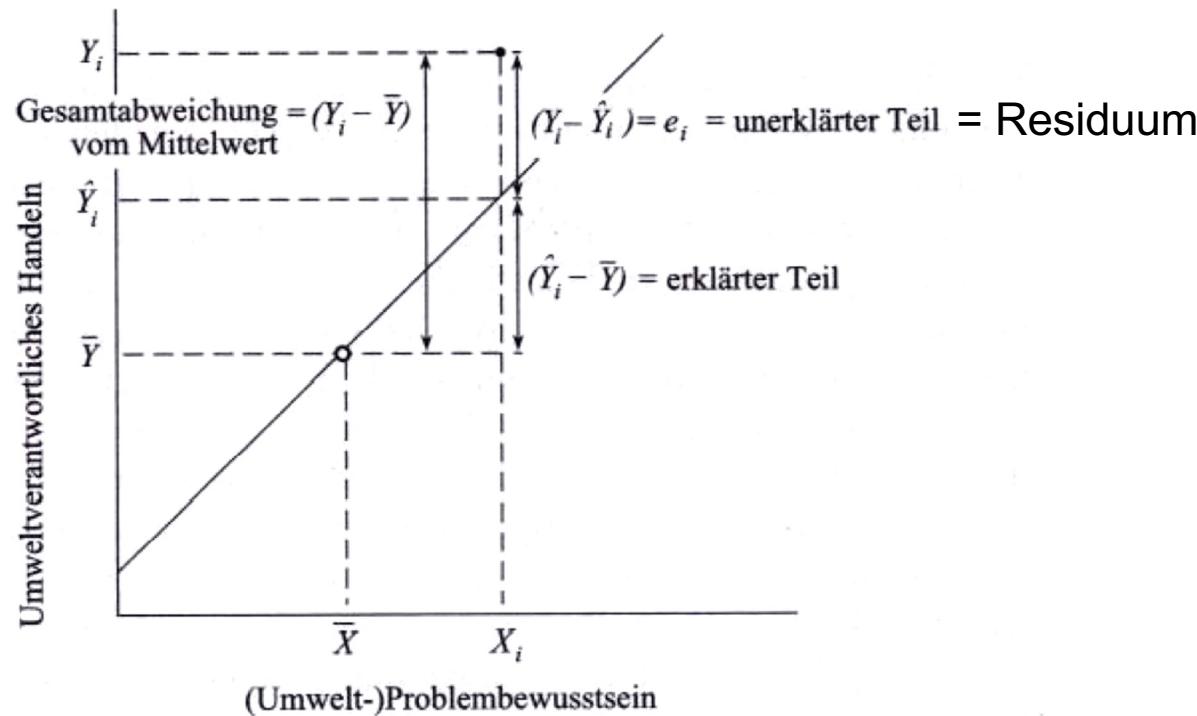


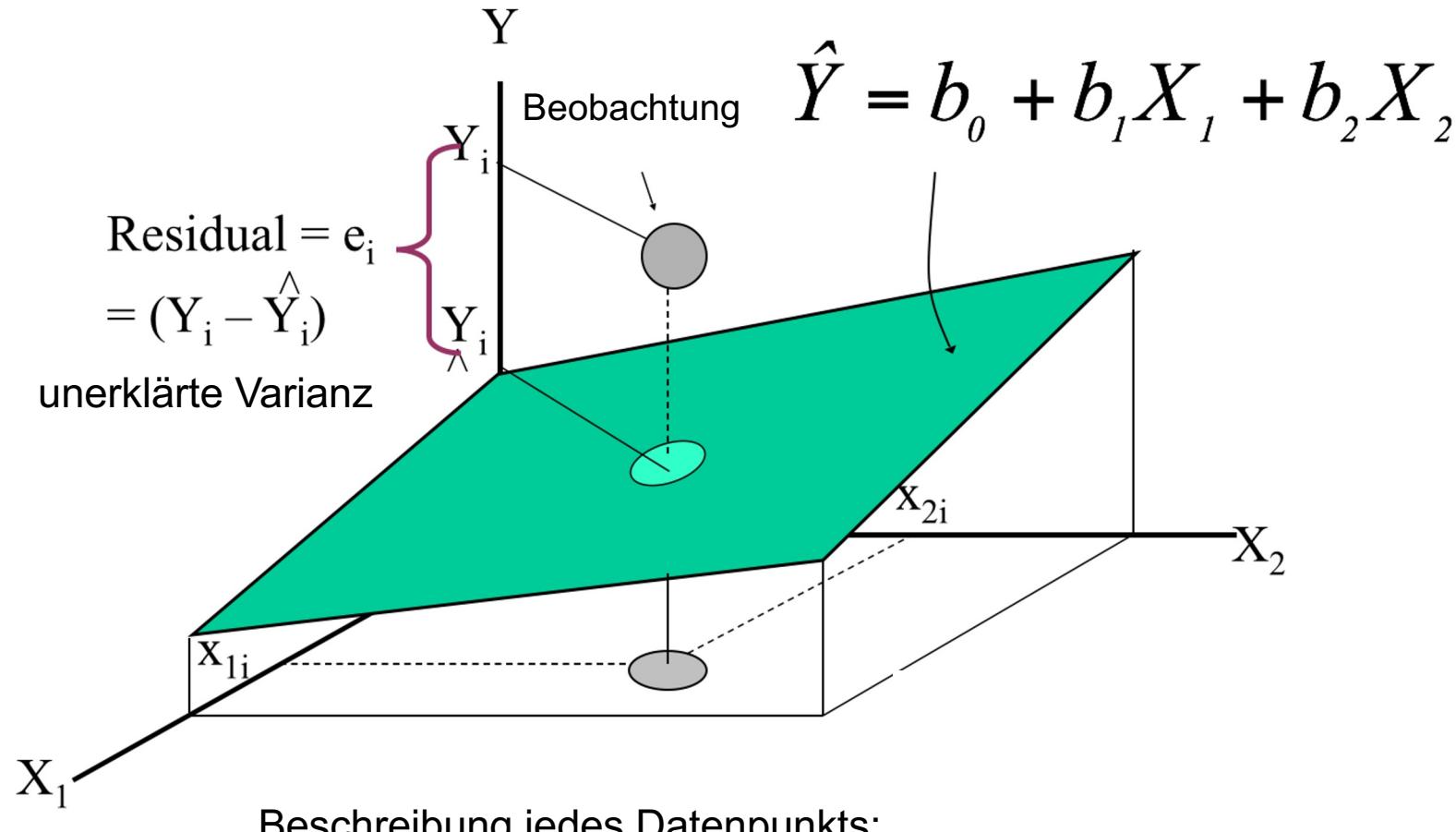
Abbildung 3.8: Variationszerlegung graphisch dargestellt

Residuen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

Der multiples Bestimmtheitsmass

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > $R^2 = 1$, wenn alle Punkte auf der Regressions-Hyperebene liegen
- > $R^2 = 0$, Modell liefert keinerlei Erklärung für die Variation von Y, Regressions-Hyperebene parallel oder senkrecht zu den X_i -Achsen
- > Bei multipler Regression nimmt R^2 mit der Anzahl der unabhängigen Variablen zu.
- > deshalb nutzt man ein **angepasstes Bestimmtheitsmass**

$$R_{\text{angepasst}}^2 = \left(R^2 - \frac{m}{n-1} \right) \left(\frac{n-1}{n-m-1} \right)$$

- > wobei m = Anzahl unabhängiger Variablen
- > das angepasste Bestimmtheitsmass ist immer kleiner als das NICHT ANGEPASSTE Bestimmtheitsmass und nimmt mit zunehmender Anzahl an Variablen ab, falls diese nichts zur Erklärung der Varianz beitragen

Zusammenfassung Varianzzerlegung

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Angepasstes multiples Bestimmtheitsmass (siehe auch R Ausgabe für ein Beispiel)

$$R_{\text{angepasst}}^2 = \left(R^2 - \frac{m}{n-1} \right) \left(\frac{n-1}{n-m-1} \right) = 1 - (1 - R^2) \frac{n-1}{n-m-1}$$

- > Beispiel: Wie verändert sich das angepasste R^2 , wenn nur 3 statt 14 Variablen in die multiple Regression eingehen?

$$R_{\text{angepasst}}^2 = 1 - (1 - 0.5) \frac{31 - 1}{31 - 14 - 1} = 0.06$$

$$R_{\text{angepasst}}^2 = 1 - (1 - 0.5) \frac{31 - 1}{31 - 3 - 1} = 0.44$$

Wie viele unabhängige Variablen?

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Wie kann man testen, ob mehr unabhängige Variablen wirklich besser sind als ein einfacheres Modell?

Beispiel:

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3 + \widehat{\beta}_4 X_4 + \widehat{\beta}_5 X_5 + \widehat{\beta}_6 X_6$$

Im Gegensatz zu:

$$\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3$$

T-Test mit H_0 : kein signifikanter Einfluss der Variablen auf das Modell, z.B. hier $H_0: \beta_4 = \beta_5 = \beta_6 = 0$.

Multiple Regression mit R

U^b

```
> summary(lm(temp[1:365]~s1[1:365]+s2[1:365]+bew[1:365]))
```

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Call:

```
lm(formula = temp[1:365] ~ s1[1:365] + s2[1:365] + bew[1:365])
```

Residuals:

Min	1Q	Median	3Q	Max
-8.586	-2.248	-0.110	2.228	8.008

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	p-Werte
(Intercept)	17.16540	0.32029	53.594	<2e-16 ***	
s1[1:365]	-2.61006	0.23655	-11.034	<2e-16 ***	
s2[1:365]	-8.50760	0.25868	-32.889	<2e-16 ***	
bew[1:365]	-0.77605	0.08875	-8.744	<2e-16 ***	

Regressionskonstante und -koeffizienten

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Bestimmtheitsmass: Regressionsmodell erklärt 83% der Variabilität in Y

Residual standard error: 3.196 on 361 degrees of freedom

Multiple R-squared: 0.8319, Adjusted R-squared: 0.8305

F-statistic: 595.5 on 3 and 361 DF, p-value: < 2.2e-16

Wie viele unabhängige Variablen?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > ACHTUNG: Variablenauswahl mit Sachverstand, d.h. es muss ein Kausalzusammenhang vorhanden sein
- > Statische Auswahl mittels „Forward Selection“ oder „Backward Rejection“
- > **Ziel: So wenige unabhängige Variablen wie möglich** um „OVERRFITTNG“ zu vermeiden

F-Test des gesamten Regressionsmodels

- > T-Test für Regressionskoeffizienten
- > F-Test für die gesamte Regressionsgleichung (mit k erklärenden Variablen für multiple Regression)

$$F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$$

$$F = \frac{\text{erklärte Varianz}}{\text{unerklärte Varianz}}$$

- > Der F-Wert sagt, ob das Model besser ist als einfach die Annahme des Mittelwerts
- > D.h., ob $H_0: R^2 = 0$ abgelehnt werden kann

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

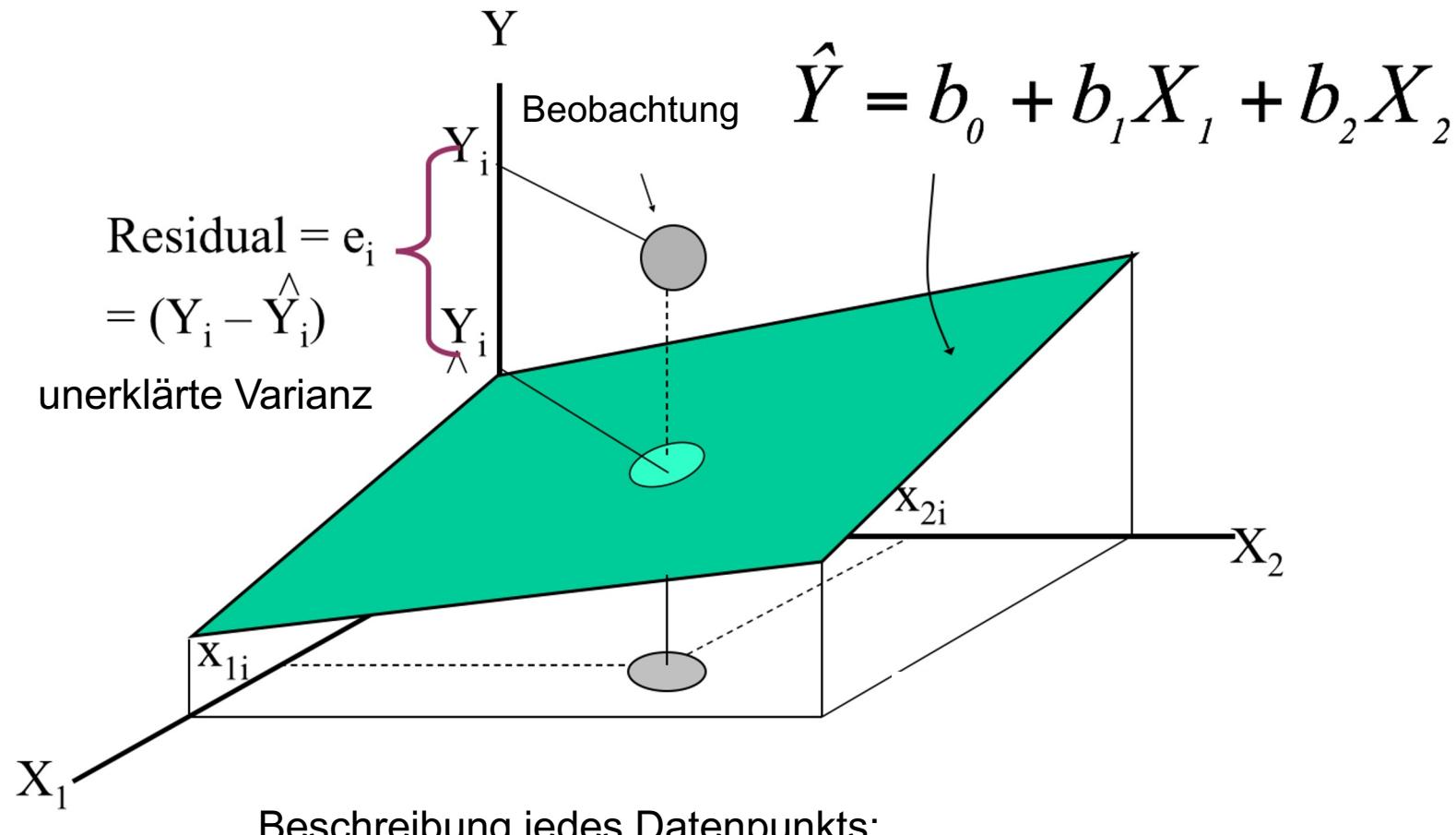
ANWENDUNGSBEDINGUNGEN DER REGRESSIONSANALYSE

Residuen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



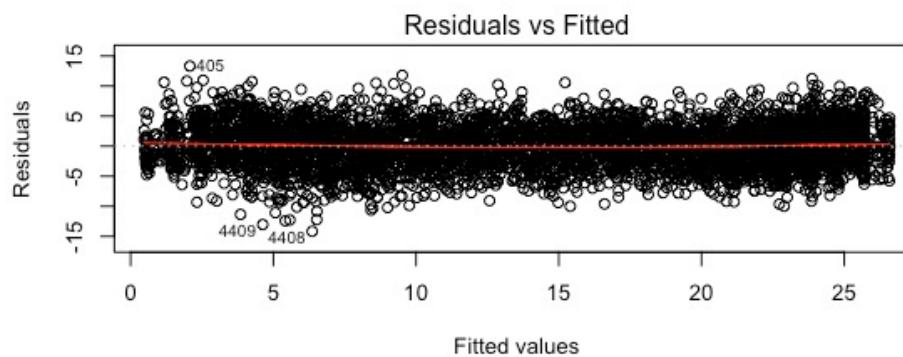
$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

Residuendiagnostik

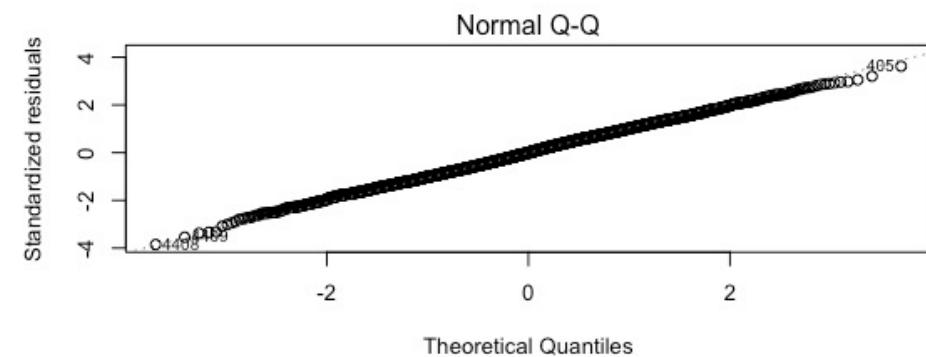
Anwendungsbedingungen ERFÜLLT

- > fit <- lm(y~x)
- > plot(fit)

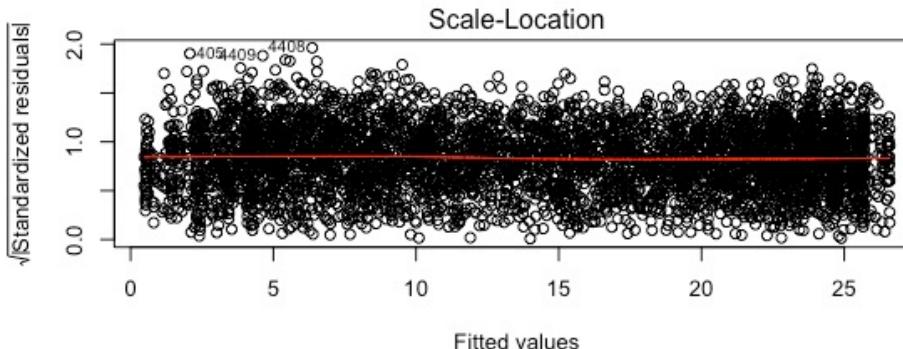
1. Linearität



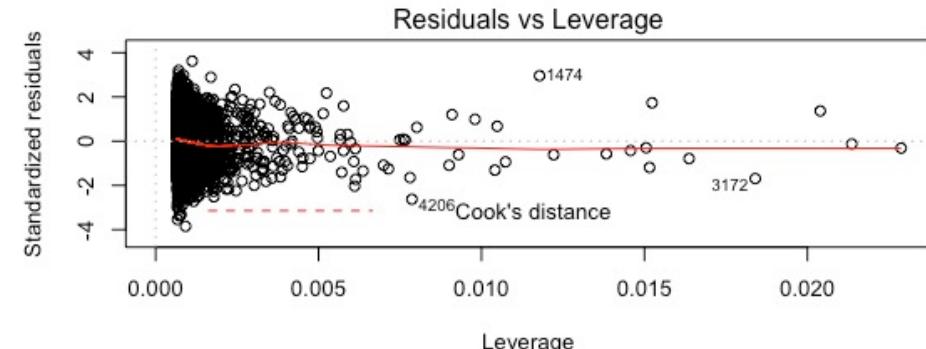
2. Normalverteilung der Residuen



3. Varianzhomogenität



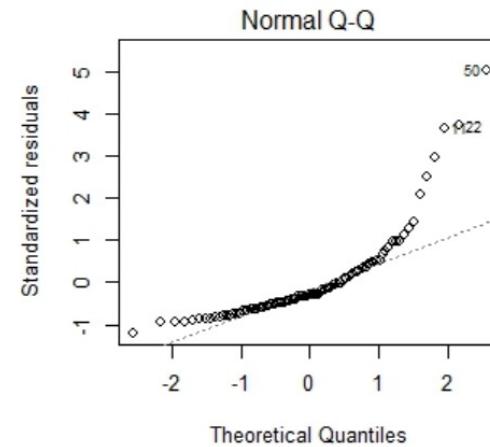
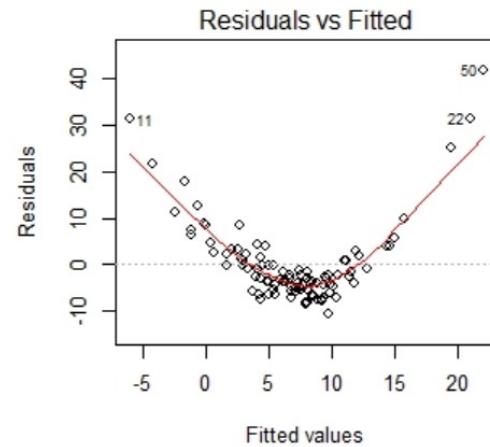
4. Ausreisser mit starken Einfluss?



Residuendiagnostik

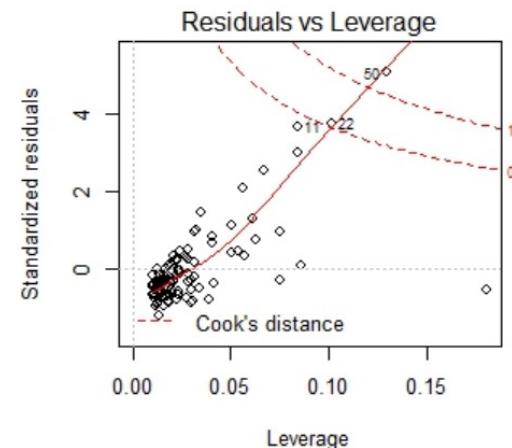
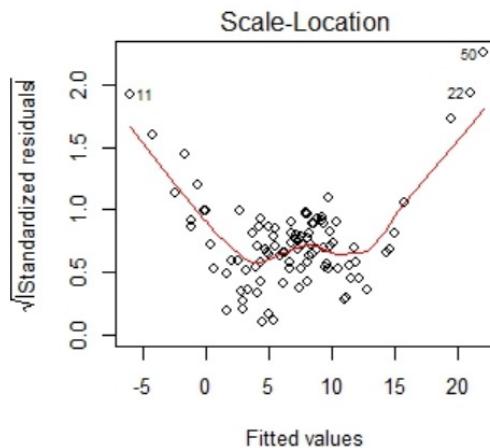
Anwendungsbedingungen NICHT erfüllt

1. Linearität



2. Normalverteilung der Residuen

3. Varianzinhomogenität



4. Ausreißer mit starken Einfluss?

Normalverteilung der Residuen

- > **ACHTUNG: Werte der Variablen müssen NICHT normalverteilt sein,** z.B. bei der Nutzung der linearen Regression zur Bestimmung des Trends über die nicht normalverteilte Zeit!
- > Normalverteilung der Residuen in der Grundgesamtheit ist Voraussetzung für t- und F-Test der Regressionskonstante und des Regressionskoeffizienten
- > Dazu Verteilung aus Stichprobe schätzen

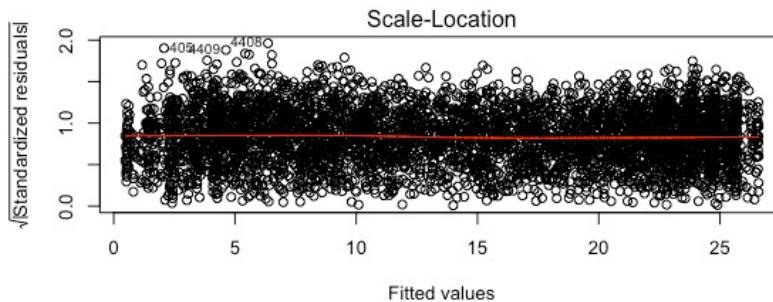
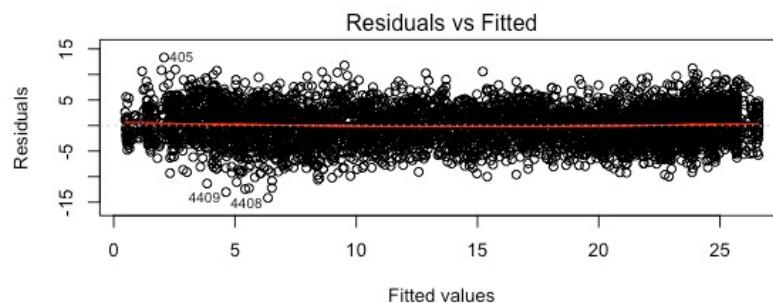
Voraussetzung:

- > Residuenverteilung kann nur geprüft werden, wenn Linearität und Varianzhomogenität gegeben

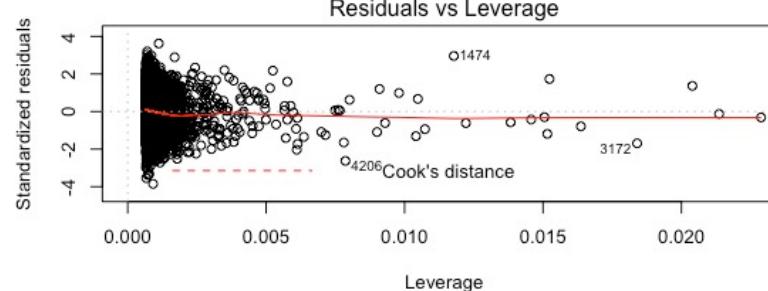
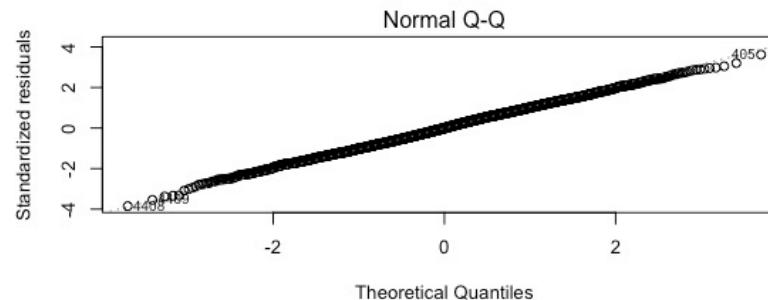
Normalverteilung der Residuen

Prüfung:

1. Quantil-Quantil (QQ) Diagramm wird in R angezeigt unter `plot(fit)` bzw. `plot(lm(y~x))`
2. Shapiro-Wilks Test



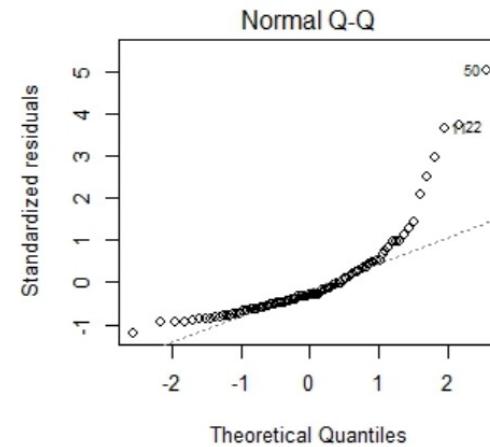
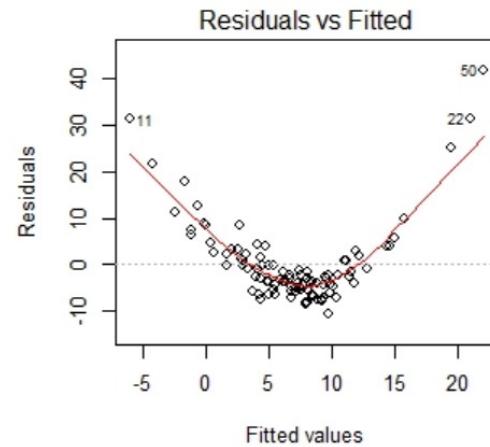
2. Normalverteilung der Residuen



Residuendiagnostik

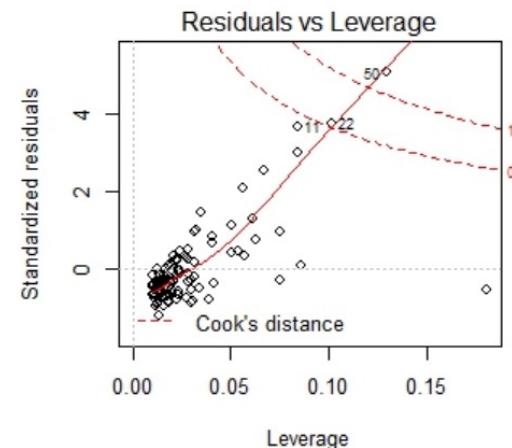
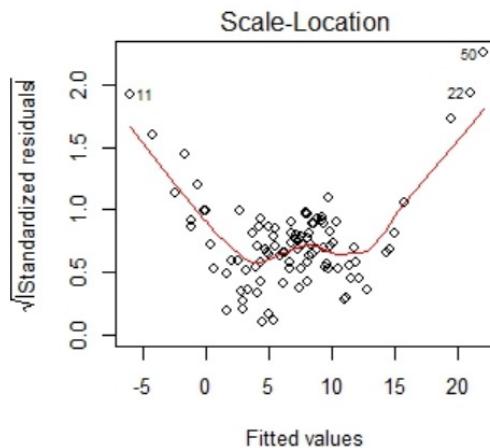
Anwendungsbedingungen NICHT erfüllt

1. Linearität



2. Normalverteilung
der Residuen

3. Varianz-
inhomogenität



4. Ausreißer mit
starken Einfluss?

Ursachen und Folgen von Heteroskedastizität (Varianzinhomogenität)

Ursachen

- > Messfehler werden über die Zeit kleiner
- > Befragungen vor und nach Lernprozess
- > Verhalten abhängig vom Einkommen, so dass Reichere mehr Wahlmöglichkeiten haben als Ärmere
- > Bei aggregierten Werten sind Klassen mit kleinem n unsicherer und streuen mehr als Klassen mit grossem n

Tests:

- > Zerlegung der Daten und Vergleich von Subsets (z.B. Zeitperioden)
- > Goldfeld-Quandt-Test (univariat)
- > White-Test (multivariat)

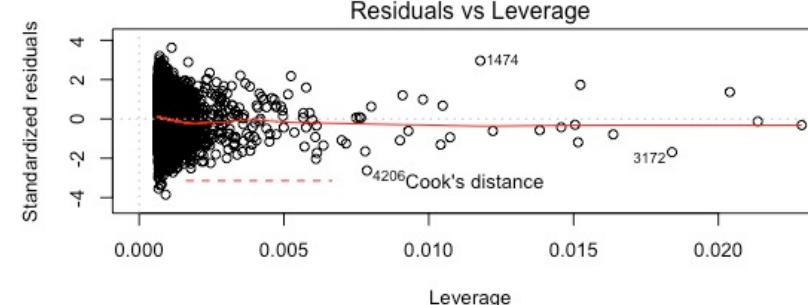
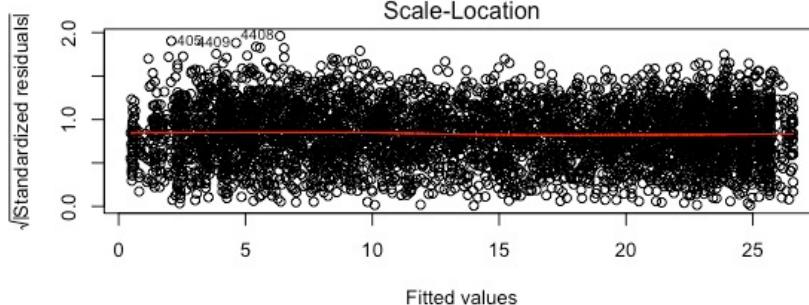
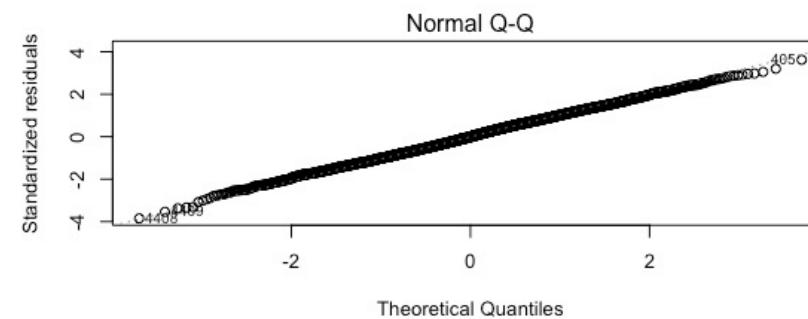
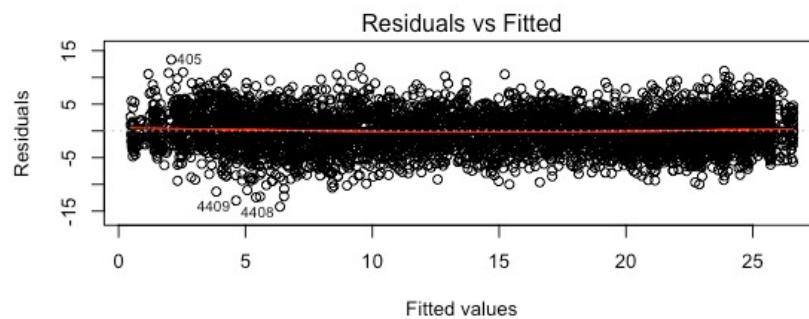
Behandlung:

- > Methode der kleinsten **gewichteten** Quadrate benutzen (weigthed least squares, WLS) Werte bekommen dort weniger Gewicht, wo die Streuung am grössten ist

Prüfung von Heteroskedastizität

Prüfung:

- > Residuenplot: Standardisierte Residuen werden gegen die standardisierten geschätzten y-Werte geplottet.

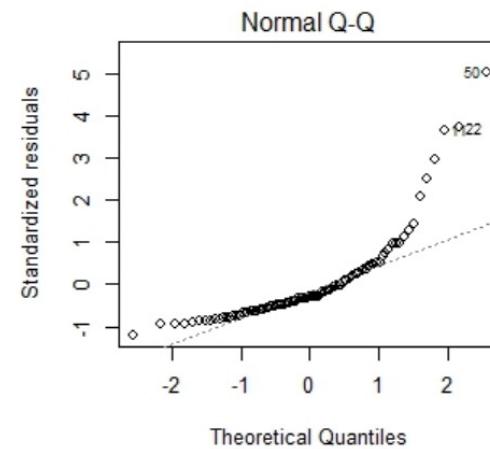
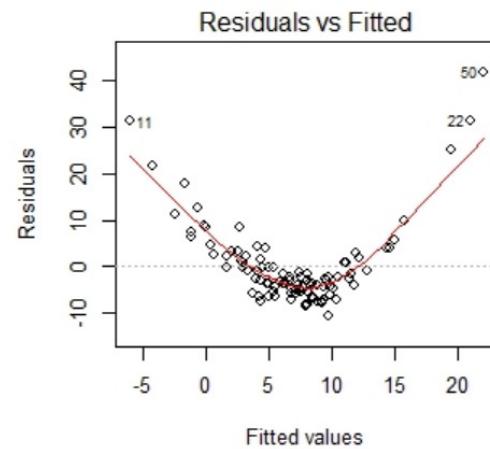


3. Varianzinhomogenität

Residuendiagnostik

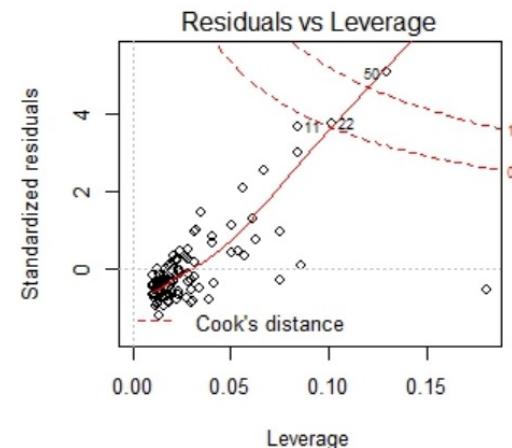
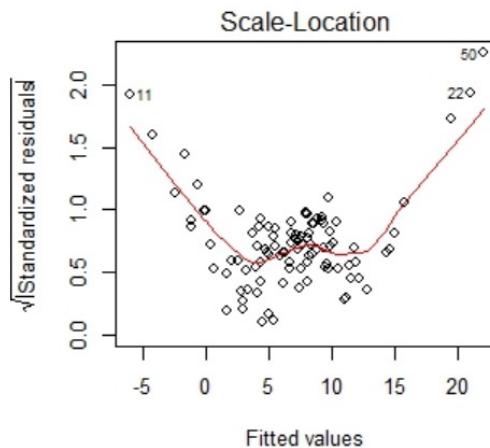
Anwendungsbedingungen NICHT erfüllt

1. Linearität



2. Normalverteilung der Residuen

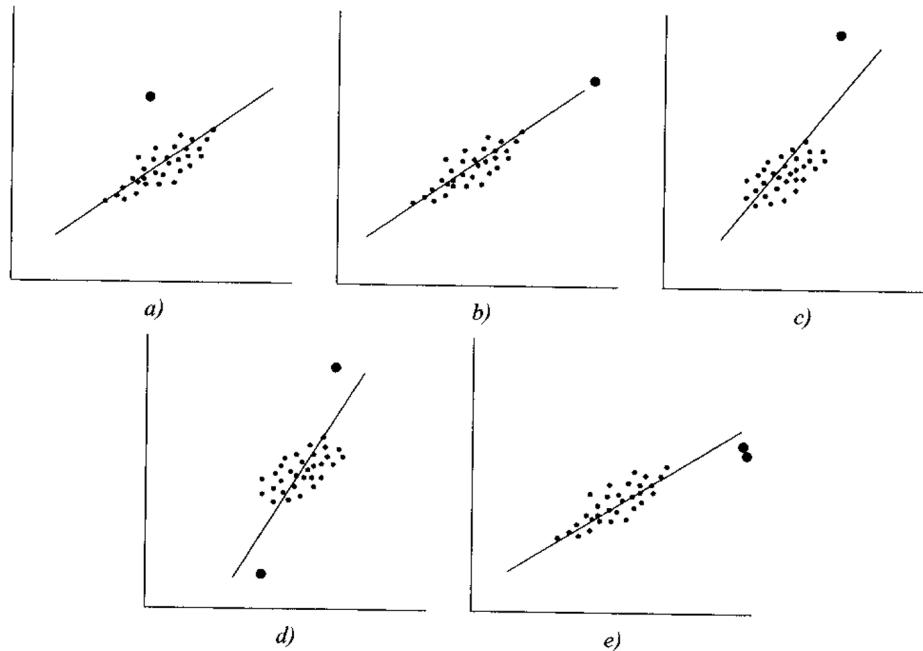
3. Varianzinhomogenität



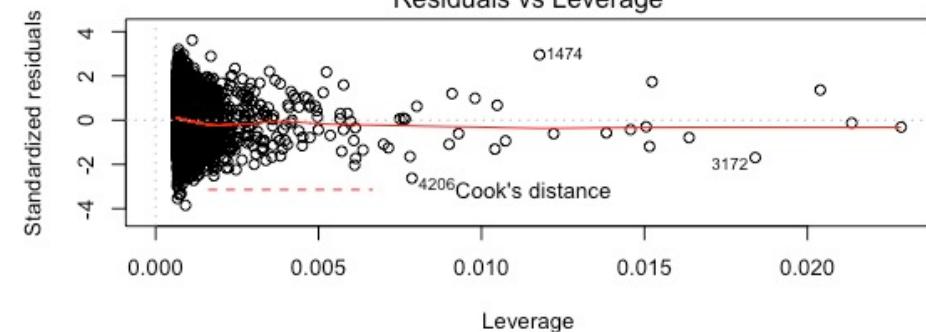
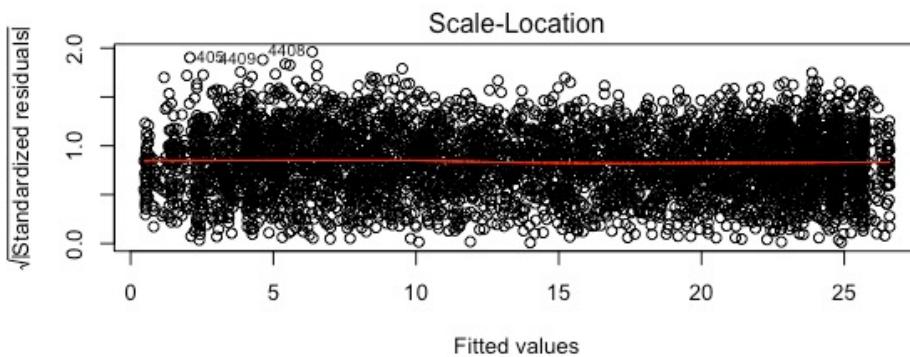
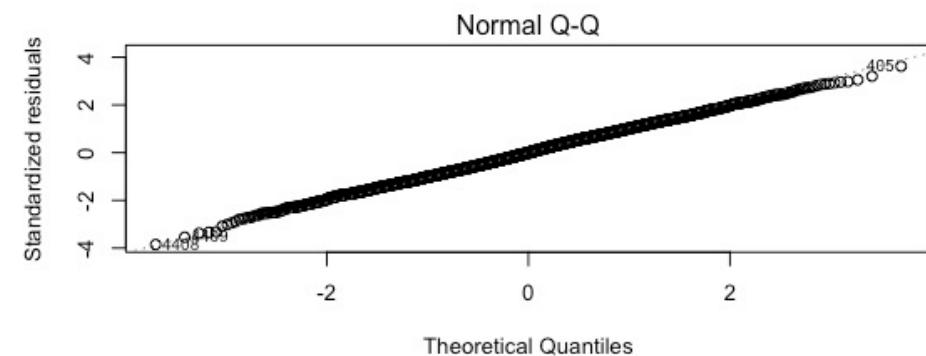
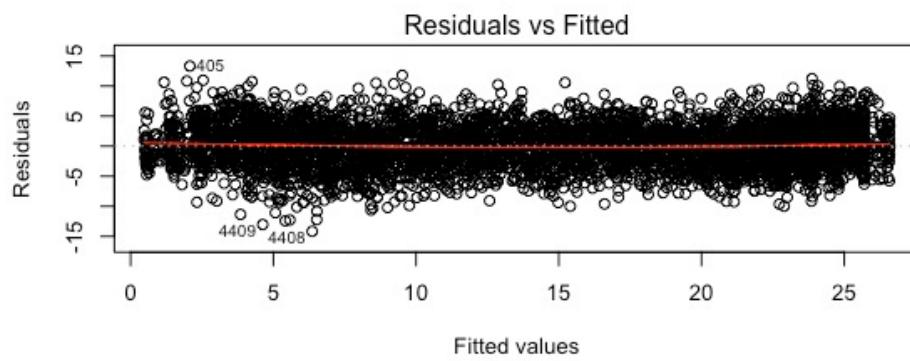
4. Ausreißer mit starken Einfluss?

Ausreißer

- > Regressionanalyse empfindlich genauso wie Korrelationsanalyse



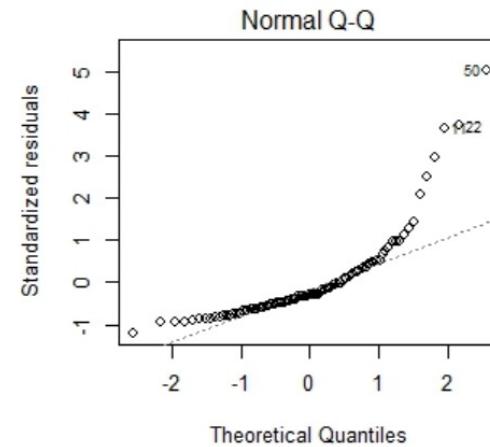
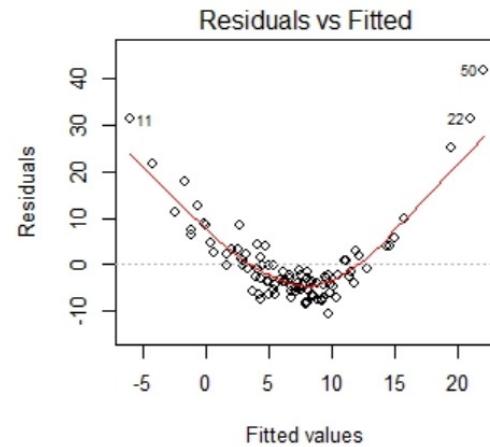
Ausreisser



Residuendiagnostik

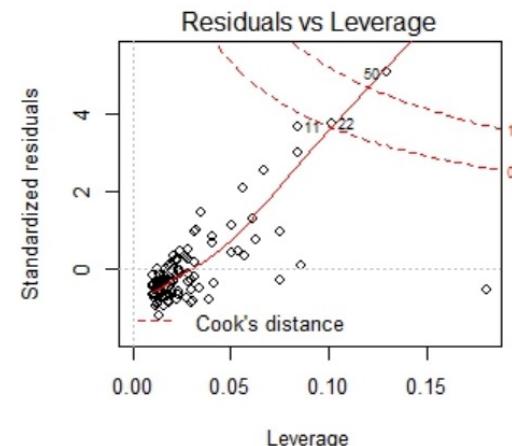
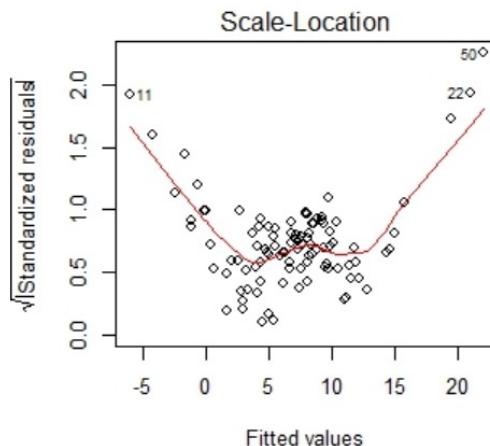
Anwendungsbedingungen NICHT erfüllt

1. Linearität



2. Normalverteilung der Residuen

3. Varianzinhomogenität



4. Ausreißer mit starken Einfluss?

(Multi-)Kollinearität

- > Die unabhängigen Variablen dürfen untereinander nicht perfekt korreliert sein, d.h. keine Linearkombinationen anderer Variablen
- > sonst ist Matrix nicht invertierbar
- > in Praxis meist nur annähernde Kollinearität

Folgen

- > Schätzungen der Regressionsparameter unzuverlässig
- > Standardfehler der Regressionskoeffizienten wird größer
- > Bei perfekter Multikollinearität ist keine Schätzung möglich

Anzeichen

- > Resultate werden stark von Weglassen einer Beobachtung beeinflusst
- > Vorzeichen der Regressionskoeffizienten ist anders als erwartet

Kollinearität

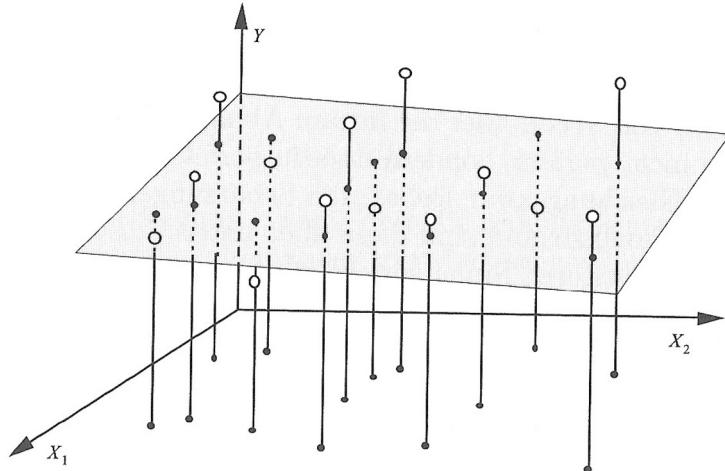


Abbildung 7.1: Beispiel für Situation ohne Kollinearität

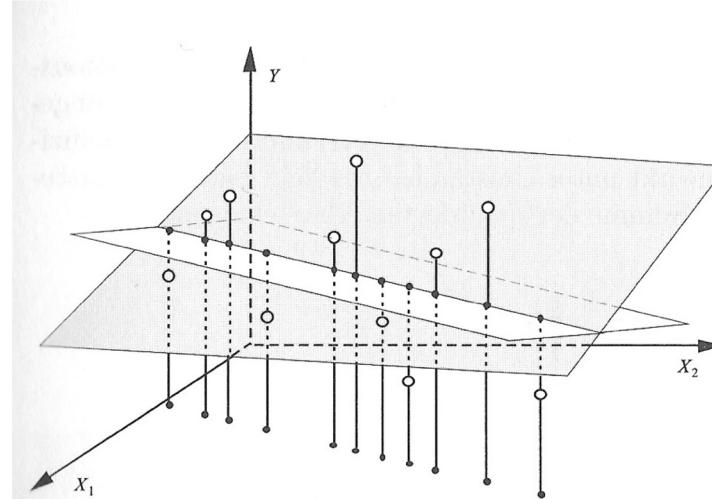


Abbildung 7.2: Beispiel für Situation mit perfekter Kollinearität

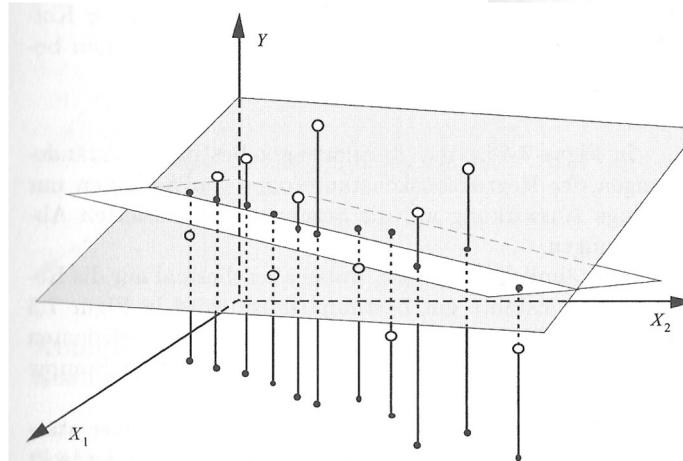


Abbildung 7.3: Beispiel für Situation mit starker, aber nicht perfekter Kollinearität

Prüfung auf Kollinearität

- > **Korrelationsmatrix:** Hohe Korrelationen der unabhängigen Variablen ($> \sim 0.8$ oder < -0.8) deuten auf mögliche Kolinearität hin.
- > **Varianzinflationsfaktor (VIF):** Misst Abhängigkeit der Varianz des geschätzten Regressionskoeffizienten aufgrund der Korrelation zwischen unabhängigen Variable. VIF-Werte über 10 gelten als kritisch.

Lösung:

- > **Variablen eliminieren:** Eine der hochkorrelierten Variablen aus dem Modell entfernen.
- > **Hauptkomponentenanalyse (PCA):** Reduktion der Dimensionalität der Daten durch PCA, um unkorrelierte Hauptkomponenten zu erstellen (nächste Woche!).

u^b

b
UNIVERSITÄT
BERN

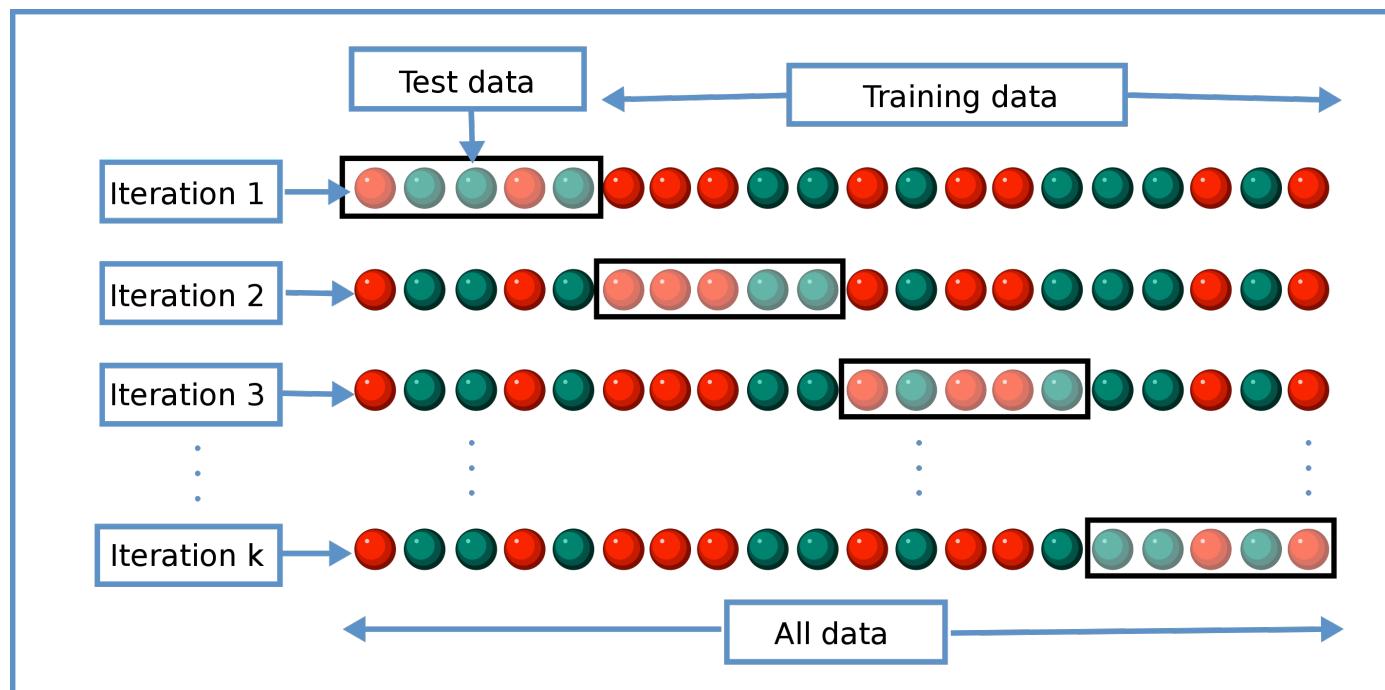
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

MODELLVALIDIERUNG

Modell-Validierung

Kreuzvalidierung:

- > Datensatz in mehrere Teile aufteilen.
- > Regressionsmodell mit einem Teil der Daten erstellen
- > Regressionsmodell mit anderem, unabhängigen Teil der Daten testen (z.B. Korrelationskoeffizient, mittlerer quadratischer Fehler)

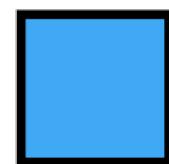


Modell-Validierung

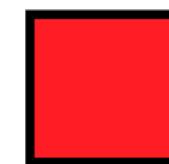
Leave-one-out Kreuzvalidierung:

- > Für kleine Datensätze
- > Nacheinander jeweils einen anderen Wert bei der Modellerstellung auslassen
- > Jeweils testen wie gut sich der ausgelassene Wert vorhersagen lässt

$n = 8$

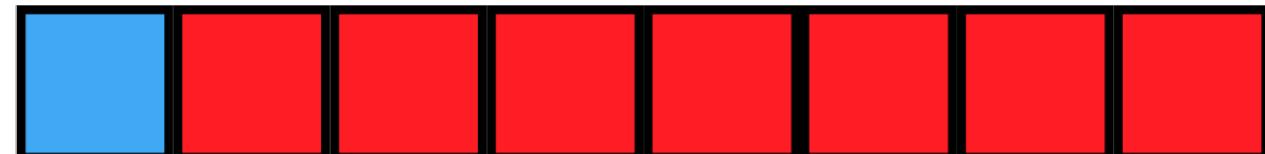


Test



Train

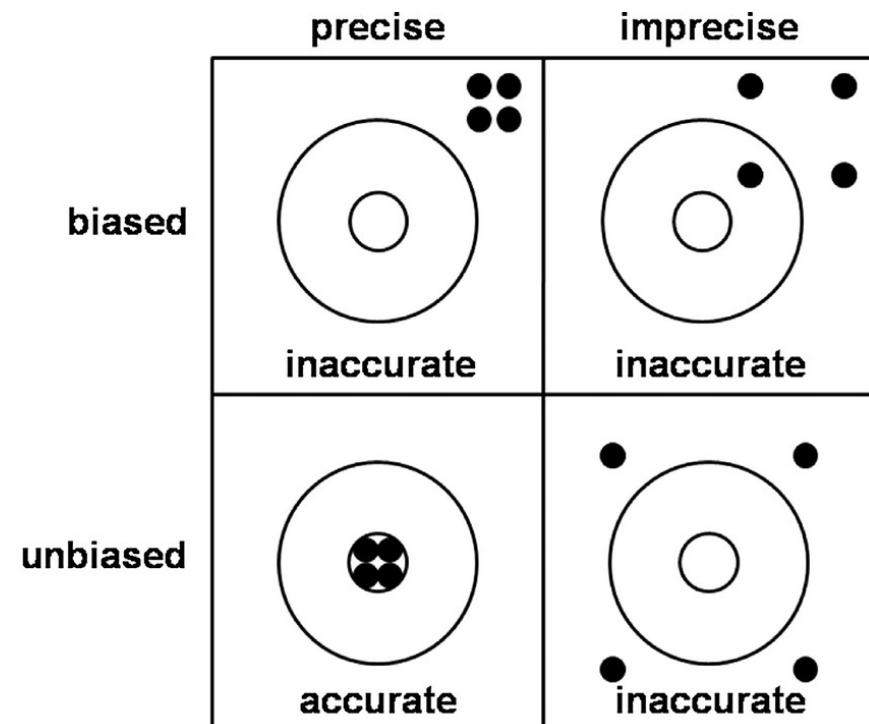
Model 1

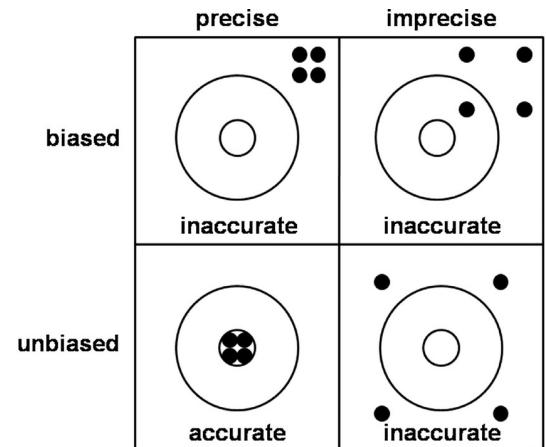


Bias

Bias (mittlerer oder systematischer Fehler):

- > $Bias_{additiv} = \overline{\text{Vorhersage}} - \overline{\text{Beobachtung}}$
- > $Bias_{multiplikativ} = \overline{\text{Vorhersage}} / \overline{\text{Beobachtung}}$





Einfache Fehlermasse

Mittlerer absoluter Fehler/Error (MAE)

> $MAE = \frac{1}{N} \sum_{i=1}^N |Vorhersage_i - Beobachtung_i|$

Mittlerer quadratischer/squared Fehler/Error (MSE)

- > $MSE = \frac{1}{N} \sum_{i=1}^N (Vorhersage_i - Beobachtung_i)^2$
- > Oft wird noch die Wurzel gezogen, um wieder gleiche Einheiten zu haben: **Root Mean Square Error (RMSE)**
- > $RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Vorhersage_i - Beobachtung_i)^2}$
- > Hier bekommen grosse Abweichungen mehr Gewicht als beim MAE

Skill Score (SS)

Vergleich mit einer Referenzvorhersage

$$MSE_{SS} = 1 - \frac{MSE_{forecast}}{MSE_{ref}}$$

- > In der Meteorologie oder Hydrologie ist die Referenz oft die Klimatologie, also der Mittelwert der Beobachtungen
- > Wertebereich: $\infty < SS < 1$
- > Perfekte Vorhersage: $SS = 1$
- > Vorhersage so gut wie Annahme des Mittelwerts: $SS = 0$
- > Auch *Nash-Sutcliffe Efficiency* genannt in der Hydrology

Take-home messages

- > Residuen des Regressionsmodells mindestens visuell prüfen
 - Linearität
 - Normalverteilung
 - Homoskedastizität
 - Gewicht von Ausreisern
- > **Validierung mit unabhängigen Daten!**

Grösste Fehler, die es zu vermeiden gilt

1. Zusammenhang ist nicht linear
2. Korrelation bedeutet nicht Kausalzusammenhang
3. Wichtige erklärende Variable im multiplen Regressionsmodell vergessen
4. Hoch miteinander korrelierte Variablen im Regressionsmodell
5. Extrapolation weit ausserhalb der Daten, auf denen das Modell beruht
6. Modell nicht mit unabhängigen Daten auf Allgemeingültigkeit getestet

Beispiel Prüfungsfragen

- > Welche Aussage/n zur Variablenauswahl bei der multiplen Regression ist/sind korrekt?
 - Es sollten möglichst viele Variablen genutzt werden, um jeden Information maximal auszunutzen
 - Es sollten nur verfügbaren Variablen genutzt werden, bei denen ein Kausalzusammenhang auch theoretisch erklärbar ist
 - Es sollte nur die unabhängige Variable genutzt werden, die am meisten erklärt
- > Welche Aussage/n zum Bestimmtheitsmass ist/sind korrekt?
 - bei der multiplen Regression muss das Bestimmtheitsmass und die Anzahl unabhängiger Variablen angepasst werden.
 - bei der multiplen Regression muss das Bestimmtheitsmass und die Anzahl abhängiger Variablen angepasst werden
 - bei der multiplen Regression muss das Bestimmtheitsmass und die Stichprobengrösse angepasst werden
- > Interpretiert die Ausgabe der Zusammenfassung einer multiplen Regression mit der lm() Funktion in R (siehe Folie oben und Übung)

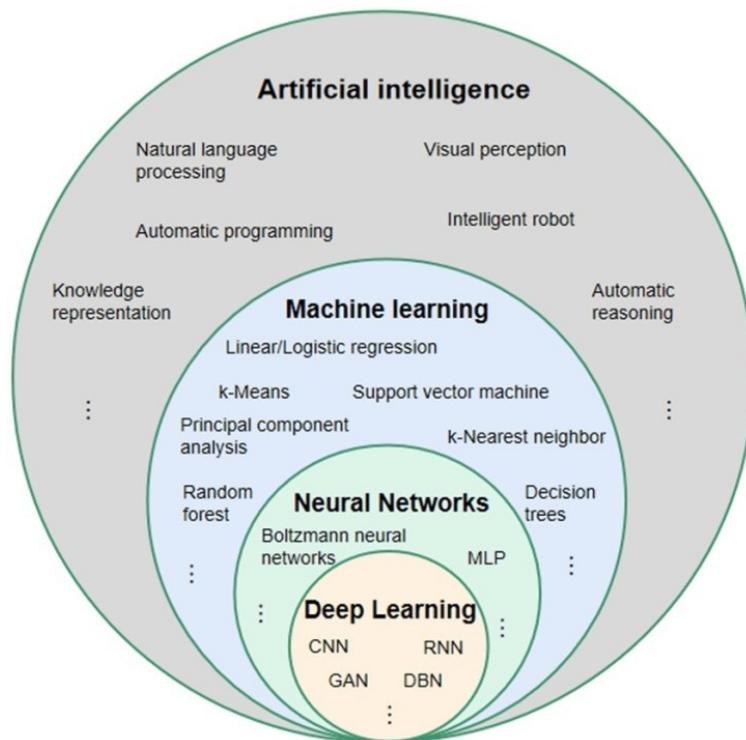
Beispiele Prüfungsfragen

- > Welches sind Bedingungen, damit die lineare Regression angewendet werden kann?
 - unabhängige Variablen müssen normalverteilt sind
 - abhängige Variablen müssen normalverteilt sind
 - Residuen müssen normalverteilt sind
- > Welches Mass ist weniger von Ausreissern beeinflusst?
 - Mittlerer quadratischer Fehler
 - Mittlerer absoluter Fehler
- > Welche Schritte gehören zu einer Regressionsanalyse?
 - Modell nicht mit unabhängigen Daten testen
 - Regressionmodell mit allen verfügbaren Daten erstellen
 - alle verfügbaren Variablen ins Regressionsmodell, auch hoch miteinander korrelierte

Daten Zusammenfassen

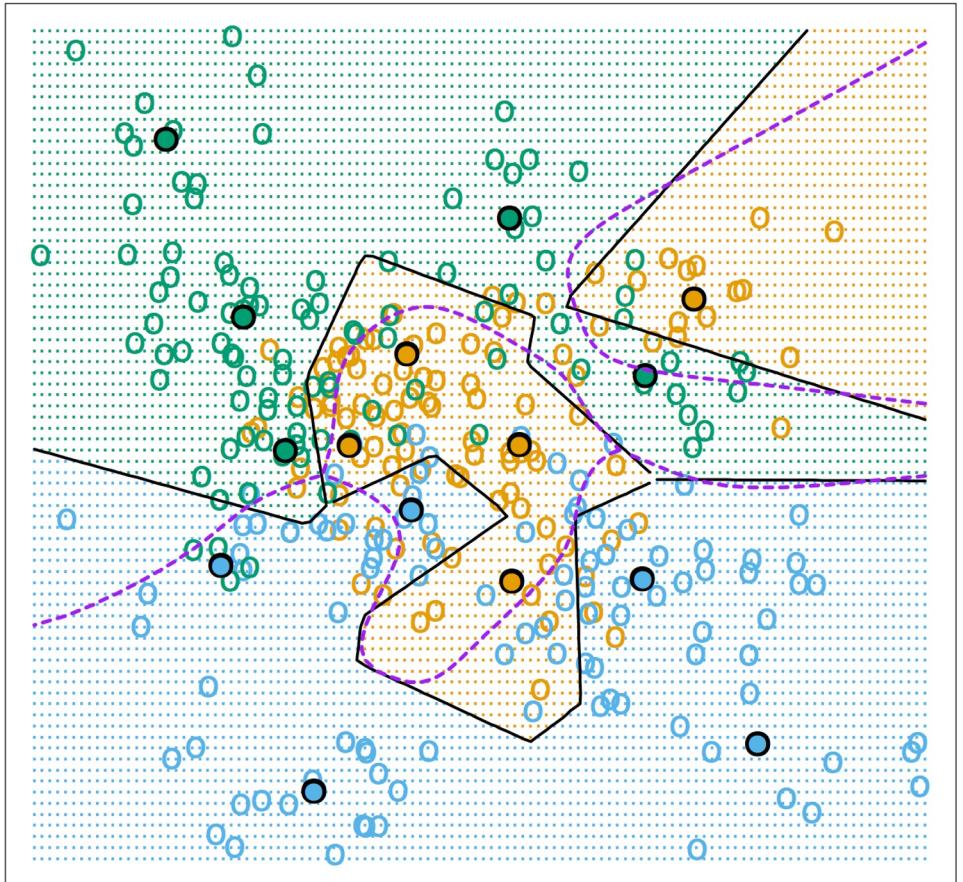
Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen, Modellvalidierung
weiterführende Methoden	Daten zusammenfassen	Extremwertstatistik	
Fallen der Statistik	Hauptkomponentenanalyse	Clusteranalyse	Neural Networks

“Künstliche Intelligenz”



- **Künstliche Intelligenz**: jede Technik, die menschliches Verhalten oder menschliche Entscheidungsfindung nachahmt
- **Machine Learning**: ein Teilbereich der KI. Verwendet statistische Methoden, um aus Daten zu lernen.
- **Neuronale Netze** sind das Herzstück der aktuellen KI-Revolution
- **Deep Learning**: ein Teilbereich des maschinellen Lernens, der komplexe NN-Architekturen (CNNs, RNNs, ...) umfasst

Clusteranalyse

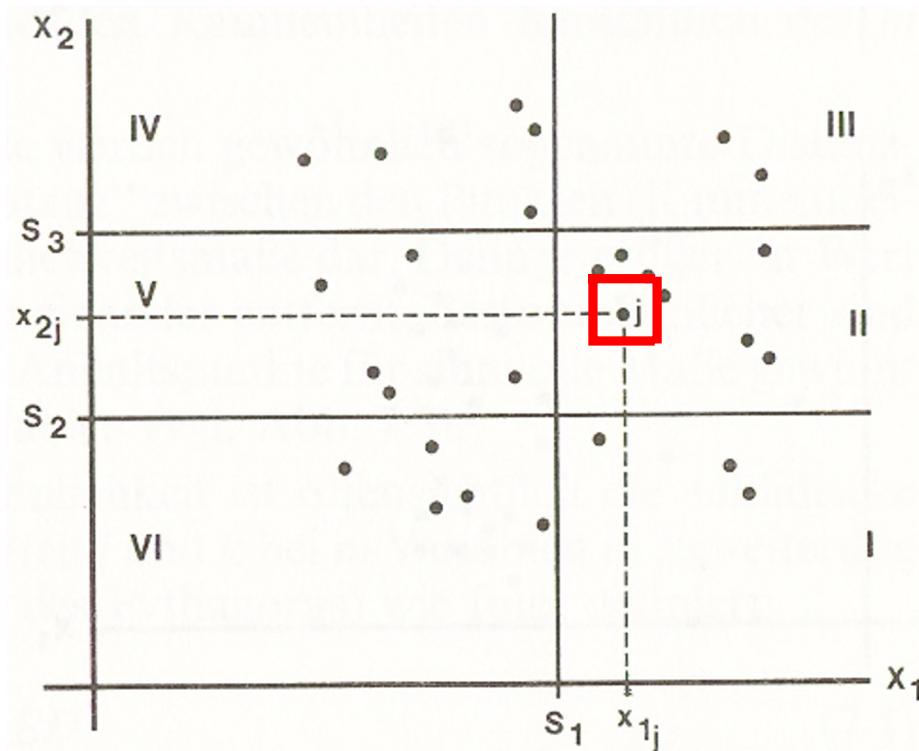


Klassifiziert Bezugs/Raumeinheiten auf der Grundlage von Ähnlichkeitsmassen, z.B. Klimatypen, Stadttypen, ...

- > Zusammenfassung/Gruppierung von ähnlichen Bezugseinheiten
- > Eindeutige Zuordnung jeder Bezugseinheit zu einem Typ/Cluster

Clusteranalyse

Einfachster Ansatz zur Klassifizierung ist die **Schwellenwertmethode**:



Nachteil:

- > setzt theoretische oder empirische Kenntnisse voraus
- > subjektive Schwellenwerte

Clusteranalyse

... automatisches, d.h. nicht subjektives Klassifikationsverfahren

- > Ähnlichkeit nur über die Lage zueinander im durch die Variablen aufgespannten m-dimensionalen Koordinatensystem definiert

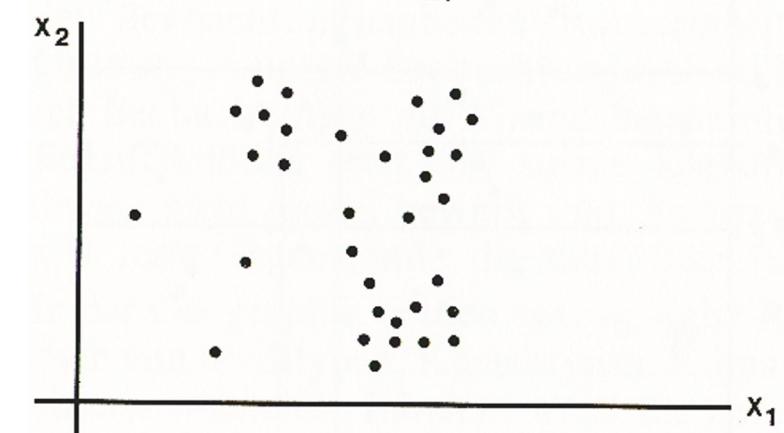
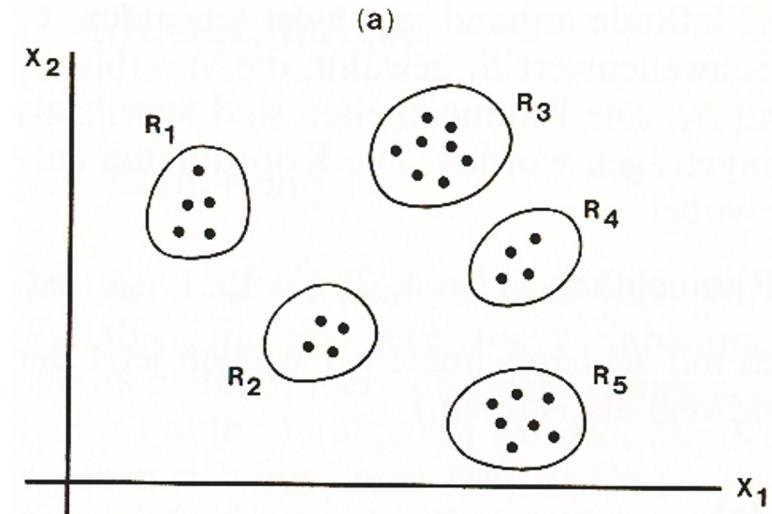
Fall 1 (oben):

- > deutliche Cluster-Bildung, die bereits visuell vorgenommen werden kann

Fall 2 (unten):

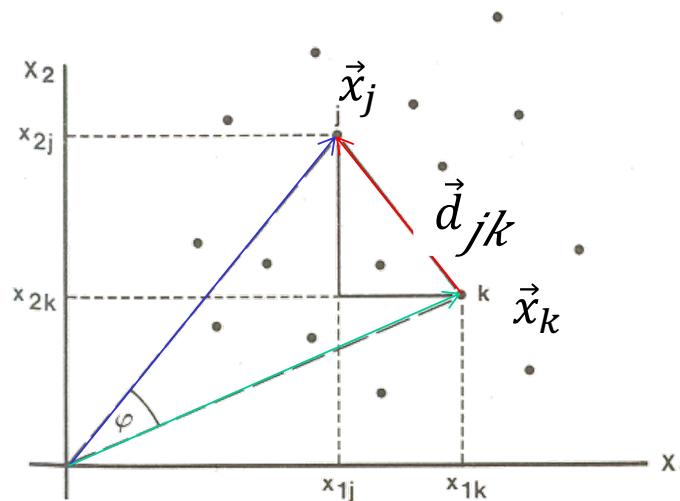
- > unstrukturierte Konstellation
- > weder Zuordnung noch Anzahl der Cluster visuell abzuschätzen
- > häufiger Fall Praxis
- > formales Verfahren gefordert, dass auch im höher dimensionalen Raum mit mehr als 2 Variablen angewendet werden kann

Beispiel mit 2 Variablen



Ähnlichkeitsmasse

$$|\vec{d}_{jk}| = \sqrt{\sum_{i=1}^m (x_{ji} - x_{ki})^2} = \text{Euklidische Distanz}$$



Skalieren

Wenn Variablen nicht skaliert sind,

- > hat die Variable mit grösster Spannweite das meiste Gewicht,
- > hängt die Distanz von der Skala ab

Skalieren gibt den Variablen gleiches Gewicht, daher skalieren wenn:

- > Variablen in unterschiedlichen Einheiten vorliegen (km, m, kg, ...)
- > man explizit gleiches Gewicht für alle Variablen möchte

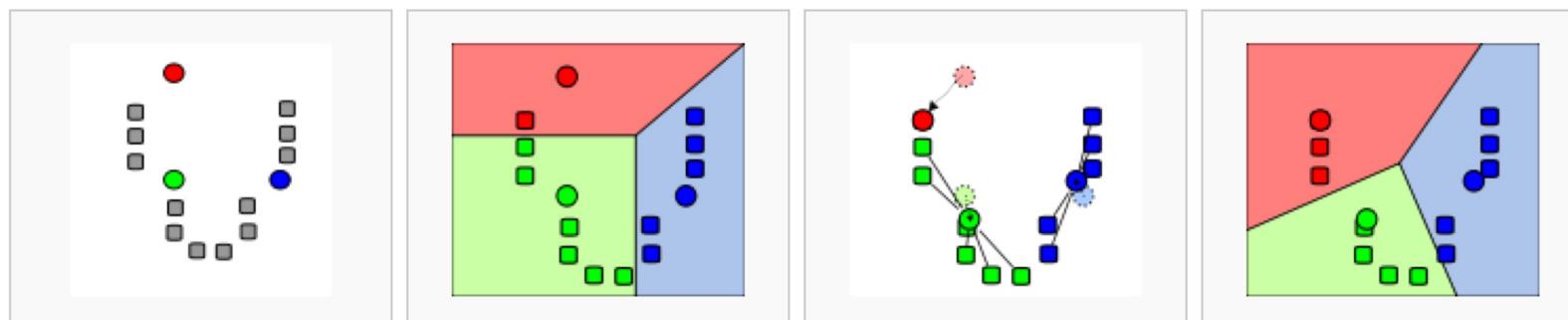
Nicht skalieren, wenn:

- > alle Variablen gleiche Einheiten haben.

NICHT-hierachisches Clustering

Beispiel: k-Means

- 1. Initialisierung:** k Clusterzentren auf k zufällige, aber unterschiedliche Positionen im p-dimensionalen Raum setzen, möglichst so dass die Abstände zwischen initialen Clusterzentren maximal sind. Jedem Clusterzentrum wird eine eindeutige Klassennummer (1 bis k) zugewiesen.
- 2. Klassifizierung:** Finde für jeden Datenpunkt das nächste Clusterzentrum und weise dem Datenpunkt die Klassennummer dieses Clusterzentrums zu.
- 3. Clusterzentren berechnen:** Berechne die Position der Clusterzentren neu, in dem alle Datenpunkte die zu einer bestimmten Klasse gehören gemittelt werden.
- 4. Iteration:** Wiederholung ab Schritt 2, bis die Klassifizierung stabil ist



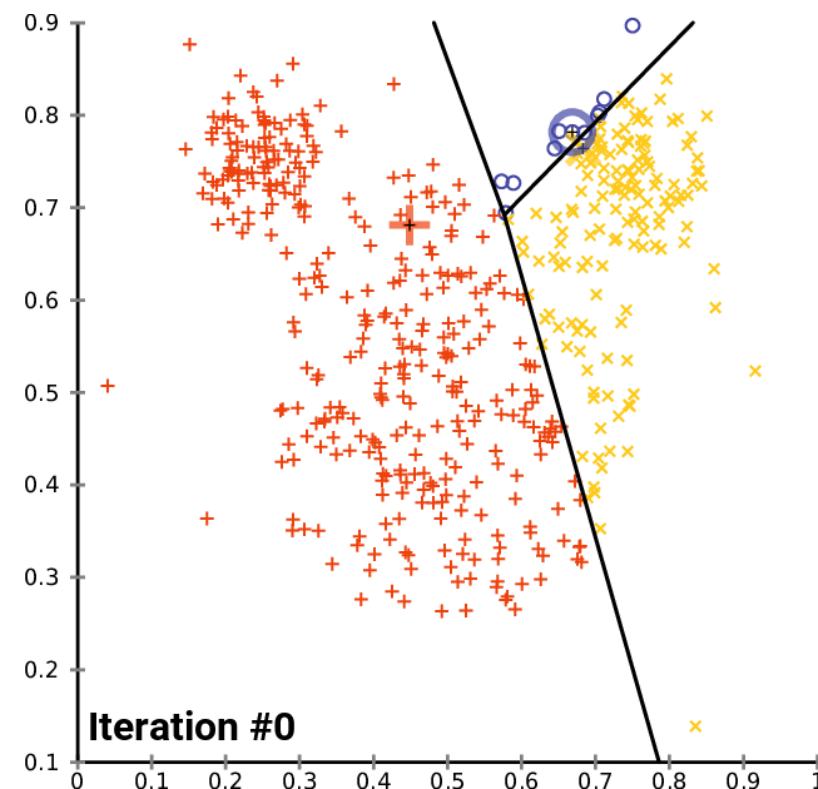
K-Means

Form des Machine Learning

u^b

^b
UNIVERSITÄT
BERN

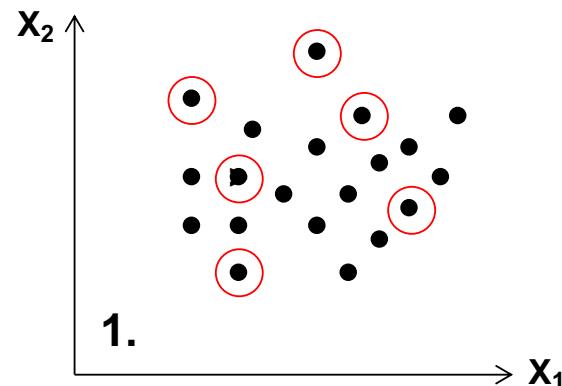
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



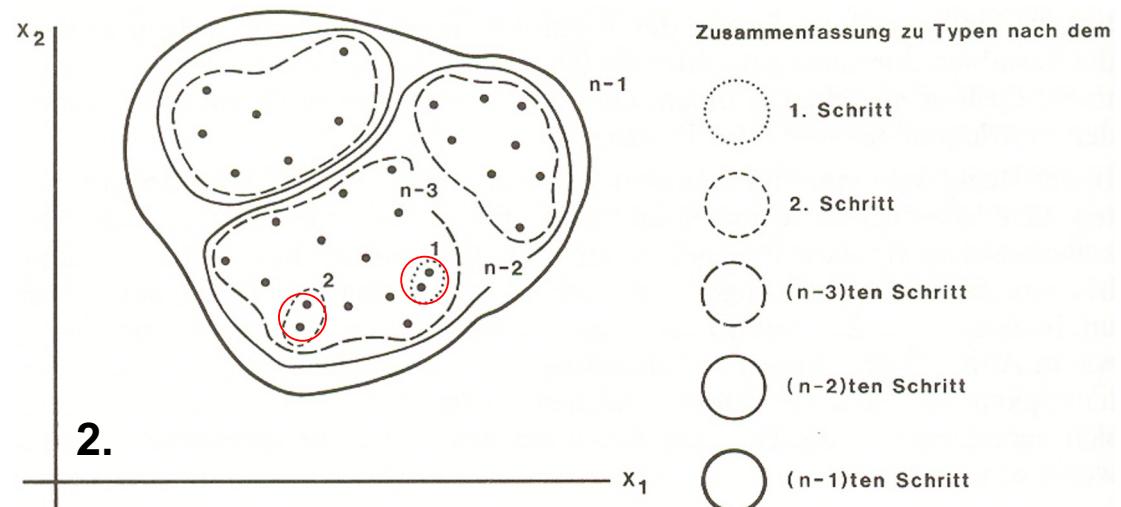
Hierarchische Cluster-Bildung

Iteratives Verfahren der Cluster-Bildung:

1. Ausgangspunkt sind $j=1..n$ Bezugseinheiten, die alle einem von n Clustern zugeordnet werden



2. im ersten Schritt werden die zwei Bezugseinheiten zu einem gemeinsamen Cluster zusammengefasst, die sich gemäss eines Ähnlichkeitsmaßes am allernächsten sind

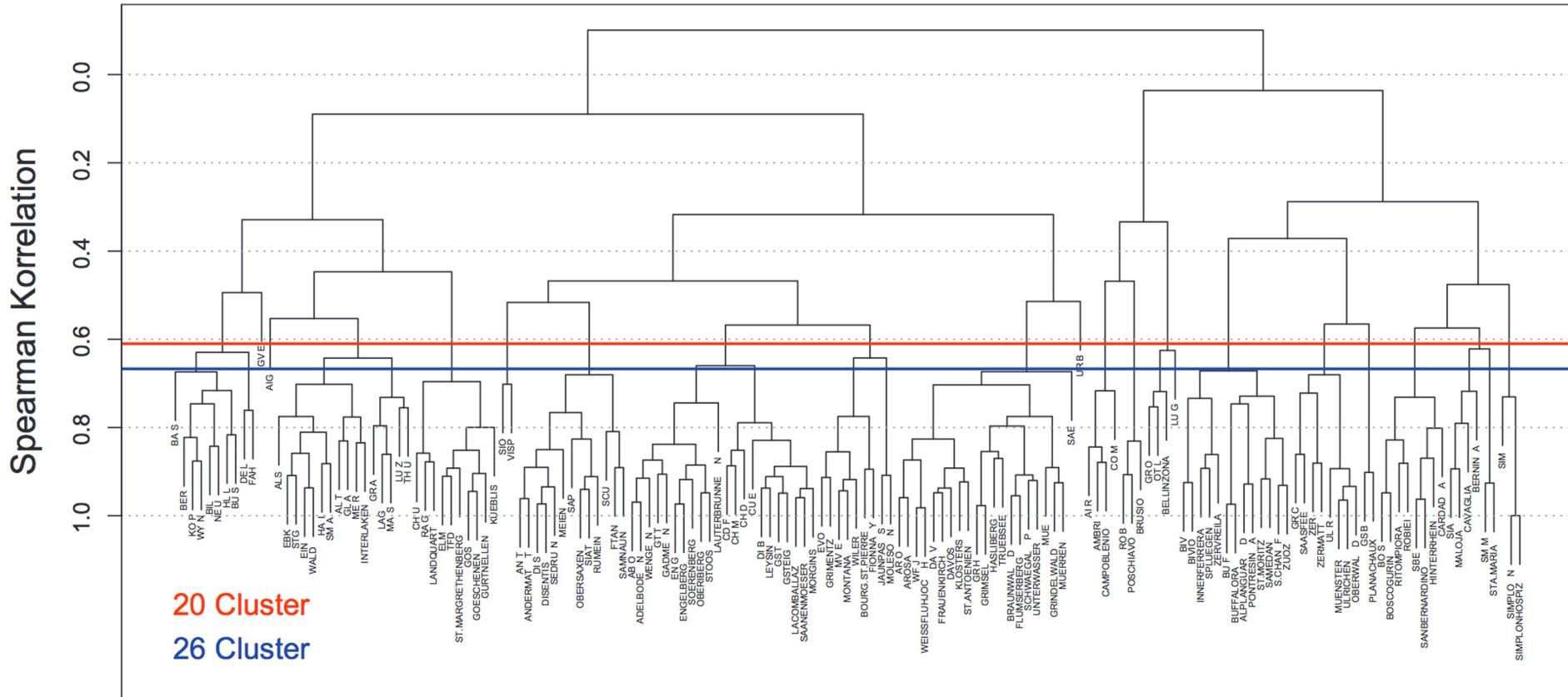


Dendrogramm Neuschnee

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

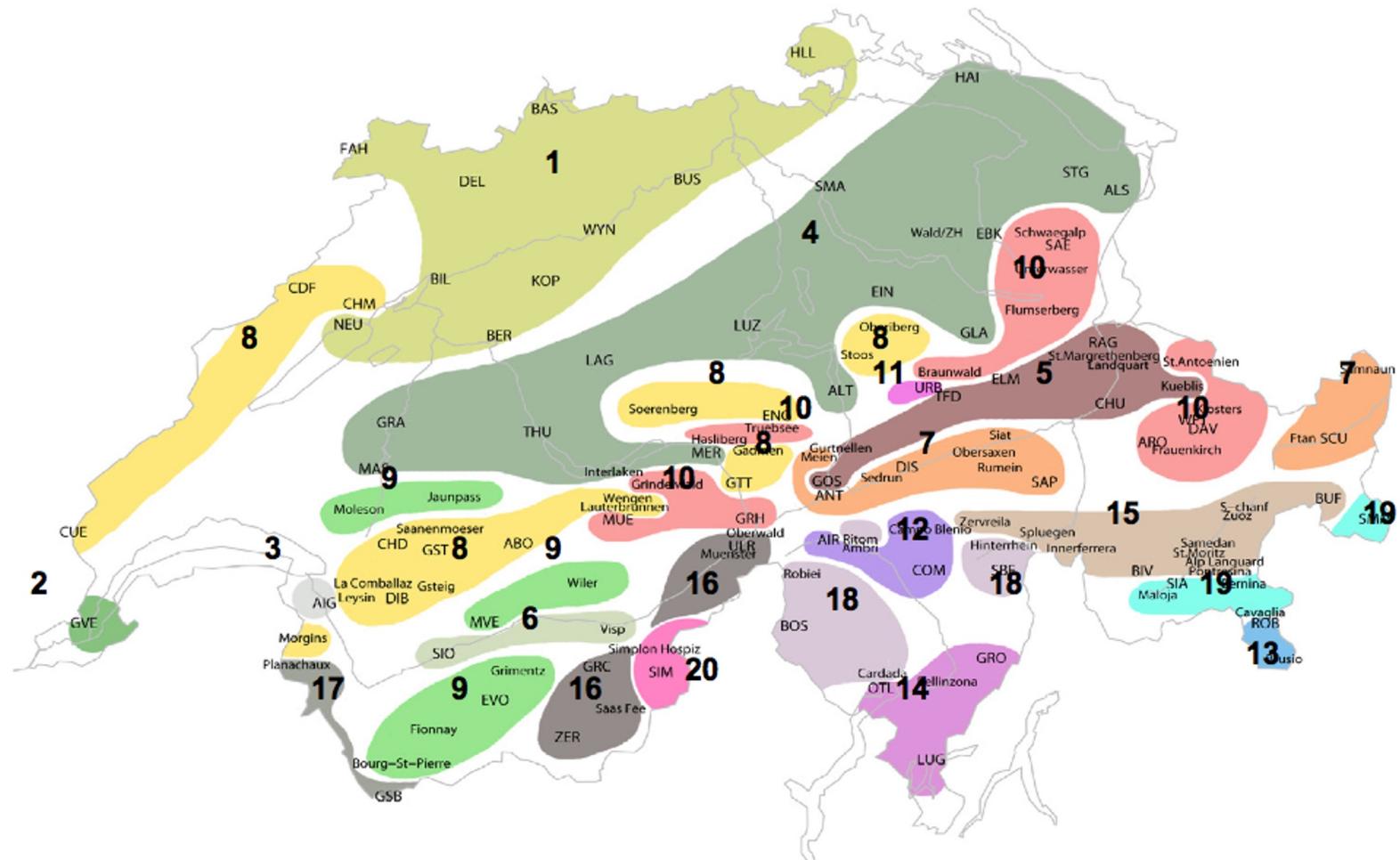


Clusteranalyse Neuschnee

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



1. nicht-hierarchische Verfahren:

- > Werte so lange immer wieder auf die Cluster verteilt, bis die Summe der Abstände von zugehörigen Cluster-Mitten minimal ist
- > Vorteil: Werte werden flexibel auf Cluster verteilt
- > Nachteil: Anzahl der Cluster muss a-priori vorgegeben werden

2. hierarchische Verfahren:

- > iterative Vorgehensweise, bei der die Cluster des letzten Schrittes immer weiter zusammengefasst werden
- > Nachteil: Einmal klassifizierte Werte verbleiben in dem Cluster auch wenn sich die Eigenschaften des Cluster im Laufe der Verfahrensschritte verändern
- > Vorteil: Anzahl der Cluster muss nicht a-priori vorgegeben werden

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Principal Component Analysis (PCA)

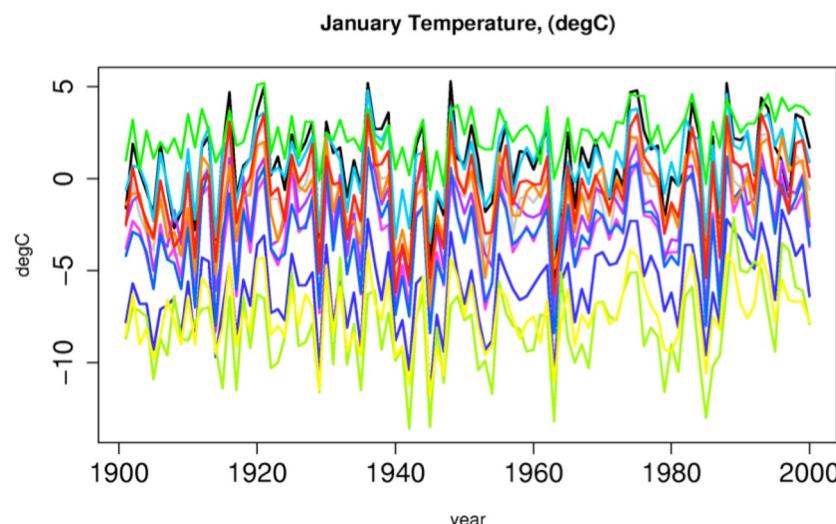
HAUPTKOMPONENTENANALYSE

Motivation

Ziel ist eine Dimensionsreduktion, d.h. „Relevante“ Informationen mit einem Redundanz- und Rauschfilter zu extrahieren

Es gibt viele **redundante** Informationen, z.B. weil Messungen benachbarter Stationen korreliert sind

Die Daten enthalten nicht nur das Messsignal sondern auch Fehler

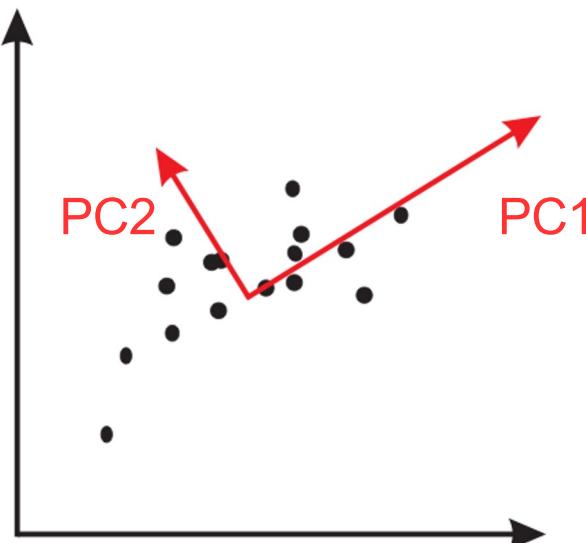


Basel, Bern, Chateaux d'Oex, Chaumont,
Davos, Engelberg, Geneve, Lugano,
Säntis, Segl Maria, Sion, Zürich

Hauptkomponentenanalyse

- > Transformation der Datenmatrix
- > $Y^T = X^T W$

- > so dass das Ergebnis die Varianz in absteigender Reihenfolge der PCs (Principal Components) maximiert.



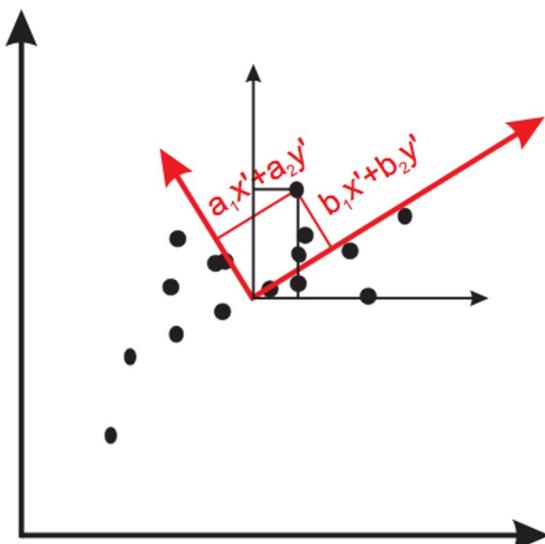
Hauptkomponentenanalyse

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Transformation der Datenmatrix
- > $Y^T = X^T W$
- > so dass das Ergebnis die Varianz in absteigender Reihenfolge der PCs maximiert.



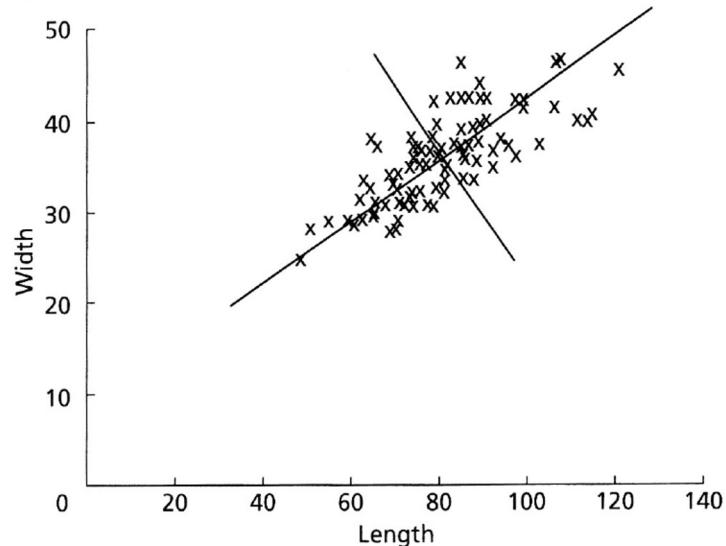
Original: (x, y)
Standardisiert: (x', y')
Transformiert: $(a_1x' + a_2y', b_1x' + b_2y')$

Hauptkomponentenanalyse

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



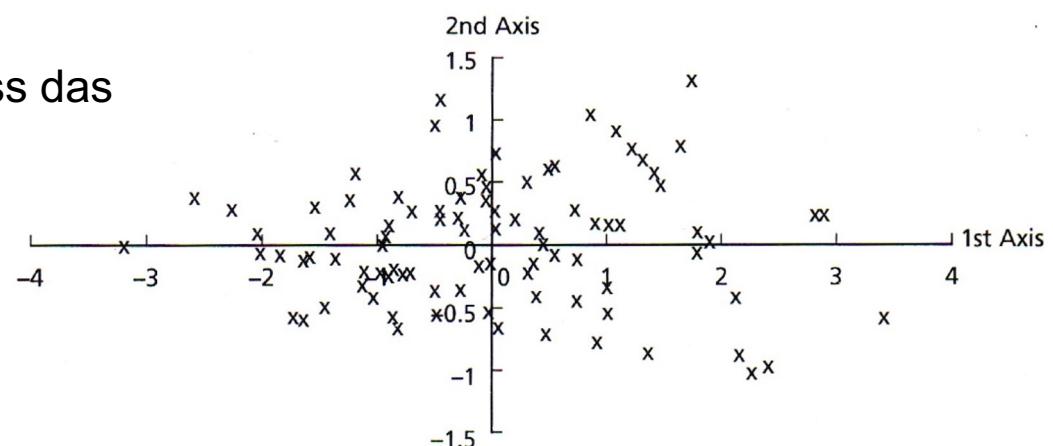
Einfachster Fall:

- > 2 Variablen (**Länge und Breite**)
- > Neues Koordinatensystem mit Nullpunkt in Mittelpunkt von x und y

Transformation der Datenmatrix, so dass das Ergebnis die Varianz in absteigender Reihenfolge der PCs maximiert:

1. Achse ???

2. Achse ???

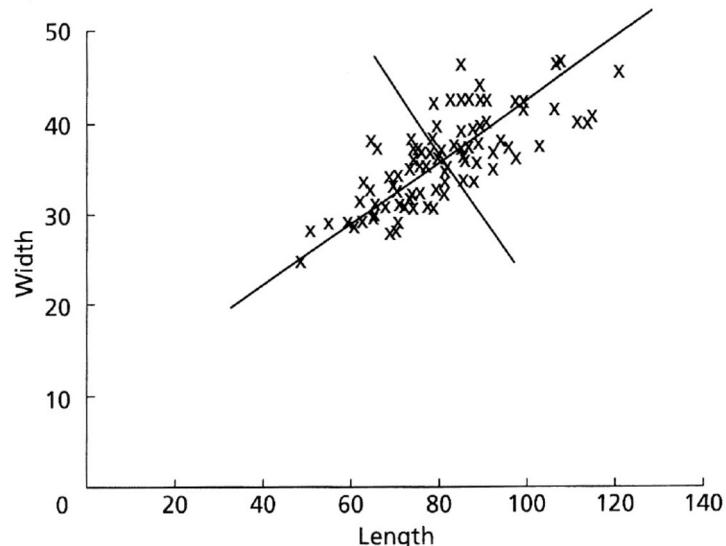


Hauptkomponentenanalyse

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

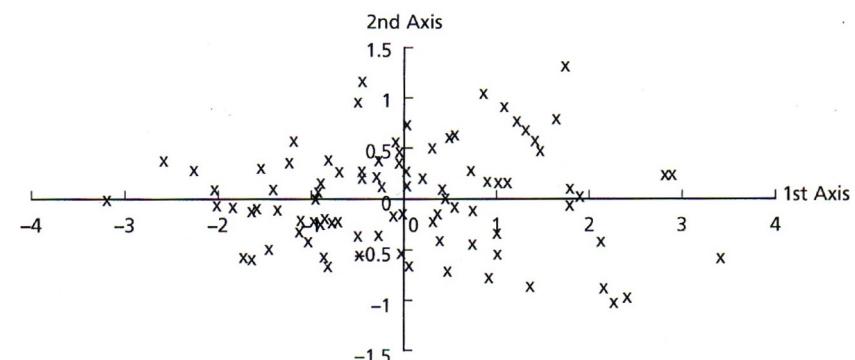


Einfachster Fall:

- > 2 Variablen (**Länge und Breite**)
- > Neues Koordinatensystem mit Nullpunkt in Mittelpunkt von x und y

1. Achse könnte man “Grösse” nennen: kleine Länge und kleine Breite links

2. Achse könnte “Form” genannt werden: Je weiter von der 1. Achse entfernt, desto mehr weicht das Verhältnis von Länge zu Breite vom Stichprobenmittel ab



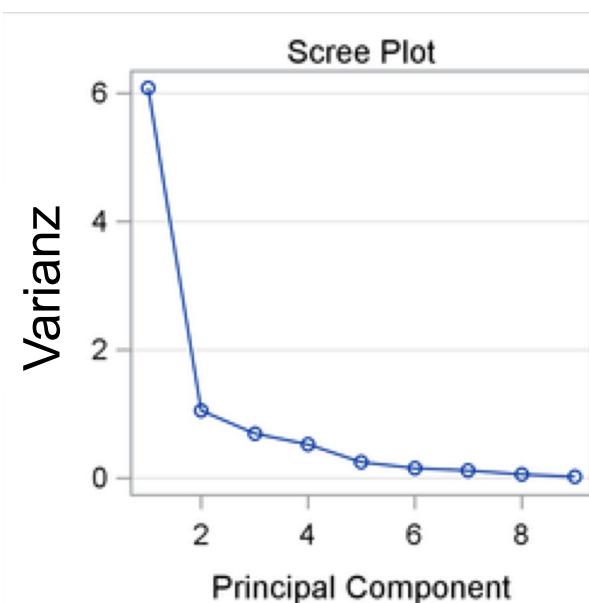
Wie viele Hauptkomponenten soll man betrachten?

U^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- Um die gesamte Variabilität des Systems zu reproduzieren braucht man auch alle Hauptkomponenten
- Aber: Die ersten wenigen Hauptkomponenten erklären oft einen Hauptanteil der Variabilität
- Damit ist unser Ziel der Reduktion der Dimensionen erreicht



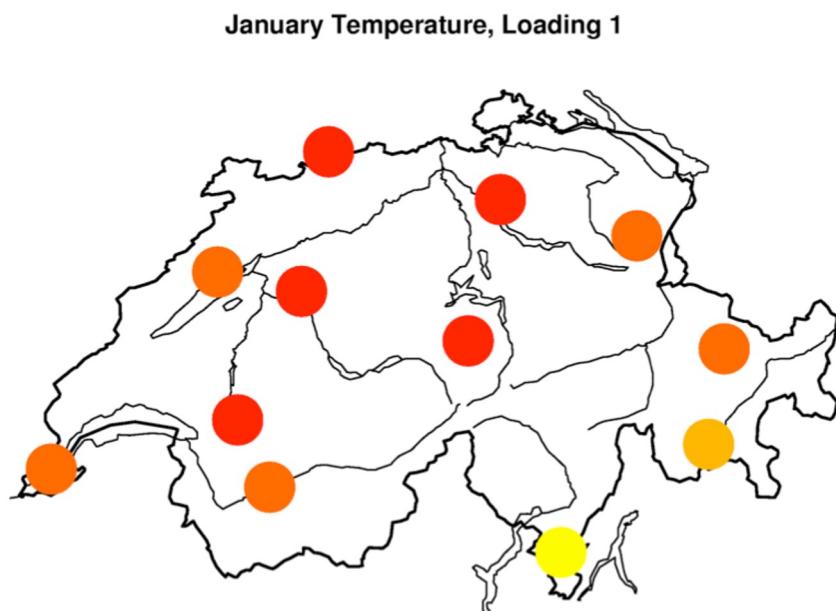
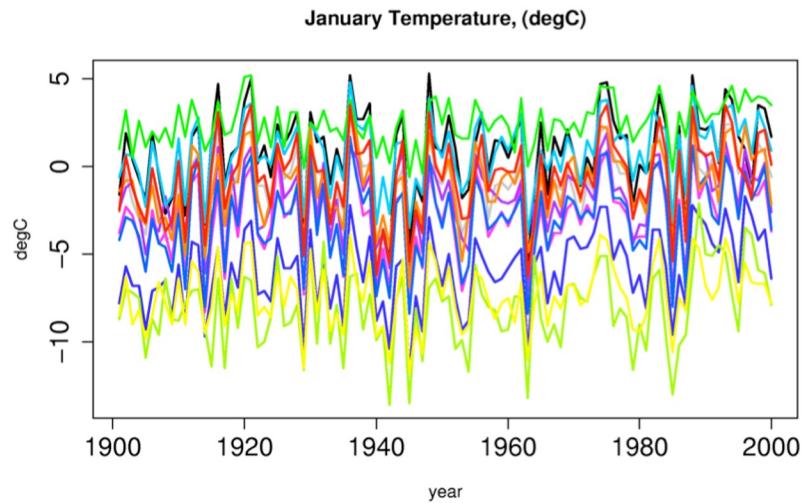
Daumenregel:
PCs bis zum “Knie” sind bedeutend

Ladungen

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Basel, Bern, Chateaux
d'Oex, Chaumont, Davos,
Engelberg, Geneve,
Lugano, Säntis, Segl Maria,
Sion, Zürich

Ladungen

- > sind die Korrelationskoeffizienten zwischen den PCs und den Originalvariablen
- > messen die Wichtigkeit jeder Variablen für die Varianz der jeweiligen PC.
- > Oft kann man an den Ladungen Gruppen erkennen
- > mit Hilfe der Ladungen kann man wichtige Variablen von solchen trennen, die nicht viel zur Erklärung der Varianz des Datensatzes beitragen

Probleme der Hauptkomponentenanalyse

Nicht-physikalisch

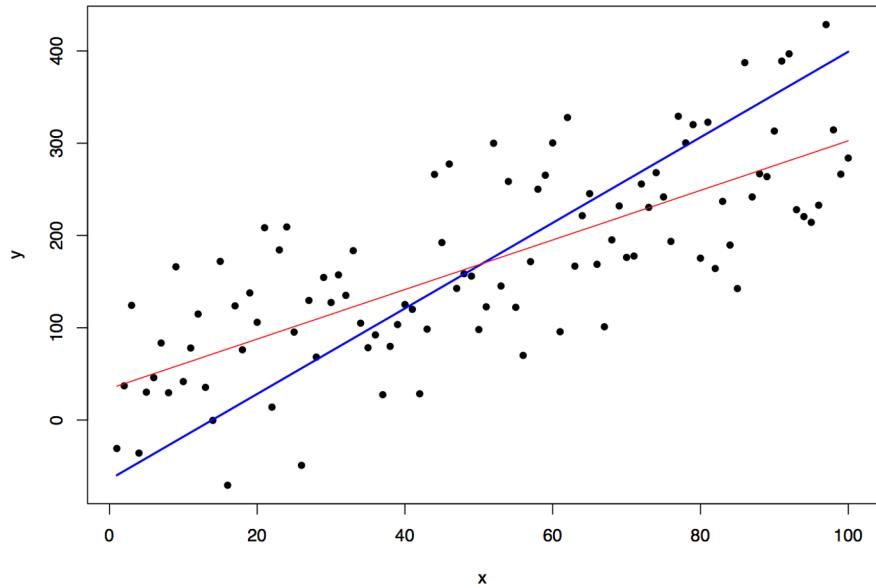
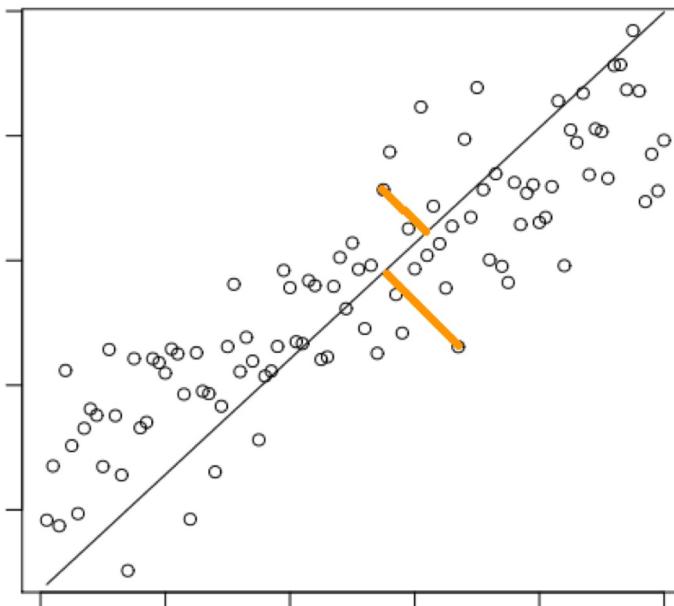
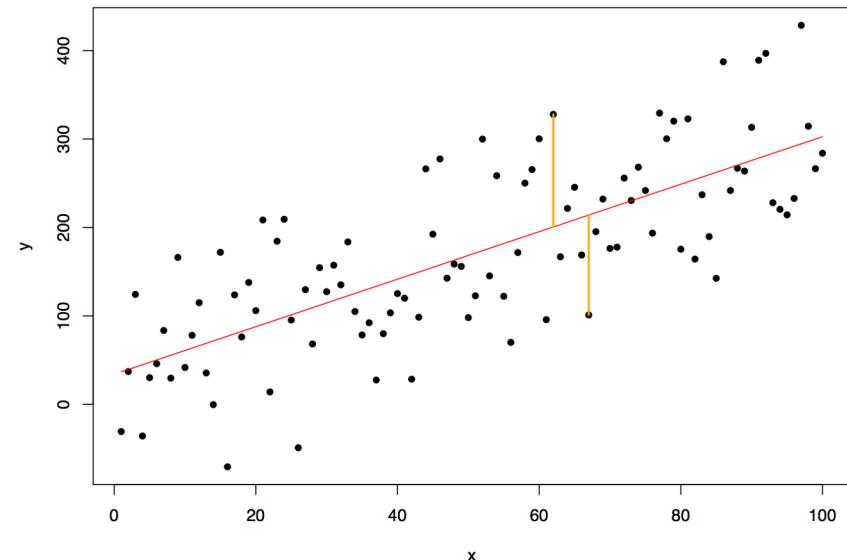
- > Orthogonalität
- > sequentielle Maximierung der Varianz
- > daher nicht immer physikalisch sinnvoll
- > **Lösung:** Vorsicht bei der Interpretation

Unregelmässige Beprobung, z.B. im Raum

- > Regionen mit mehr Daten sind im PC überrepräsentiert
- > **Lösung:** Daten vor der PC-Analyse interpolieren

ACHTUNG: Vorzeichen der Ladungen arbiträr

Regression vs. PCA



Erste Hauptkomponente (blau)
und Regression (rot)
unterscheiden sich.

Zusammenfassung Hauptkomponenten

U^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Eine Hauptkomponente ist eine Linearkombination aller beobachteten Variablen
- > Es gibt so viele Hauptkomponenten wie beobachte Variablen
- > Die Hauptkomponenten sind unkorreliert
- > Die Hauptkomponenten werden so bestimmt, dass sie sukzessive maximale Varianz aufweisen
- > Da PC1 die gemeinsame Kovariation ausdrückt, kann sie sinnvoller sein als das arithmetische Mittel
- > ACHTUNG: PCs sind KEINE Prozesse und machen physikalisch oft keinen Sinn! Man hofft nur, dass sich die ersten Hauptkomponenten gut interpretieren lassen und man so die Daten besser verstehen kann

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

34

EXTREMWERTSTATISTIK

Statistische Datenanalyse

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen, Modellvalidierung
weiterführende Methoden	Daten zusammenfassen	Extremwertstatistik	
	Hauptkomponentenanalyse	Clusteranalyse	Neural Networks
Fallen der Statistik			

Motivation

Schutz

- > der Gesellschaft
- > vor finanziellen Schäden
- > von Infrastruktur

vor seltenen x-jährlichen Ereignissen

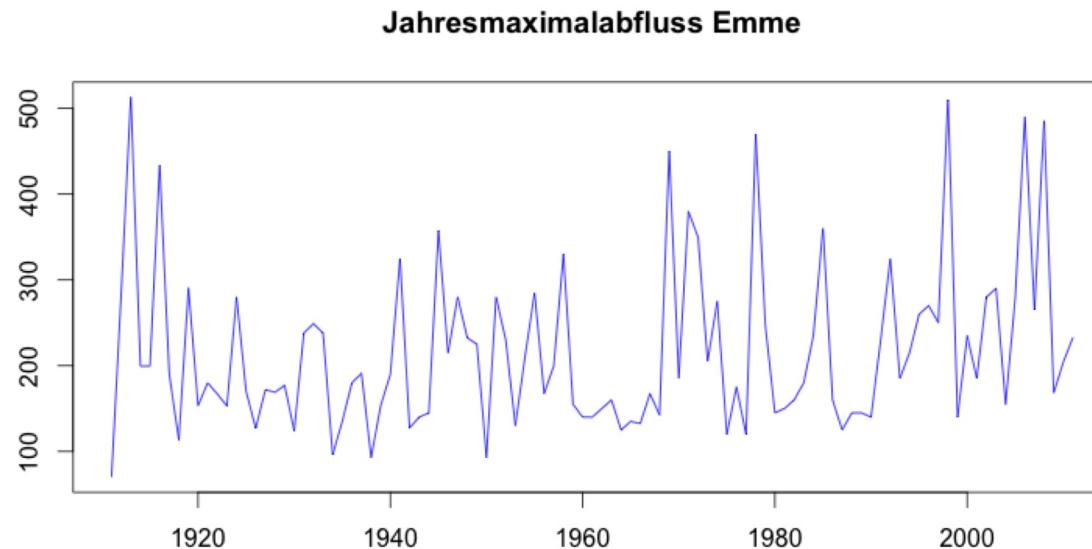
mittels Berechnung von

- > Wiederkehrwerte (Return level) und
- > Wiederkehrperiode (Return period)



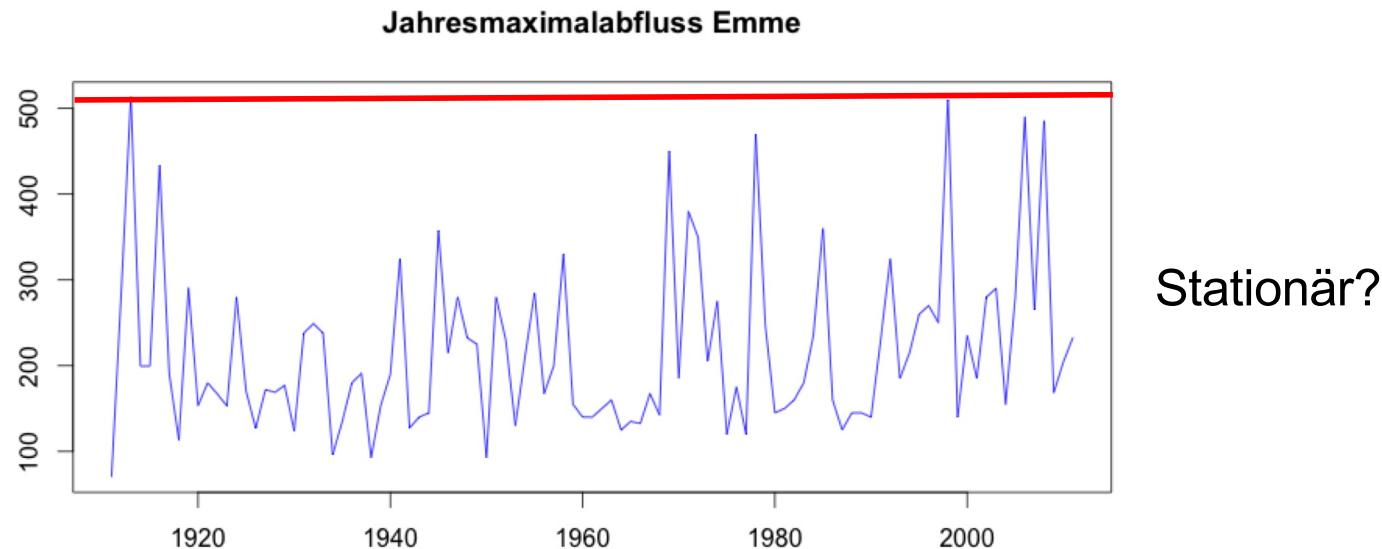
Wiederkehrwerte und Wiederkehrperiode

- > $X(T)$: Wiederkehrwert X der Wiederkehrperiode T
- > Wenn T die Einheit Jahre hat, dann ist X der Wert der jedes Jahr mit einer Wahrscheinlichkeit von $1/T$ überschritten wird



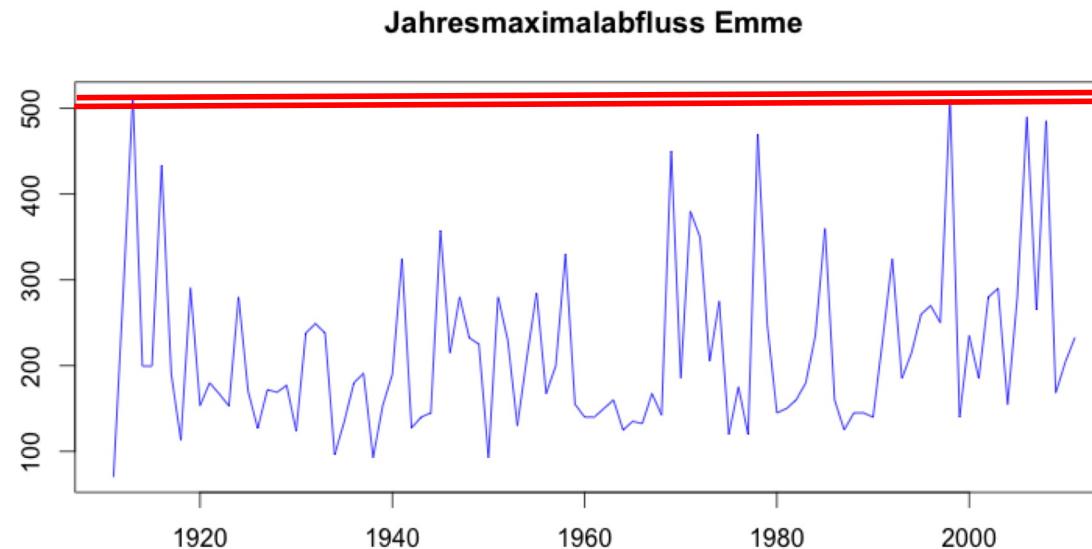
Wiederkehrwerte und Wiederkehrperiode

- > $X(T)$: Wiederkehrwert X der Wiederkehrperiode T
- > Wenn T die Einheit Jahre hat, dann ist X der Wert der jedes Jahr mit einer Wahrscheinlichkeit von $1/T$ überschritten wird



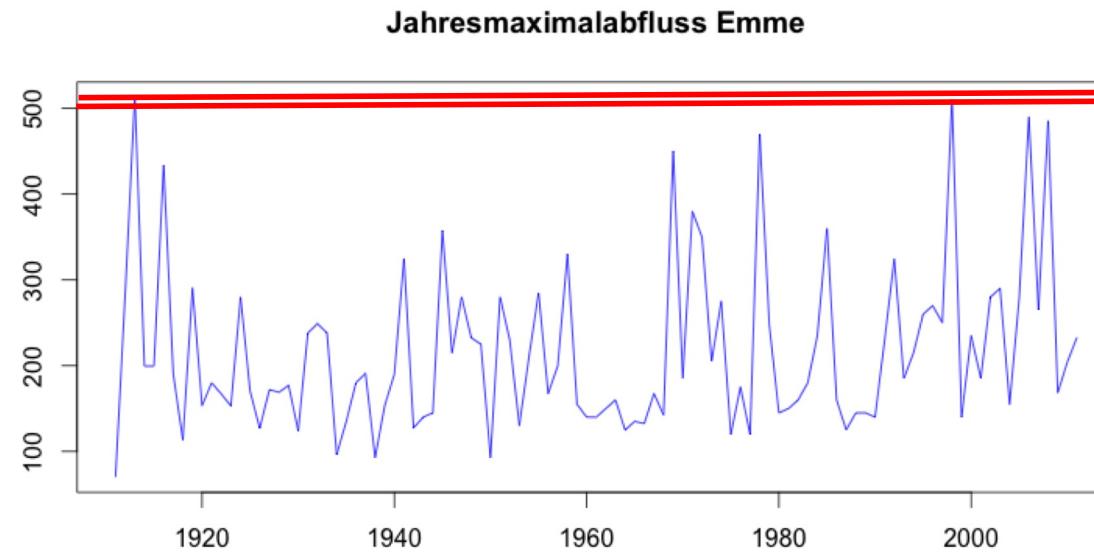
Wiederkehrwerte und Wiederkehrperiode

- > $X(T)$: Wiederkehrwert X der Wiederkehrperiode T
- > Wenn T die Einheit Jahre hat, dann ist X der Wert der jedes Jahr mit einer Wahrscheinlichkeit von $1/T$ überschritten wird



Wiederkehrwerte und Wiederkehrperiode

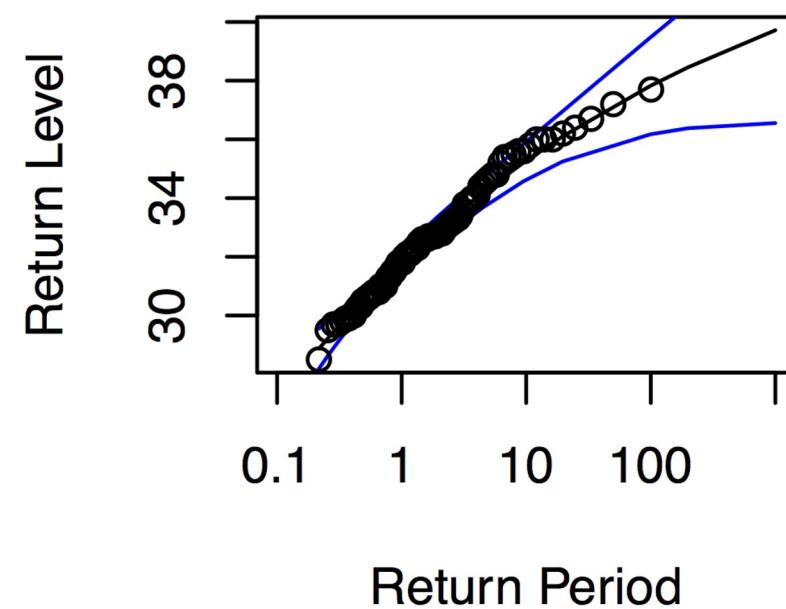
Aber wie kann man ein 1000-jähriges Ereignis bestimmen?



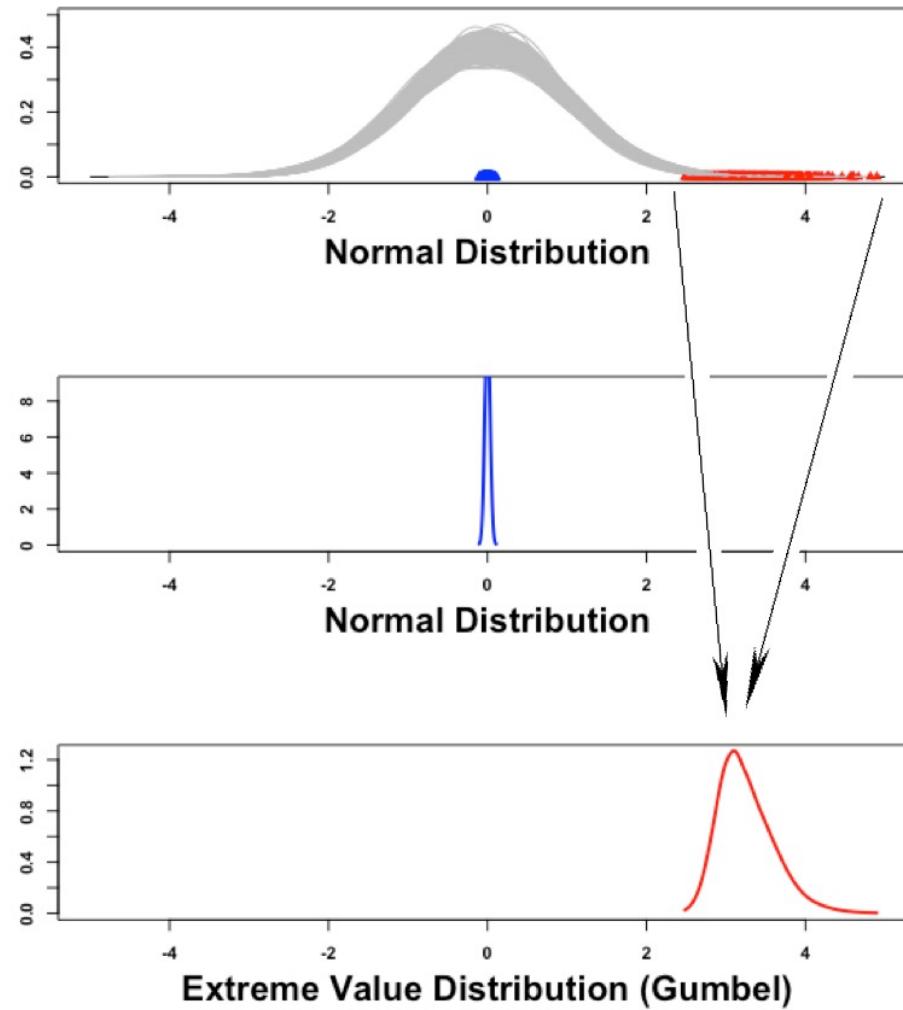
Schätzung von Wiederkehrwerten

Aber wie kann man ein 1000-jähriges Ereignis bestimmen?

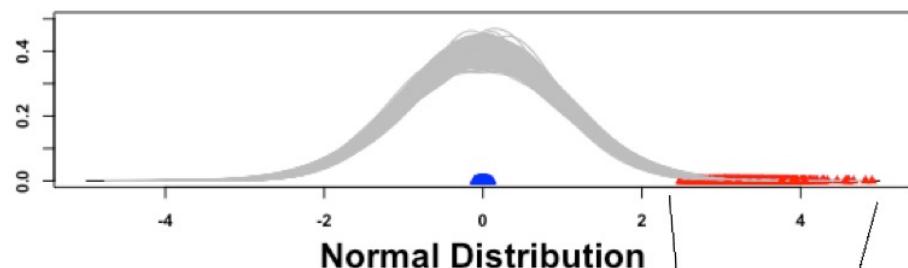
1. Funktion anpassen, die das Verhältnis von Wiederkehrwert und -periode in bestehenden Daten beschreibt
2. Diese Funktion extrapolieren
3. Unsicherheiten schätzen



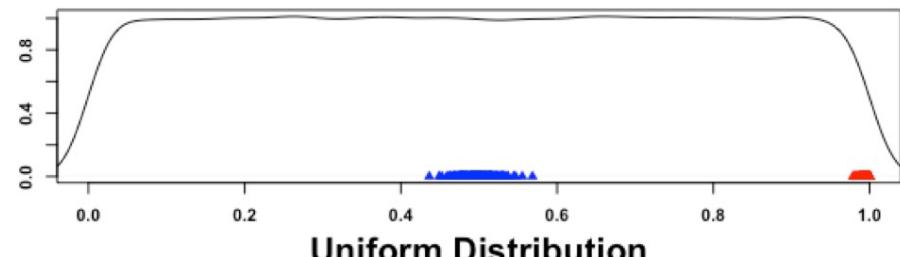
Extremwertverteilungen



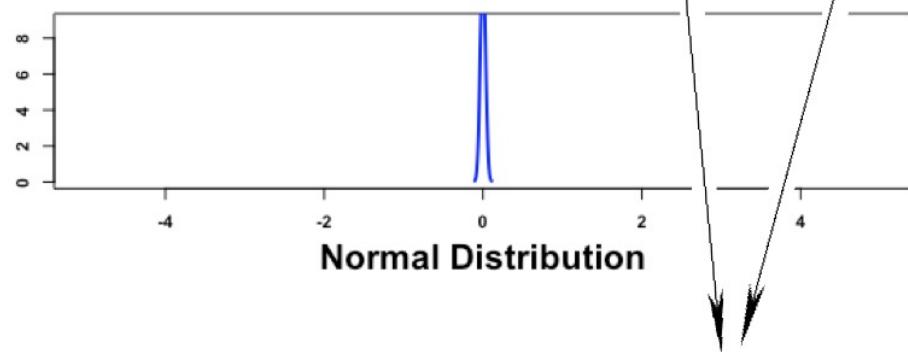
Extremwertverteilungen



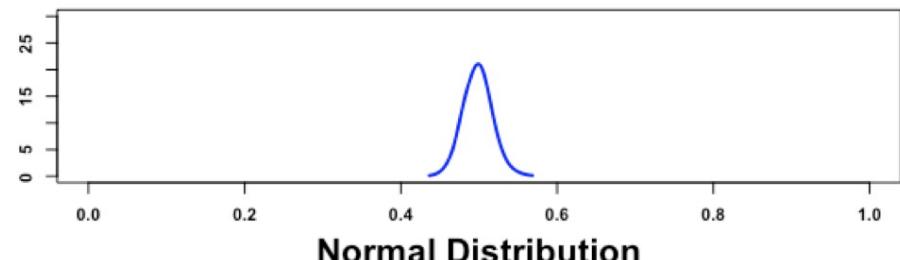
Normal Distribution



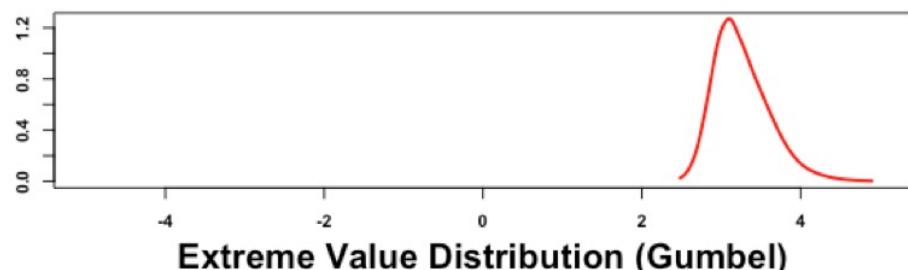
Uniform Distribution



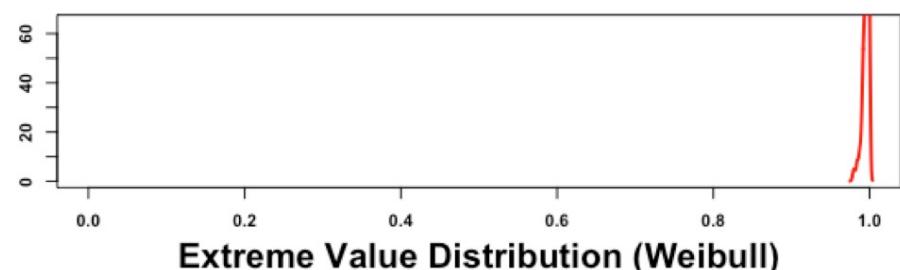
Normal Distribution



Normal Distribution



Extreme Value Distribution (Gumbel)



Extreme Value Distribution (Weibull)

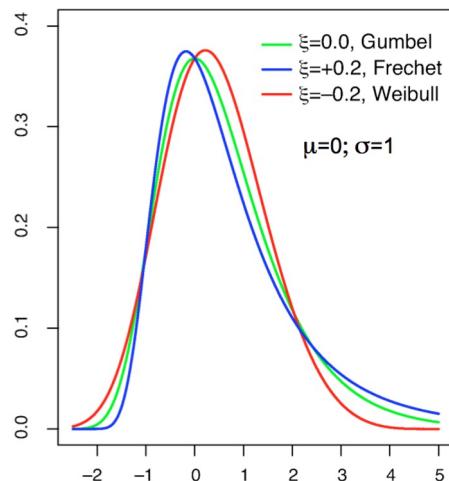
Extremwertverteilungen

Zentraler Grenzwertsatz

- > Das Mittel einer grossen Summe an Zufallsvariablen ist normalverteilt, *unabhängig von der Verteilung der Ausgangsdaten.*

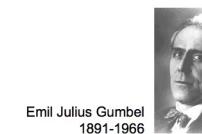
Theorem für Extremwertverteilungen

- > Das Maximum ein grossen Anzahl an Zufallsvariablen ist *Gumbel-, Fréchet- oder Weibull-verteilt, unabhängig von der Verteilung der Ausgangsdaten.*



- **The Gumbel distribution**

$$G(x) = \exp(-\exp(-x))$$



Emil Julius Gumbel
1891-1966

- **The Fréchet distribution**

$$G(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0, \alpha > 0 \end{cases}$$



Maurice René
Fréchet
1878-1973

- **The Weibull distribution**

$$G(x) = \begin{cases} \exp(-(-x)^\alpha) & x < 0, \alpha > 0 \\ 1 & x \geq 0 \end{cases}$$



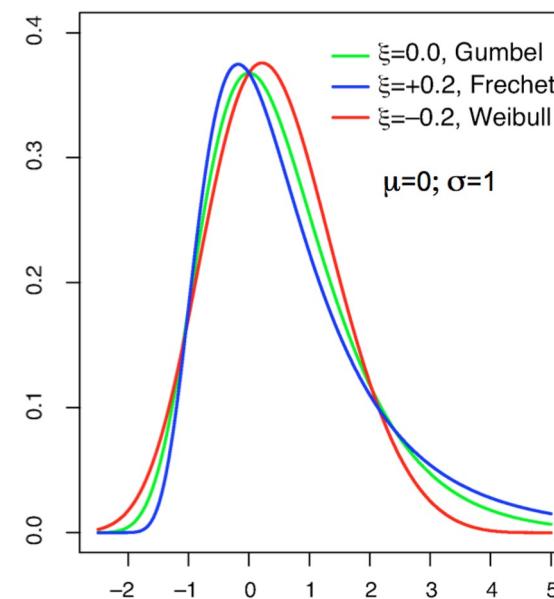
Ernst Hjalmar
Waloddi Weibull
1887-1979

Generalized Extreme Value Distribution (GEV)

Generalisierung dieser drei Verteilungen

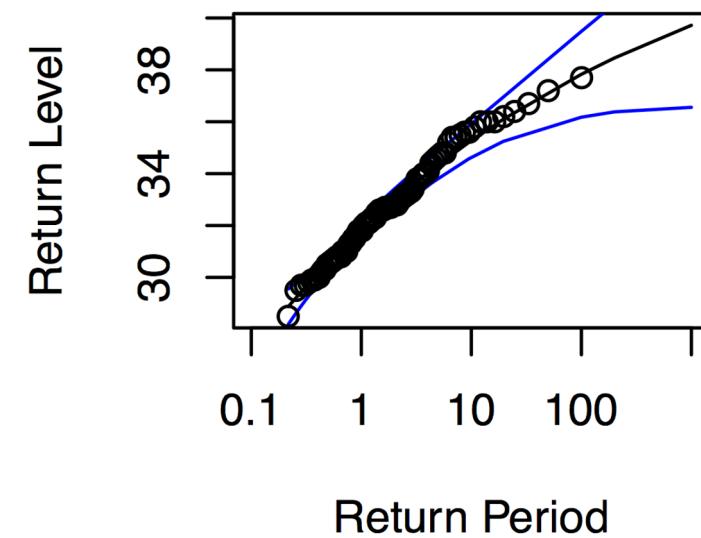
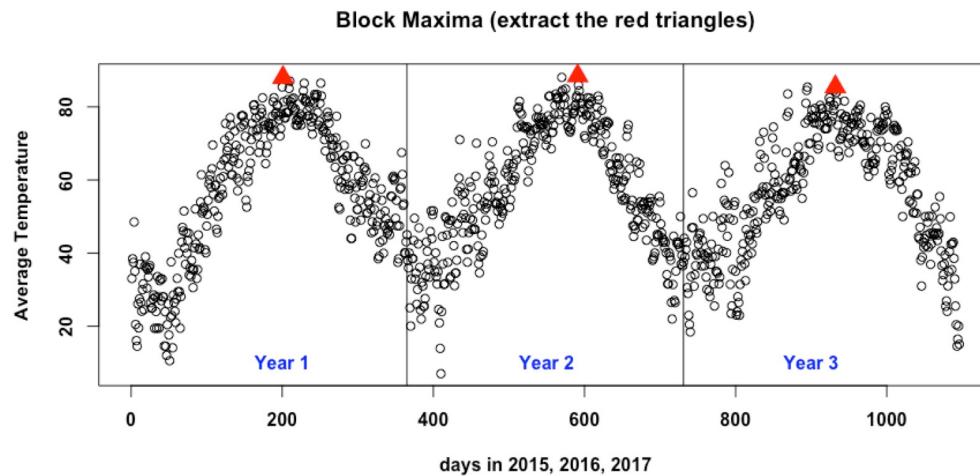
$$GEV(x; \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

- > Location Parameter (μ),
- > Scale Parameter (σ),
- > Shape Parameter (ξ)
 - $\xi = 0$: *Gumbel*, unbegrenzt
 - $\xi > 0$: *Fréchet*, untere Grenze
 - $\xi < 0$: *Weibull*, obere Grenze



Block Maxima

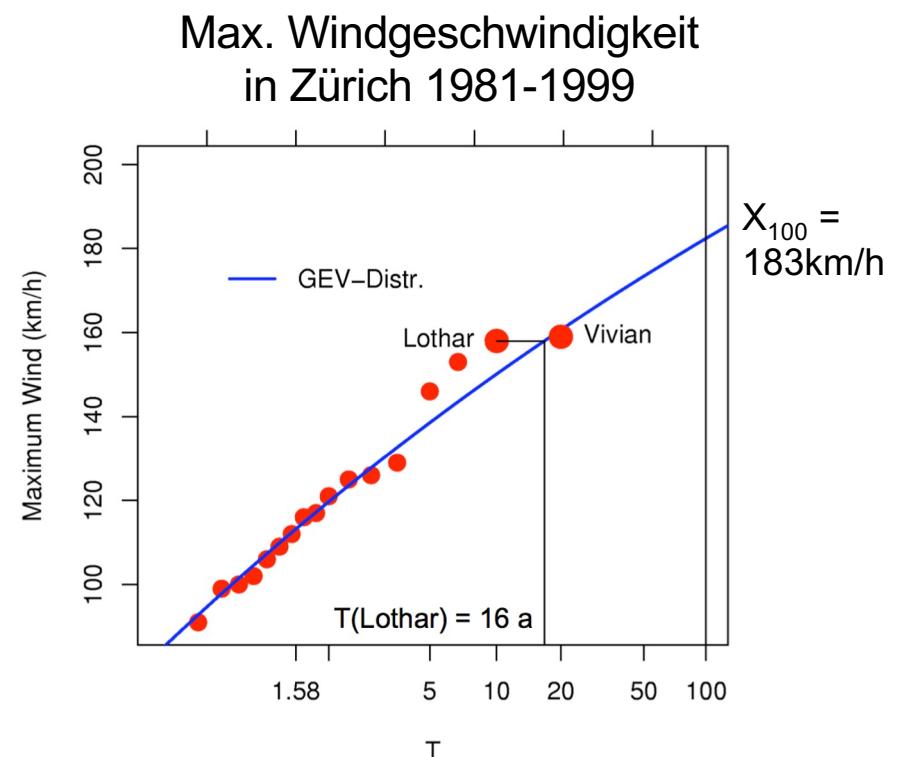
1. Daten in gleich grosse Blöcke unterteilen, z.B. Jahre
2. Maximum jedes Blocks berechnen
3. GEV an die Block-Maxima anpassen
4. Unsicherheiten schätzen



Annahmen und Interpretation

1. Stationarität, d.h. kein Trend über die Zeit
2. Unabhängige Daten, d.h. keine Autokorrelation

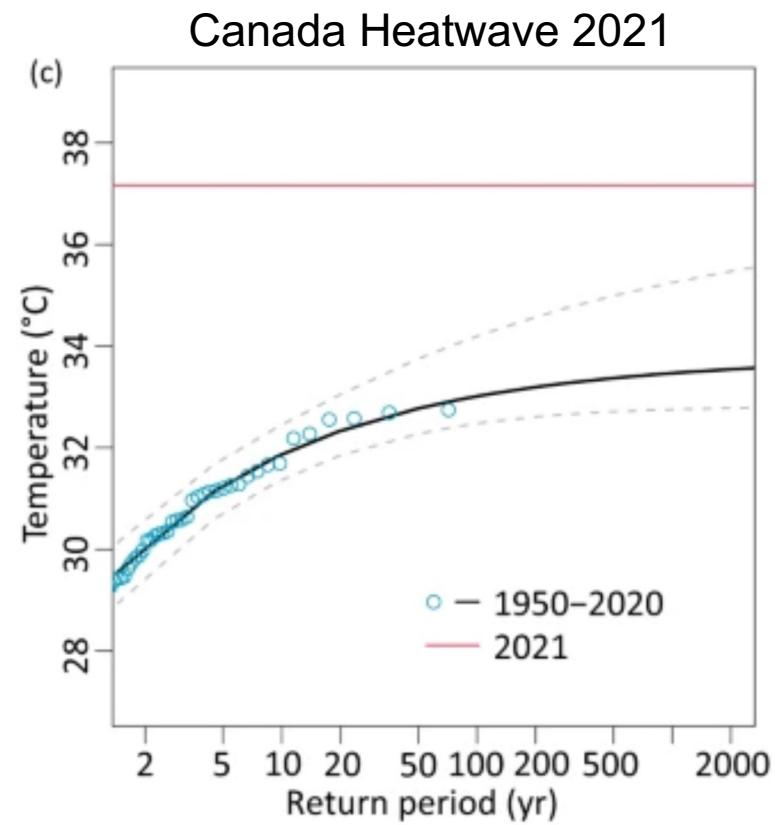
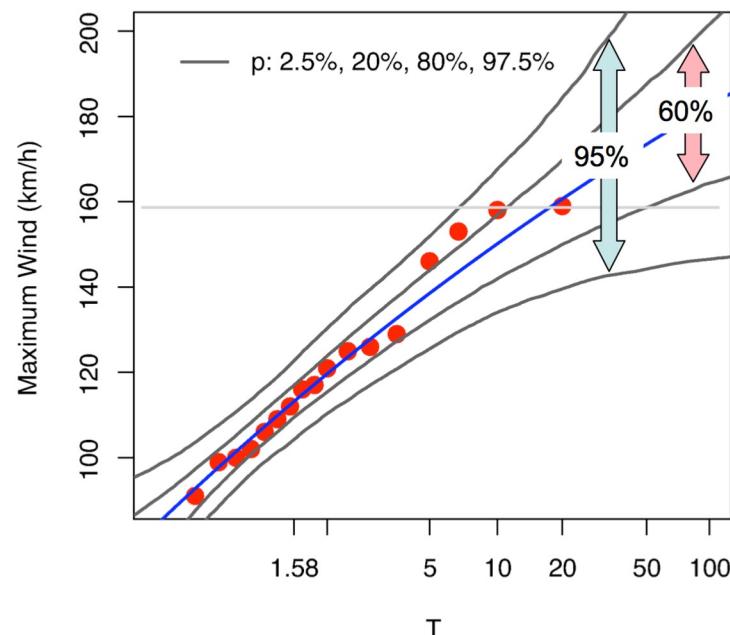
Wie schätzt man nun die Unsicherheit in den GEV Parametern?



Unsicherheit Schätzen

Konfidenzintervalle bestimmen

- ✓ delta Methode
- ✓ L-moments
- ✓ Maximum Likelihood



Fischer et al. 2023

Peak-Over-Threshold Methode

Vorteil:

- > Besser geeignet für kurze Zeitreihen, da nicht nur Jahresmaxima genutzt werden

Difficulties:

- > Ereignisse müssen getrennt werden, z.B. hoher Abfluss über mehrere Tage
- > «Thresholds» müssen subjektiv gewählt werden

Beispiel Prüfungsfragen

- 1. Welcher der folgenden Nachteile trifft auf das K-Means-Clustering zu?**
 - a. Es erfordert einen rechnerisch aufwendigen iterativen Prozess
 - b. Es ist empfindlich gegenüber der anfänglichen Platzierung der Zentroiden
 - c. Es erfordert, dass die Anzahl der Cluster im Voraus festgelegt wird
- 2. Wie wird die erste Hauptkomponente in der PCA bestimmt?**
 - a. Sie ist die Komponente mit der geringsten Varianz
 - b. Sie ist die Komponente, die die meiste Varianz der Daten erklärt
 - c. Sie ist die Komponente, die die Daten in Cluster gruppiert
 - d. Sie ist die Komponente, die die Daten normalisiert
- 3. Welche Parameter werden in der GEV-Verteilung verwendet?**
 - a. Mittelwert, Varianz, Schiefe
 - b. Lageparameter, Skalenparameter, Formparameter
 - c. Median, Modus, Standardabweichung
 - d. Erwartungswert, Varianz, Kurtosis

u^b

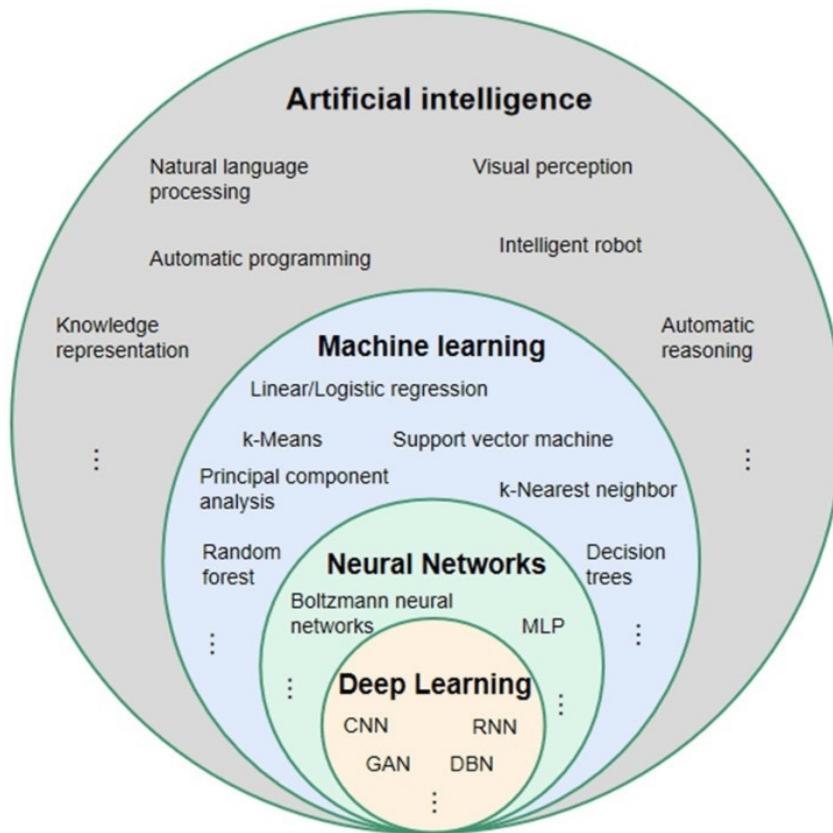
b
UNIVERSITÄT
BERN

Neuronale Netze (NN)

Neuronale Netze (NN)

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
	Statistische Tests Konfidenzintervalle	Korrelation	Regression
	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Gibt es gemeinsame gleich- oder entgegengerichtete Variationen	Kausalzusammenhänge für Vorhersagen oder Interpolationen nutzen
weiterführende Methoden	Daten zusammenfassen		Extremwertstatistik
Fallen der Statistik	Hauptkomponentenanalyse	Clusteranalyse	Neural Networks

„Künstliche Intelligenz“



- **Künstliche Intelligenz:** jede Technik, die menschliches Verhalten oder menschliche Entscheidungsfindung nachahmt
- **Machine Learning:** ein Teilbereich der KI. Verwendet statistische Methoden, um aus Daten zu lernen.
- **Neuronale Netze** sind das Herzstück der aktuellen KI-Revolution
- **Deep Learning:** ein Teilbereich des maschinellen Lernens, der komplexe NN-Architekturen (CNNs, RNNs, ...) umfasst

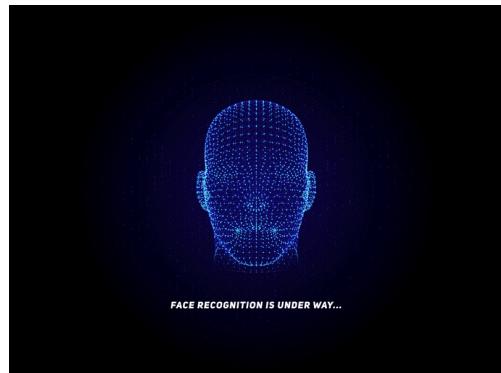
Importance of NNs in industrial and science

U^b

b
UNIVERSITÄT
BERN



apple.com



pinterest.com

Since 2015, AI beats humans
in visual recognition

5

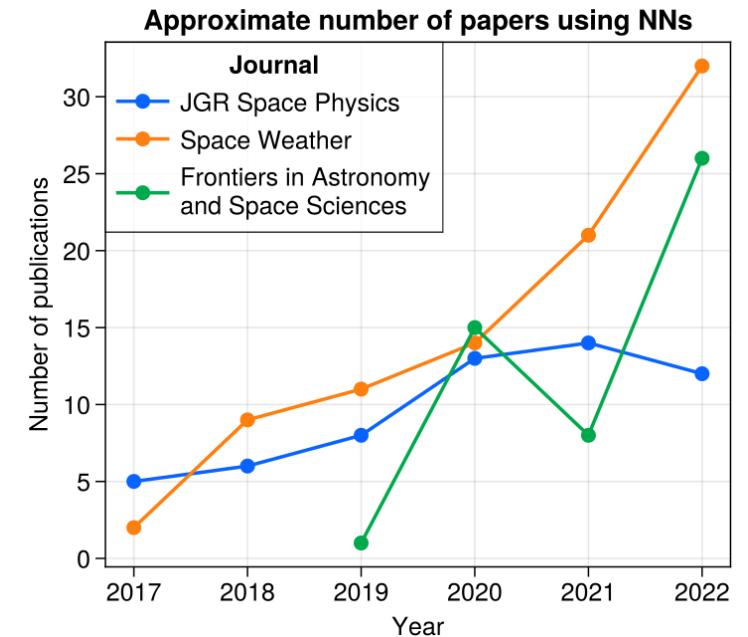


google.com



freecodecamp.org

Since 1997, AI beats humans
in the game of chess



Was ist ein Neurales Netzwerk?

u^b

b
UNIVERSITÄT
BERN

- Sind eine breite Familie mathematischer Modelle
- bestehen aus einer Reihe miteinander verbundener Knotenpunkte (auch Neuronen genannt)
- sind in Schichten organisiert
- sind von der Struktur des menschlichen Gehirns inspiriert
- haben eine sehr große Anzahl von Parametern
- können sehr komplexe nicht-lineare Beziehungen abbilden

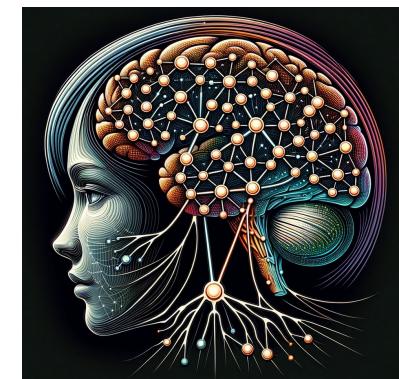
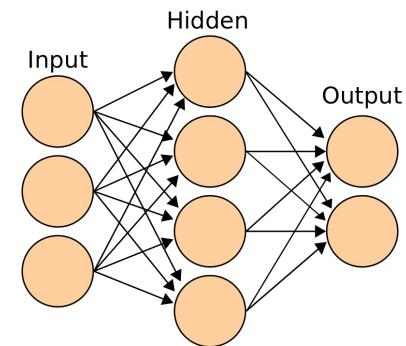


Image generated by DALL-E

u^b

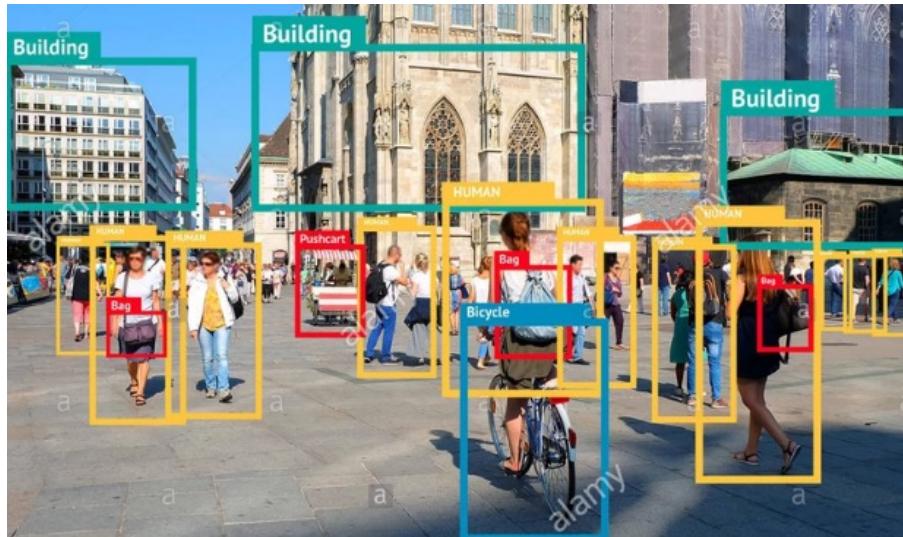
b
UNIVERSITÄT
BERN

Grundlegende Konzepte des maschinellen Lernens

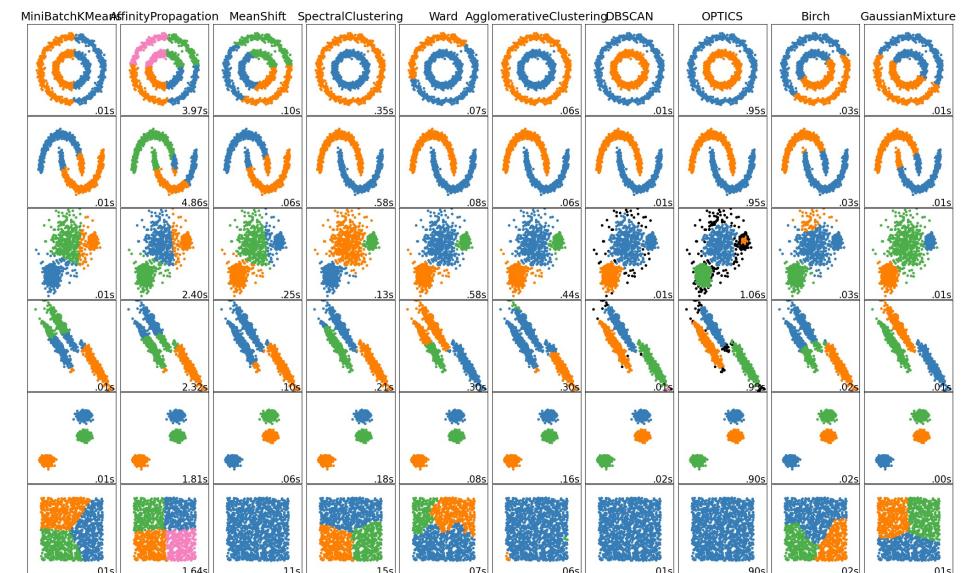
Supervised vs Unsupervised learning

Überwachtes vs Unüberwachtes Lernen

Zwei Kategorien von Methoden
Neuronale Netze können für beides verwendet werden



Objekterkennung: *supervised*



Clustering: *unsupervised*

Die Antwort ist unbekannt

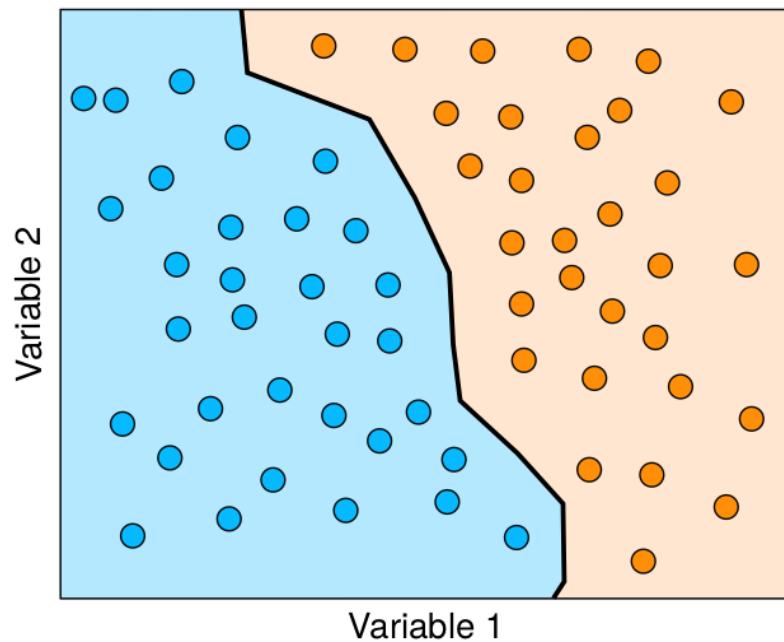


Klassifikation und Regression

u^b

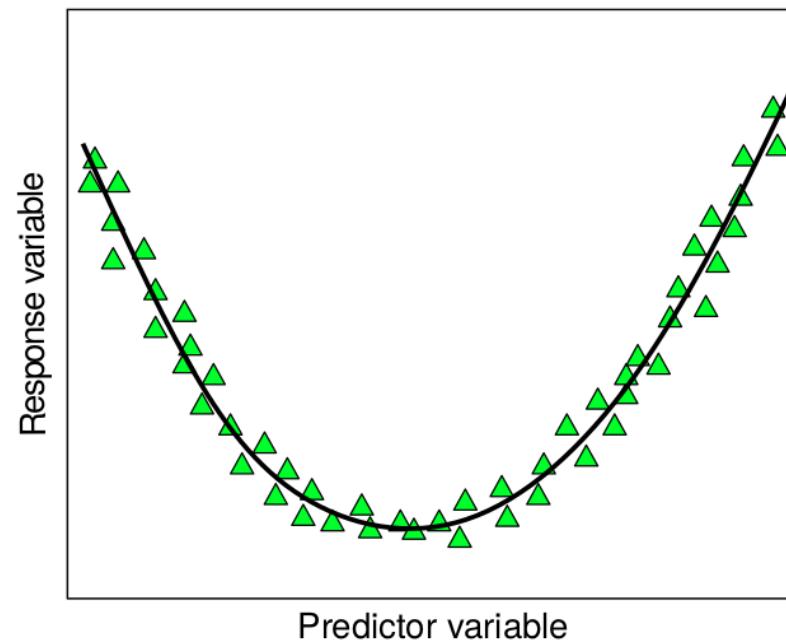
b
UNIVERSITÄT
FRANKFURT

(a) Classification



Diskrete Ergebnisse

(b) Regression

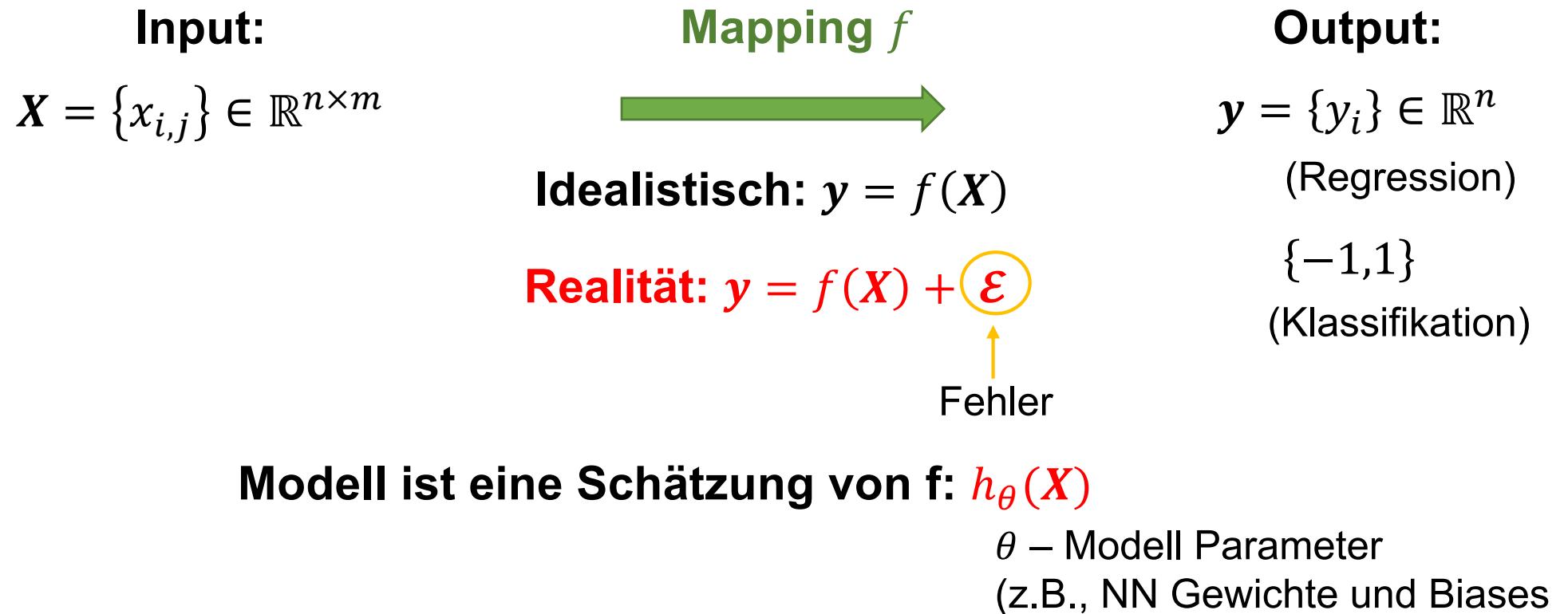


Kontinuierliche Ergebnisse

Was ist ein Modell?

u^b

b
UNIVERSITÄT
BERN



Wie finden wir ein “gutes” Modell?

u^b

b
UNIVERSITÄT
BERN

Modell: $h_\theta(X)$

θ – Modell Parameter, gelernt aus X

Modell Vorhersagen: $\hat{y} = h_\theta(X)$

Wir hoffen: $\hat{y} \approx y$

Minimierung des Unterschieds zwischen \hat{y} und y . “Cost” function $J(\hat{y}, y)$
e.g., MSE, MAE.

Wie finden wir ein “gutes” Modell?

u^b

b
UNIVERSITÄT
BERN

Aber in der Realität gibt es noch Fehler : $y = f(X) + \boxed{\varepsilon}$

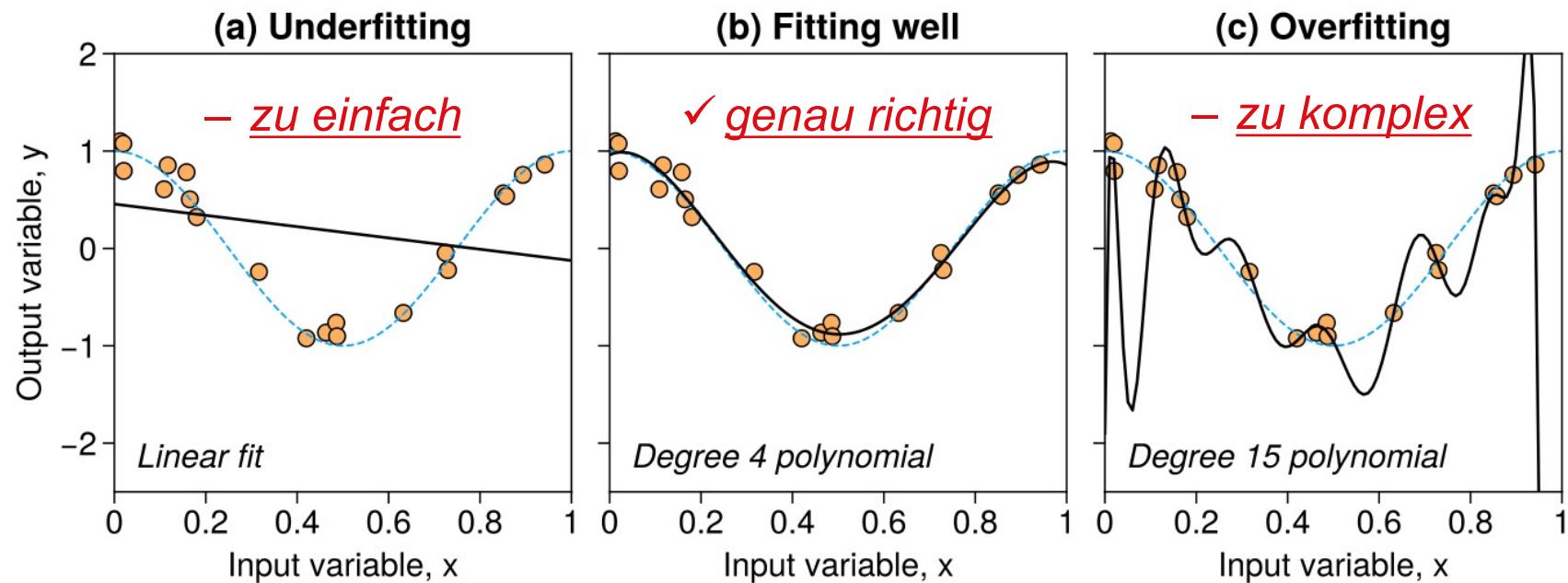
3 Szenarien:

$$\hat{y} = h_{\theta}(X)$$

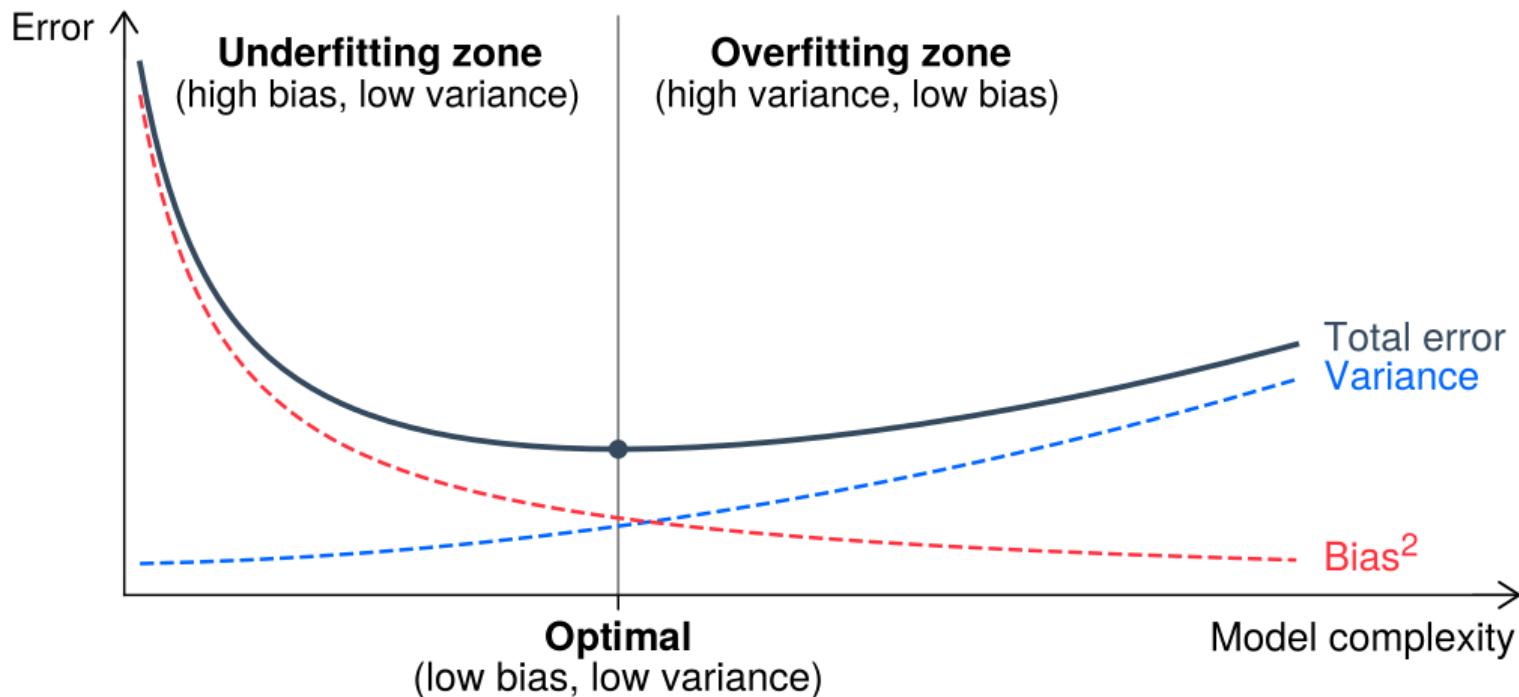
model predictions

- $\hat{y} \equiv y$, das Modell lernt die Datenfehler ε !
- $\hat{y} \neq y$, das Modell ist weit entfernt von den Daten
- $\hat{y} \approx y$, das Modell stimmt gut mit den Daten überein

Wie finden wir ein “gutes” Modell?



Bias-variance tradeoff

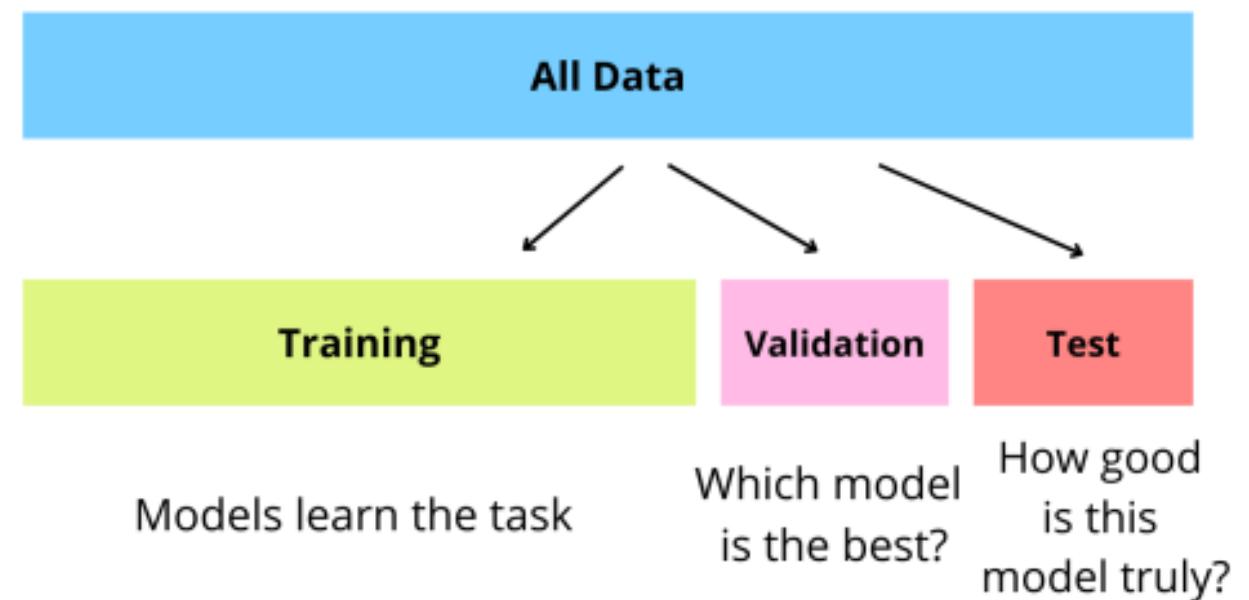


Wie können wir die Modell Qualität beurteilen?

U^b

b
UNIVERSITÄT
BERN

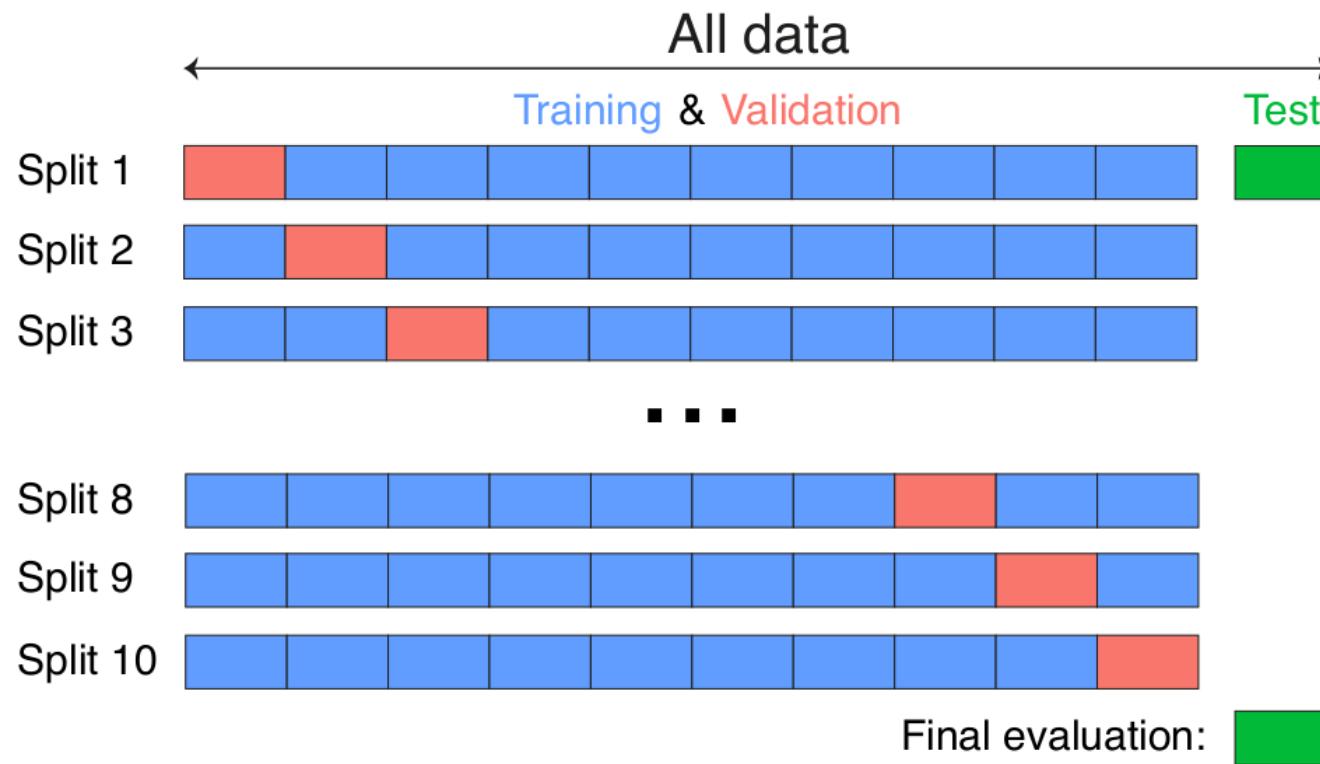
Wie müssen die Daten
in drei Teile aufteilen



[Medium.com]

Kreuzvalidierung

Daten und K Teile aufteilen und das Modell K Mal auf K-1 Teile trainieren



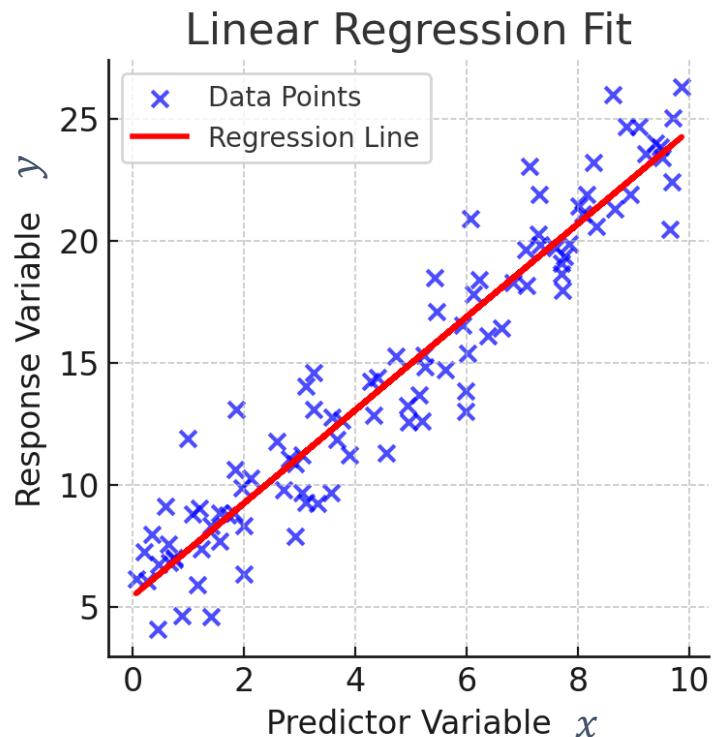
[Smirnov et al., 2020]

u^b

b
UNIVERSITÄT
BERN

Iteratives training

Linear Methoden



$$\hat{y} = h_{\theta}(x) = wx + b$$

Modell Parameter: $\theta = \{w, b\} = \{\theta_0, \theta_1\}$

Kostenfunktion: $J(h_{\theta}) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Wir müssen θ finden, so dass $J(h_{\theta}) \rightarrow \min$

Besten „Fit“ finden

u^b

b
UNIVERSITÄT
BERN

$$\text{Kostenfunktion: } J(h_\theta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

1. Kleinstes Quadrate
2. Optimierung der Kostenfunktion mittels Gradientenabstieg

Beste Anpassung mittels Gradientenabstieg

Wir konzentrieren uns auf w

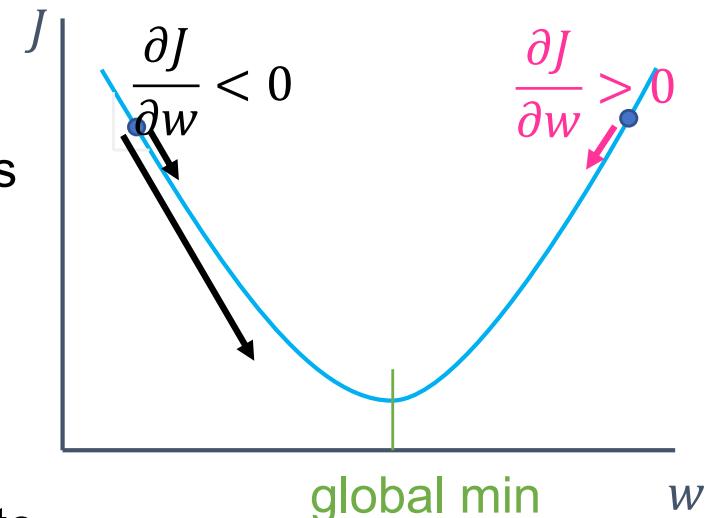
$\frac{\partial J}{\partial w}$ gibt uns eine Steigung (Richtung) des Abstiegs

Aber wie groß ist die Schrittweite?

Wir brauchen eine Skalierung, α , genannt Lernrate

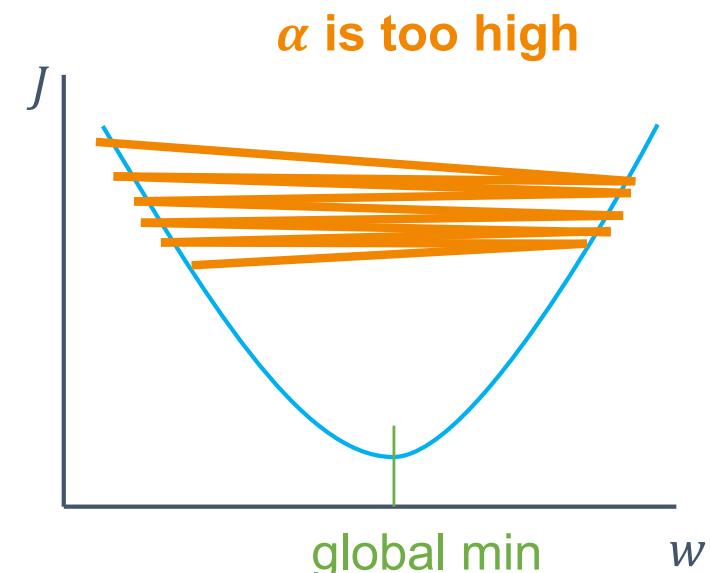
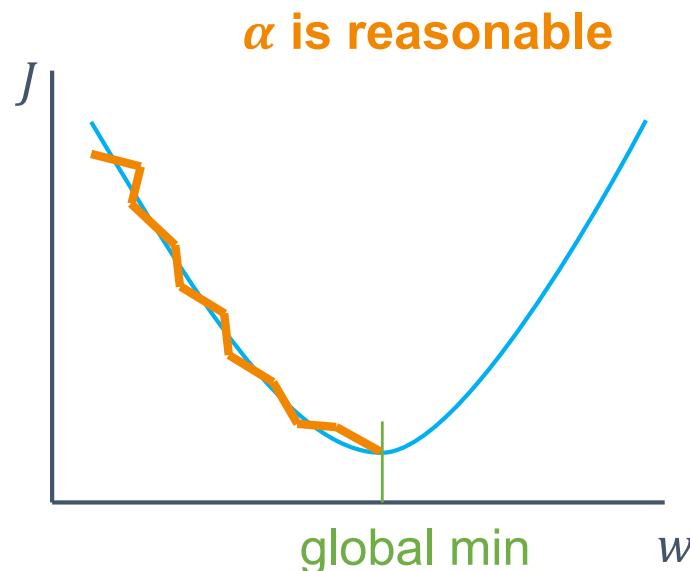
Der aktualisierte Wert ist : $w^{new} = w^{old} - \alpha \frac{\partial J}{\partial w}$

die Schrittweite



Beste Anpassung mittels Gradientenabstieg

Die Lernrate ist ein entscheidender Parameter, der abgestimmt (getuned) werden muss

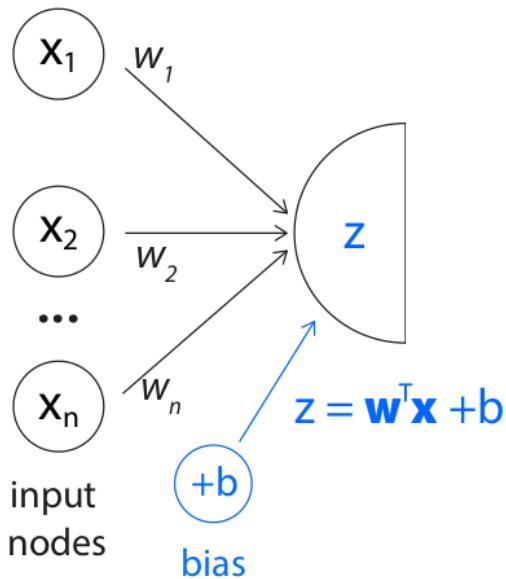


u^b

b
UNIVERSITÄT
BERN

Bausteine neuronaler Netze

Von linearer Regression zu einem Neuron

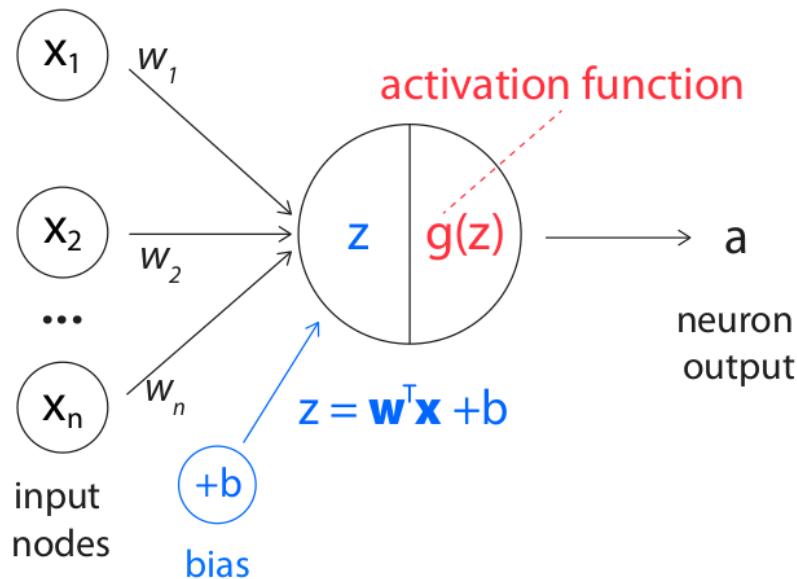


Wie kann man Nicht-Linearität einbauen?

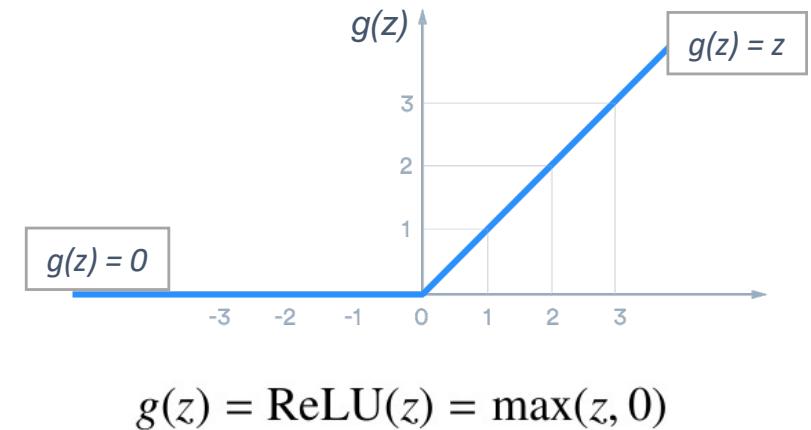
Wir müssen die Ausgabe z durch eine nichtlineare Funktion leiten

~ lineare Regression

From a linear regression to a neuron



ReLU Aktivierungsfunktion

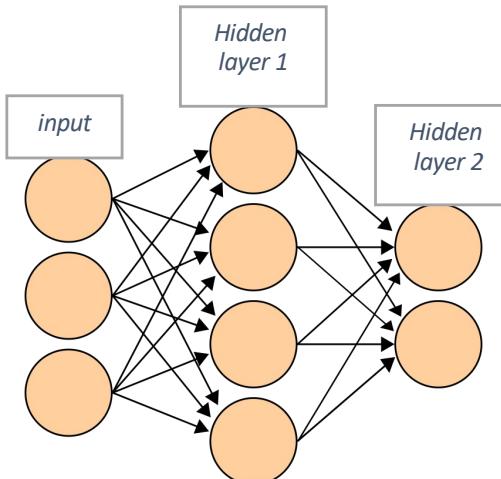
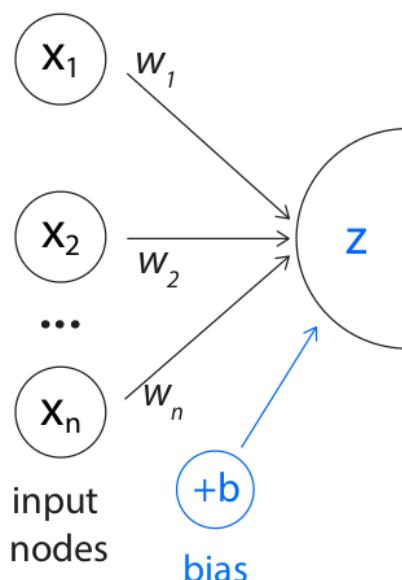


Aktivierungsfunktionen ermöglichen die Anpassung komplexer nichtlinearer Muster

Was wäre, wenn wir diese Einheit mehrmals in Schichten anordnen würden?

U^b

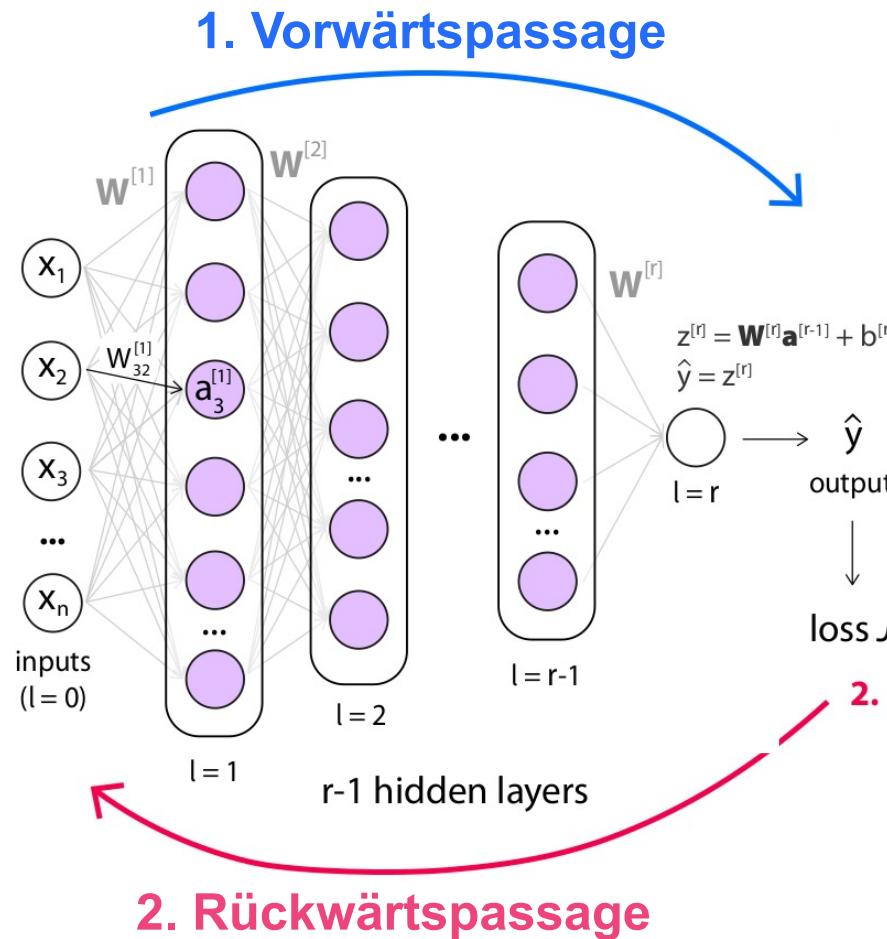
b
UNIVERSITÄT
BERN



$$1^{\text{st}} \text{ hidden layer: } z^{[1]} = W^{[1]}x + b^{[1]}$$

$$2^{\text{nd}} \text{ hidden layer: } z^{[2]} = W^{[2]}z^{[1]} + b^{[2]}$$

Die Idee hinter „Backpropagation“



Update Regel:

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \alpha \frac{\partial J}{\partial \mathbf{W}} ; \quad \mathbf{b}_t = \mathbf{b}_{t-1} - \alpha \frac{\partial J}{\partial \mathbf{b}}$$

Gewichte und Verzerrungen (Biases) werden iterativ aktualisiert, und die Leistung verbessert sich mit der Zeit.

Parameters vs Hyperparameters

u^b

b
UNIVERSITÄT
BERN

Hyperparameters kontrollieren wie W und b angepasst werden

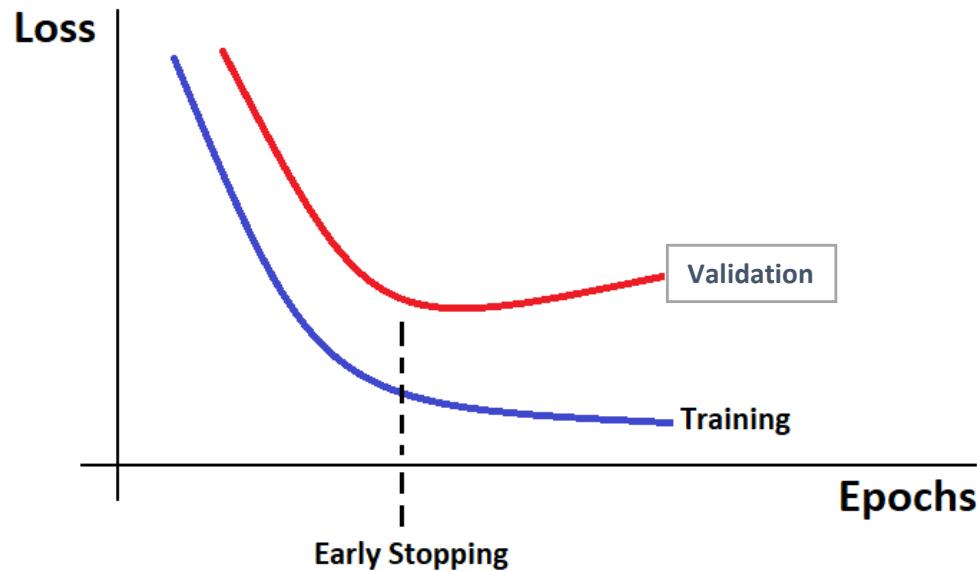
Modell Parameter:

- Weights W
- Biases b

Modell Hyperparameter:

- Lernrate α
- Anzahl Schichten (layer)
- Anzahl Neuronen
- Aktivierungsfunktion (g)
- ...

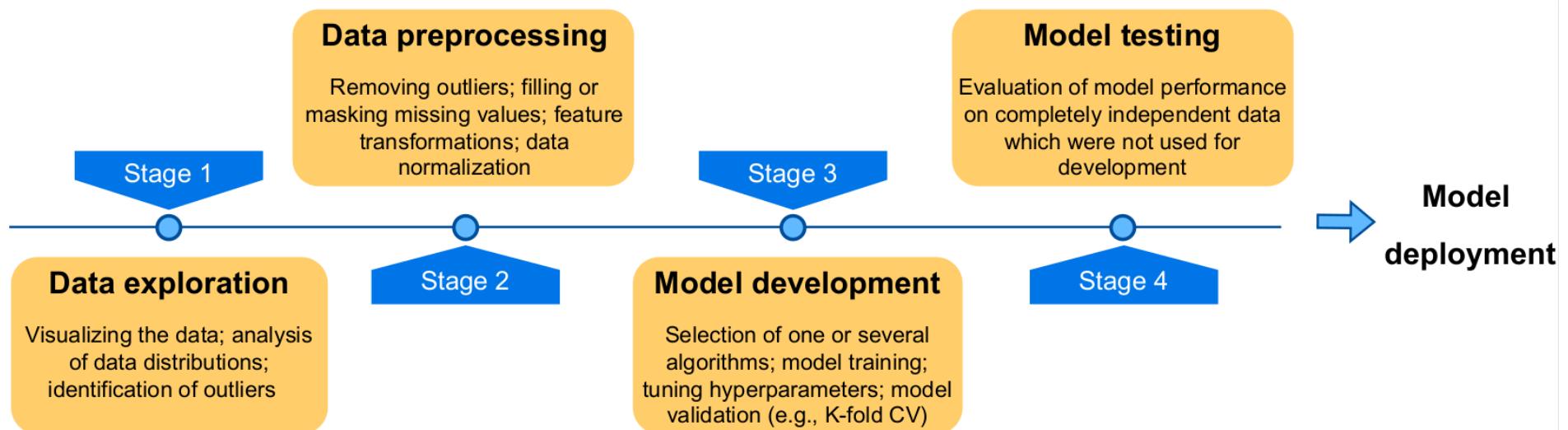
Training beenden bei „Overfitting“



Sobald die Trainings- und Validierungskurven auseinanderlaufen, kann das Training abgebrochen werden.

Das „optimale“ Modell wird gespeichert

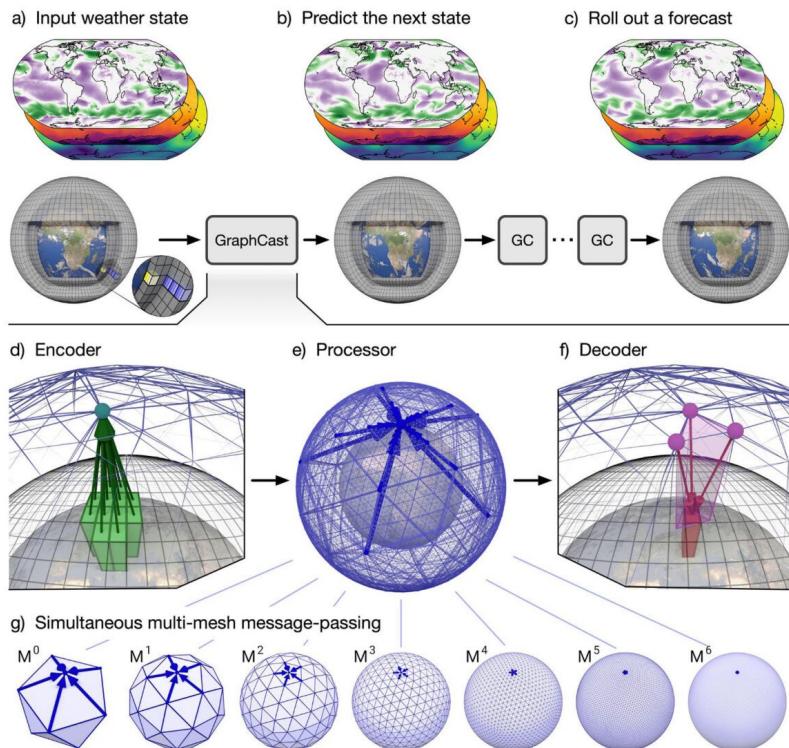
Ablauf von Machine Learning Projekten



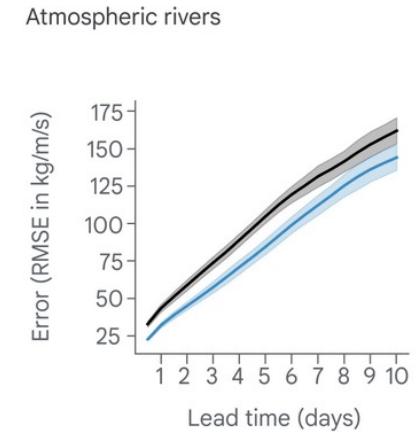
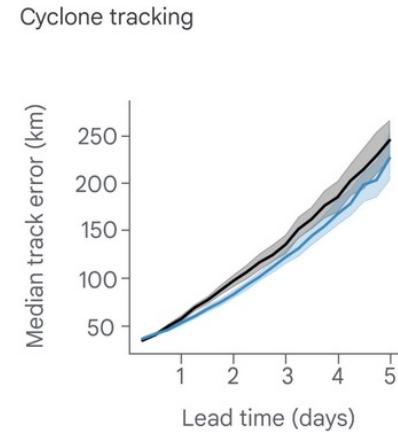
Ein Beispiel was Machine Learning heutzutage kann

\mathcal{U}^b

b
UNIVERSITÄT
BERN



Google DeepMind Wettervorhersagemodell



[Lam et al., 2024, SciAdv]

u^b

b
UNIVERSITÄT
BERN

Evaluation

[https://scanserveruls.
unibe.ch/evasys/online.php?pswd=WCX
NE](https://scanserveruls.unibe.ch/evasys/online.php?pswd=WCXNE)



Beispiel Prüfungsfragen

Neuronale Netze sind

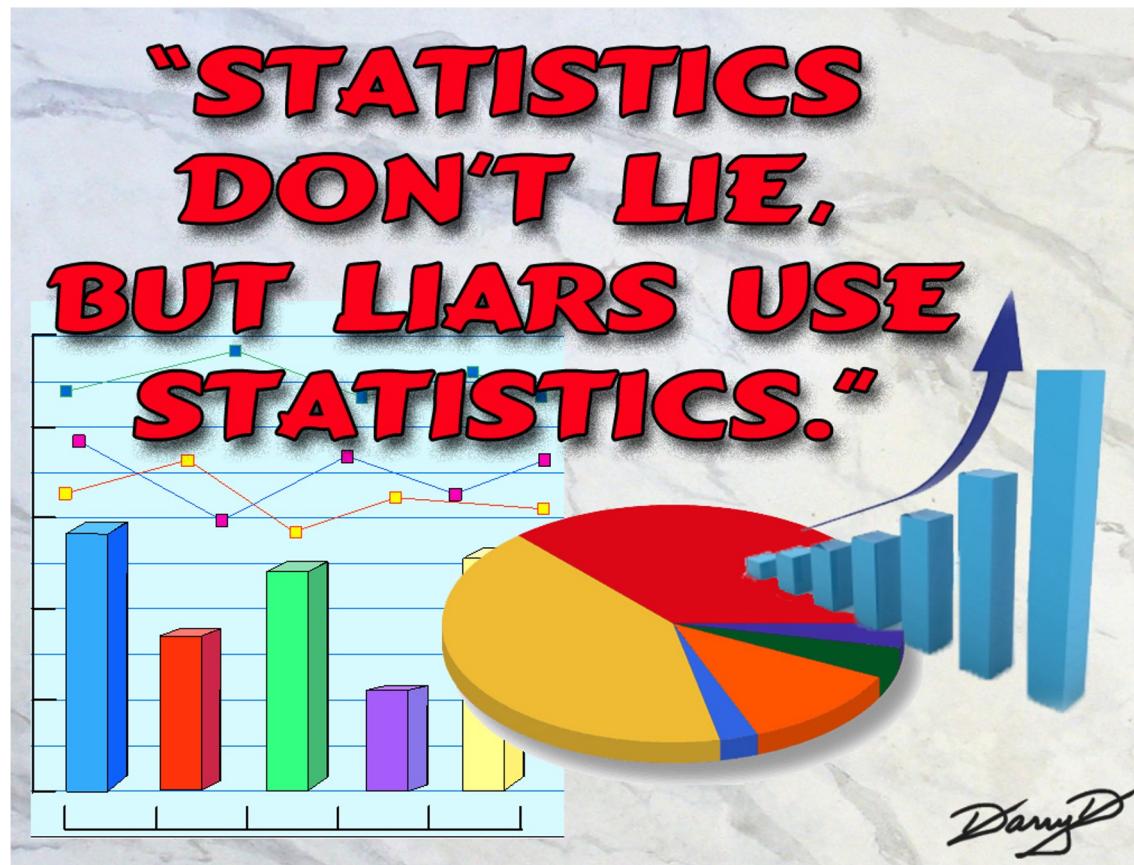
- a) ein anderen Begriff für Künstliche Intelligenz
- b) ein anderer Begriff für Maschinelles Lernen
- c) ein Teilbereich des Maschinellen Lernens
- d) Die Basis von Deep Learning

Neuronale Netze eignen sich

- a) zur Klassifikation
- b) zur nicht-linearen Regression
- c) supervised learning/überwachtes Lernen
- d) UNsupervised learning /UNüberwachtes Lernen

Welche Aussagen zu Neuronalen Netzwerken sind RICHTIG:

- a) Neuronalen Netzwerke neigen zum Overfitting
- b) Neuronalen Netzwerke neigen zum Underfitting
- c) Neuronalen Netzwerke optimieren alle (Hyper-)Parameter selbstständig
- d) Neuronalen Netzwerke erfordern die Wahl einiger (Hyper-)Parameter durch den Nutzer



Fallen in der Statistik

Fehler identifizieren und selber vermeiden

Statistische Datenanalyse

Deskriptive Statistik

Rohdaten
visualisieren

Datenqualität
prüfen

statistische
Masszahlen

Schliessende Statistik

Unterschiede
identifizieren

Zusammenhänge
identifizieren

Abhängigkeiten
modellieren

Statistische Tests
Konfidenzintervalle

Korrelation

Regression

Wie wahrscheinlich
sind die Daten der
Stichprobe, wenn
die Nullhypothese
zutrifft?

Gibt es gemein-
same gleich- oder
entgegengerichtete
Variationen

Kausalzusammen-
hänge für Vorher-
sagen oder Inter-
polationen nutzen

weiterführende
Methoden

Daten
zusammenfassen

Extremwertstatistik

Hauptkomponenten-
analyse

Clusteranalyse

Neural Networks

Fallen der Statistik

Experiment

- > Nehmt euch eine Münze und steht auf
- > Kopf oder Zahl?
- > bei Zahl setzt ihr euch hin
- > mehrere Runden

Datamining

Experiment:

- > Was ist das Geheimnis hintereinander 5x die gleiche Seite zu werfen?
- > Natürlich nur Zufall, aber die Wahrscheinlichkeit ist $1/2^5 = 1/32$ oder $p=0.03$,
- > d.h. signifikant mit dem normalerweise angesetzten $\alpha=0.05$, um unsere Nullhypothese zu widerlegen!
- > Wenn wir jetzt clevere Wissenschaftler sind, denken wir uns etwas aus wie es dazu kommen konnte: Training, gute Konzentration, ...
- > was natürlich alles Quatsch ist!
- > Deshalb erst eine Hypothese aufstellen diese dann testen.

Multiple Tests

Mehrfache Analyse derselben Daten

“It's called a 95% confidence interval because it misses 5% of the time”

C. J. Geyer, 2005

Wenn genug Hypothesen getestet werden, findet man zufällig ein positives Ergebnis

Aber Datenerhebung ist teuer und nur ein Test wäre Verschwendung, deshalb:

- > 1. vor der Studie definierte Haupthypothese aufstellen und in Publikation benennen
- > 2. später weitere explorative Hypothesen aufstellen und ebenfalls so kennzeichnen
- > eventuell Signifikanzlevel anpassen (z.B. 0.01 statt 0.05)

p-Werte

- > „The effect of the drug on blood pressure was statistically significant ($p = 0.02$).“
- > **Was sagt uns dieser Satz?**

p-Werte ohne Effektgrösse

- > „The effect of the drug on blood pressure was statistically significant ($p = 0.02$).“
- > Ein p-Wert < 0.05 führt zur Ablehnung der Nullhypothese
- > Die Nullhypothese sagt normalerweise, dass es keine Korrelation oder keinen Unterschied zwischen A und B gibt
- > ABER
- > *“It is foolish to ask ‘Are the effects of A and B different?’ They are always different for some decimal place”* (Tukey, 1991)
- > Wir sind nicht wirklich an “statistischer Signifikanz” interessiert, sondern an physikalisch oder sozial signifikanten Effekten
- > Aber wie gross war der Effekt, in welche Richtung, hat es praktische Relevanz?
- > Korrekt sollte es daher heissen: „The drug lowered diastolic blood pressure by a mean of 18mmHg, from 110 to 92mmHg (95% CI = 2 - 34; $p = 0.02$).“

- > Je mehr Stunden ich bei meinem Tennislehrer nehme, desto schlechter spiele ich!

- > Was ist an diesem Zusammenhang vermutlich falsch?

Umgekehrter Zusammenhang

- > Je mehr Stunden ich bei meinem Tennislehrer nehme, desto schlechter spiele ich!

- > Könnte sein, wenn man einen wirklich schlechten Lehrer hat,
- > vermutlich ist der Zusammenhang jedoch, dass ich mehr Stunden nehme, wenn ich merke, wie schlecht ich spiele.

Kein Effekt oder nicht aussagekräftig?

- > “*Absence of proof is not proof of absence.*” W. Cowper (1731-1800)
- > Studien mit NICHT-signifikanten Resultaten sind ergebnislos und nicht aussagekräftig, d.h. NICHT, dass sie ein negatives Resultat haben!
- > Nur weil wir statistisch nicht sagen können, dass zwei Gruppen unterschiedlich sind, ist das kein Beweis, dass sie gleich sind. Vielleicht bräuchten wir nur eine grössere Stichprobe.
- > oder vielleicht erkennt man einen positiven Trend, aber auf Grund der kurzen Zeitreihe ist dieser noch nicht statistisch signifikant

Simpson's Paradox

Klinische Medikamentenstudie:

All	Treatment group	
	Control	Treated
Alive	60	200
Dead	60	200
Rate	50%	50%

> Ergebnis?

Simpson's Paradox

Klinische Medikamentenstudie:

All	Treatment group	
	Control	Treated
Alive	60	200
Dead	60	200
Rate	50%	50%

Die Behandlung hat
keinen Effekt

Simpson's Paradox

Klinische Medikamentenstudie:

All	Treatment group	
	Control	Treated
Alive	60	200
Dead	60	200
Rate	50%	50%

	Control	Treated
Alive	40	80
Dead	30	50
Rate	43%	38%

Positiver Effekt

Getrennt nach Geschlechtern erkennen wir einen positiven Effekt bei Männern (Abnahme 43% auf 38%).

- > Die Behandlung hat
- > keinen Effekt

Was erwartet ihr nun für Frauen?

Simpson's Paradox

Klinische Medikamentenstudie:

All	Treatment group	
	Control	Treated
Alive	60	200
Dead	60	200
Rate	50%	50%

- > Die Behandlung hat
- > keinen Effekt

	Control	Treated
Alive	40	80
Dead	30	50
Rate	43%	38%

Positiver Effekt

	Control	Treated
Alive	20	120
Dead	30	150
Rate	60%	55%

Positiver Effekt

Einzelpersonen der Gruppen bekommen unterschiedliches Gewicht in Prozent

Waffen retten mehr Leben als sie vernichten

- > Kleck & Gertz (1993) haben in einer Untersuchung herausgefunden, 2.5 Millionen zivile Amerikaner pro Jahr eine Waffe zur Selbstverteidigung nutzen
- > In anderen Worten: 1% der erwachsenen Zivilbevölkerung hat im letzten Jahr eine Waffe zur Selbstverteidigung genutzt
- > **Hört sich nach einer unglaublich hohen Zahl an. Denkt ihr diese ist korrekt?**

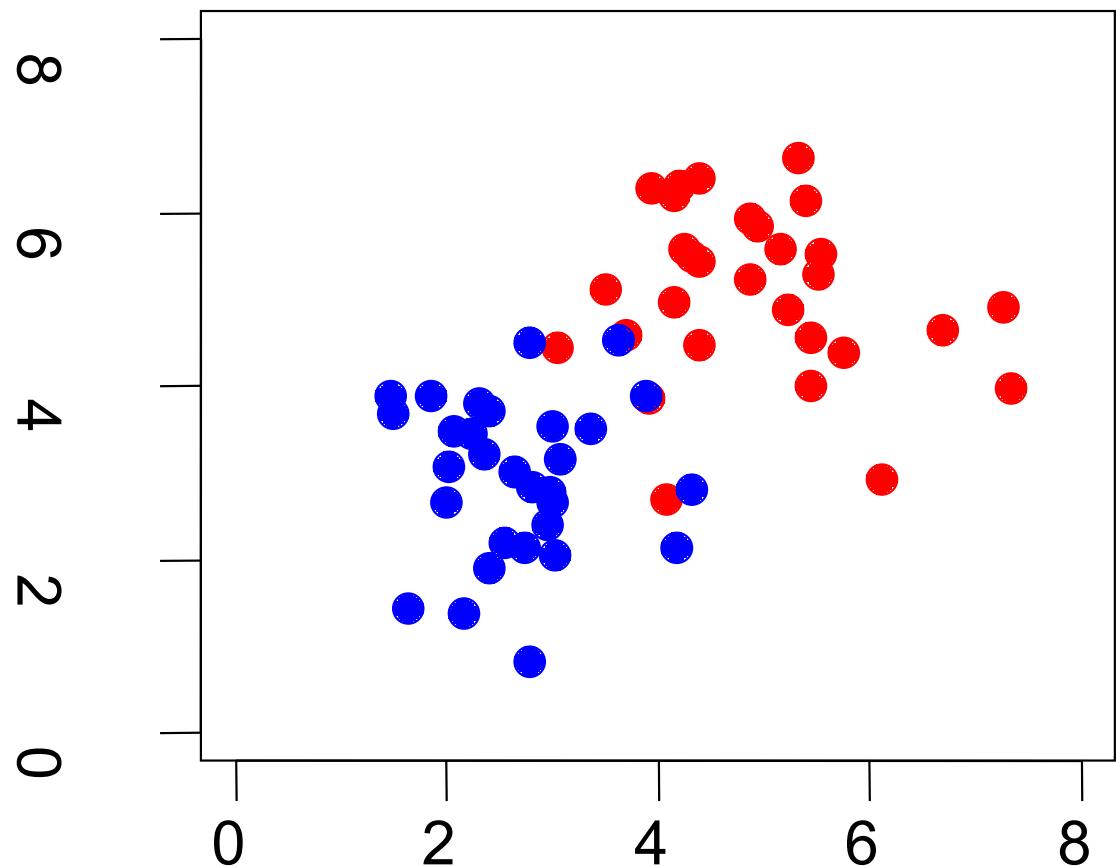
Probleme mit seltenen Ereignissen

- > Umfragen haben eine Tendenz seltene Ereignisse zu überschätzen
- > Nehmen wir an, 1% der Personen gibt eine falsche Antwort, weil z.B. die Frage missverstanden wurde.
- > Bei einem seltenen Ereignis z.B. 0.01% der Befragten antworten mit „Ja“ könnte zu 1.01% werden.
- > Bei einem häufigen Ereignis, bei dem z.B. 50% mit „JA“ antworten, heben sich falsche positive und falsche negative Antworten gegeneinander auf.

Störfaktor

Pseudo-Korrelation

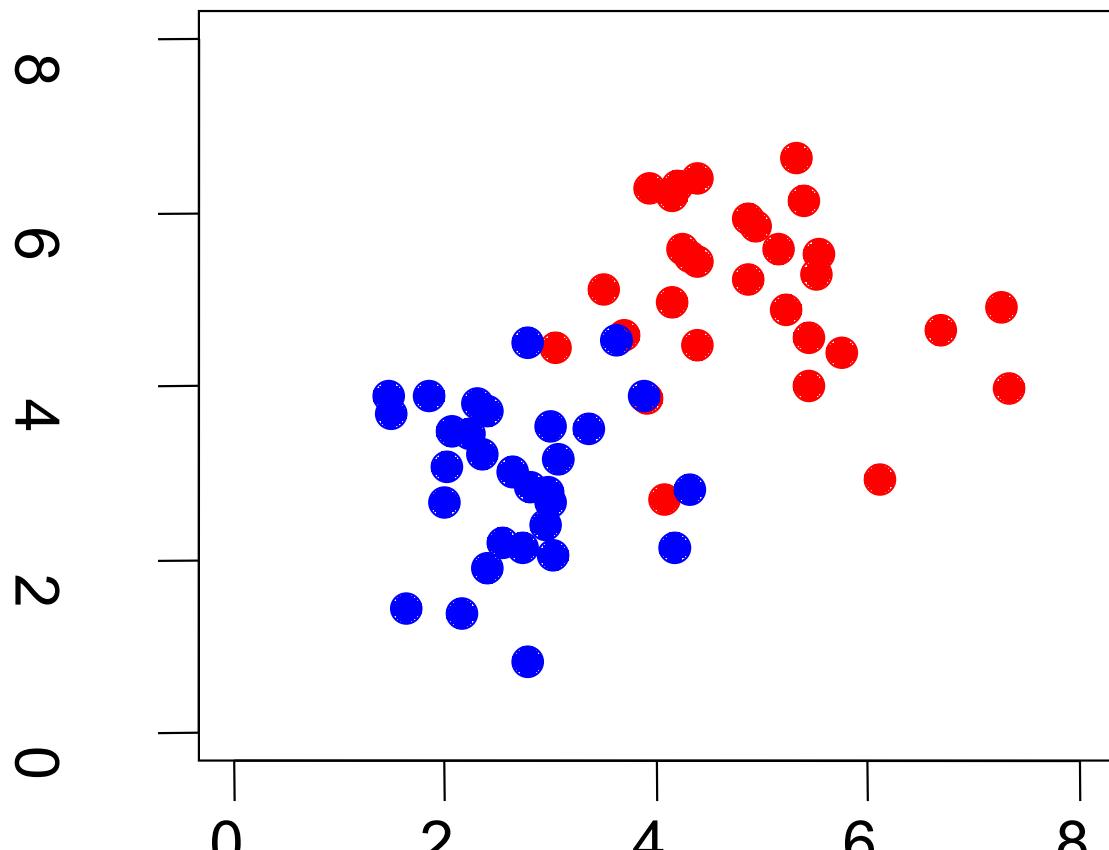
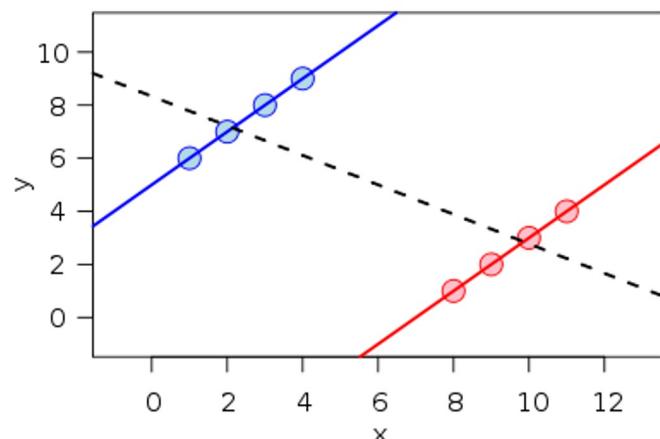
- > nur Männer
 $p = 0.99; r^2 = 0.01$
- > nur Frauen
 $p = 0.34; r^2 = 0.03$
- > beide
 $p < 0.001; r^2 = 0.30$



Störfaktor

Pseudo-Korrelation

- > nur Männer
 $p = 0.99; r^2 = 0.01$
- > nur Frauen
 $p = 0.34; r^2 = 0.03$
- > beide
 $p < 0.001; r^2 = 0.30$



Gegenteiliger Effekt

Falsche Interpretation / Wahrnehmung

Wie gross ist die Wahrscheinlichkeit, dass 2 Personen hier im Raum am gleichen Tag Geburtstag haben?

- > <1%, 1-20%, 20-40%, 40-60%, 60-80%, >80%

Falsche Interpretation / Wahrnehmung

Warum ist die Wahrscheinlichkeit so hoch?

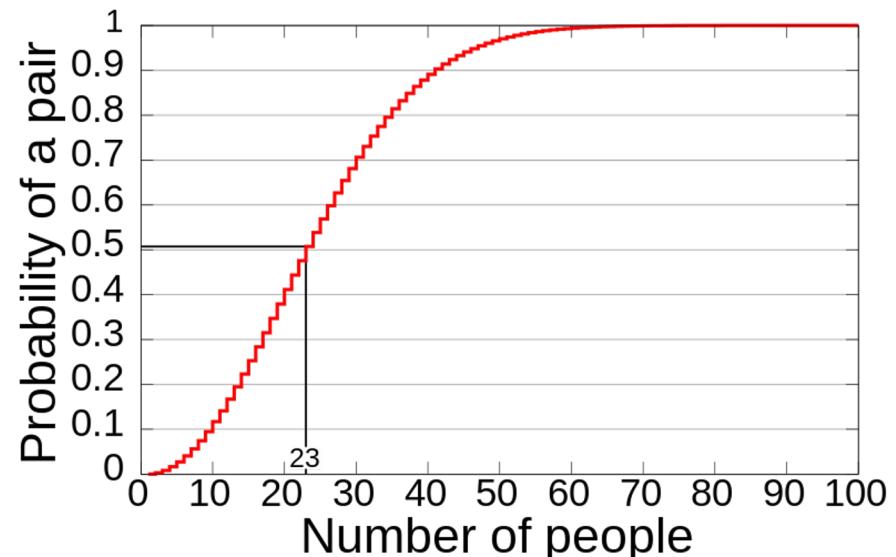
Gefragt:

- > zwei Personen an einem beliebigen Tag im Jahr Geburtstag haben!
- > interpretiert als „wie wahrscheinlich es ist, dass eine Person an einem bestimmten Tag im Jahr Geburtstag hat“

Bei k=23 Personen:

$$P(A) = 1 - \left(\frac{365}{365} \cdot \frac{364}{365} \cdot \dots \cdot \frac{343}{365} = 50.7 \right)$$

$$P(A) = 1 - \frac{365!}{(365 - k)! \cdot 365^k}$$



Ihr fragt nach dem Weg und jemand sagt:

- > „Immer geradeaus und dann nach ca. 2 km links abbiegen“.
- > Eine zweite Person kommt hinzu und sagt, „nein, du musst nach genau 1.82 km rechts abbiegen“.

Biegt ihr rechts oder links ab?

Ihr fragt nach dem Weg und jemand sagt:

- > „Immer geradeaus und dann nach 2 km links abbiegen“.
- > Eine zweite Person kommt hinzu und sagt, „nein, du must nach genau 1.75 km rechts abbiegen“.

- > Dezimalstellen gaukeln uns Präzision vor, die oft gar nicht vorhanden ist!

Beispiel: Datenerhebungstragie

Ihr sollt untersuchen, ob Rauchen Krebs verursacht?

- > 1. Vorschlag: Idealisierte Studie mit 2 Testgruppen, die sich nur durch den Faktor „Rauchen“ unterscheiden!
- > Also zufällig 100 Studenten auswählen, die Hälfte darf die nächsten 20 Jahre nicht eine Zigarette anfassen, die andere Hälfte muss jeden Tag 2 Packungen rauchen. Beim Jubiläum in 20 Jahren schauen wir, wie viele aus jeder Gruppe Krebs bekommen haben!?
- > Statistisch ein gutes Experiment. Ethisch und moralisch vielleicht nicht ganz unumstritten und braucht extrem lange Zeit!

Primärdaten / eigene Datenerhebung

- > Ausserdem wäre die Wahrscheinlichkeit, dass gesunde Personen zum Jubiläum kommen grösser. Vielleicht sind sogar schon welche gestorben und würden so aus der Statistik fallen.

- > **Habt ihr eine bessere Idee?**

Primärdaten / eigene Datenerhebung

- > Ausserdem wäre die Wahrscheinlichkeit, dass gesunde Personen zum Jubiläum kommen grösser. Vielleicht sind sogar schon welche gestorben und würden so aus der Statistik fallen.
- > Habt ihr eine bessere Idee?

Lösung:

- > Raucher vs. Nichtraucher untersuchen!
- > Wir müssen aber aufpassen, weil Raucher auch öfter mehr trinken, ungesunder Essen, nicht in allen Altersgruppen gleich viel rauchen, etc.
- > Wir müssen also andere Faktoren berücksichtigen, die das Ergebnis beeinflussen könnten. **Mit welcher Methode, die wir schon kennengelernt haben, geht dies?**

Datenerhebungstragie



- > Wir müssen andere Faktoren berücksichtigen, die das Ergebnis beeinflussen könnten. Mit welcher Methode, die wir schon kennengelernt haben, geht dies?

Multiple Regression!

- > Wir müssen viel mehr Daten erheben als nur über das Rauchen;
- > zusätzlich anderes Verhalten und Parameter, die Einfluss haben könnten, z.B.: Alter, Geschlecht, Trinkverhalten, Sport,
...

Fallen bei der Datenerhebung



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Einfach in der Theorie

- > man hat eine Urne mit Kugeln und zieht daraus zufällig eine best. Anzahl

Wenn wir z.B. die Bevölkerung untersuchen wollen, NICHT trivial!

Ist die Stichprobe wirklich repräsentativ für die Grundgesamtheit?

- > Leute, die von sich aus an einer Umfrage teilnehmen, haben möglicherweise überdurchschnittlich viel Zeit und sind besonders in einem Thema engagiert und damit nicht repräsentativ
- > Wenn man eine Telefonumfrage macht, muss man z.B. zu allen Zeiten anrufen und vielfach die gleiche Nummer bis sich endlich jemand meldet, statt einfach die nächste Nummer zu probieren
- > Viele weitere Fragen stellen, die mit dem Umfragethema nichts zu tun haben, sondern nur zur Sicherung der Repräsentativität dienen, z.B. Ethnizität, Einkommensklasse, Alter, ...

Fallen bei der Datenerhebung



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Ist die Frage richtig gestellt?

- > Bei einer Umfrage 2002, ob sie für die Todesstrafe für Mörder wären, sagen 60% der Amerikaner JA.
- > Bei einer nächsten Umfrage, ob sie für die Todesstrafe oder lebenslange Haft für Mörder wären, sagten 47% Todesstrafe und 48% lebenslange Haft.

Oft macht das gewählte Wort einen grossen Unterschied:

- > „Klimawandel“ hört sich harmloser an als „Klimakrise“
- > Eine „Steueranpassung“ wird eher akzeptiert als eine „Steuererhöhung“

Sagen die Befragten die Wahrheit?

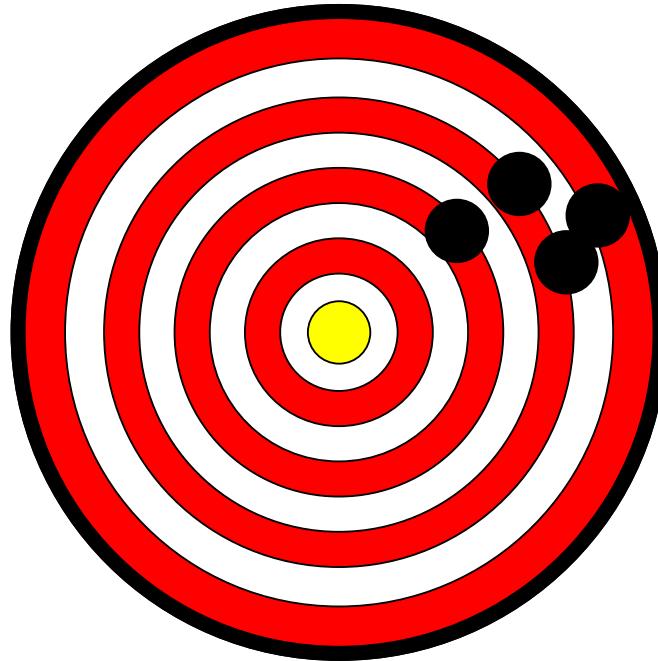
- > unpopuläre, oder sozial nicht akzeptierte Meinungen werden nicht geäussert
- > deshalb fragt man nicht, ob der Befragte eine Meinung hat, sondern z.B. ob er Personen mit dieser Meinung kennt
- > es gibt Übertreibungen wie oft man wählen geht oder wie oft man Sex hat, ...
- > deshalb fragt man nicht nur, ob die Person dieses Jahr plant wählen zu gehen, sondern auch ob sie bei der letzten Wahl war

Bias, systematische Fehler

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



schlechte Genauigkeit
(accuracy)
= systematischer Fehler

Selektions Bias

- > „die AFD kann nicht ins Parlament kommen, ich kenne niemanden, der sie wählen würde“.

Stimmt das?

Selektions Bias

- > „die AFD kann nicht ins Parlament kommen, ich kenne niemanden, der sie wählen würde“.
- > Unsere Freunde sind vermutlich keine repräsentative Stichprobe aus der Bevölkerung

Selektions Bias

- > „die AFD kann nicht ins Parlament kommen, ich kenne niemanden, der sie wählen würde“. Unsere Freunde sind vermutlich keine repräsentative Stichprobe aus der Bevölkerung

Publikations Bias

- > Positive Studien werden meist veröffentlicht, Negative nicht immer. Wird z.B. 20x wird die Wirkung eines Medikaments untersucht und einmal wird eine Wirkung festgestellt (erwartet bei $\alpha=0.05$). In der Literatur erscheint es dann so als würde es nur eine Studie geben, die eine Wirkung hätte.

Selektions Bias

- > „die AFD kann nicht ins Parlament kommen, ich kenne niemanden, der sie wählen würde“. Unsere Freunde sind vermutlich keine repräsentative Stichprobe aus der Bevölkerung

Publikations Bias

- > Positive Studien werden meist veröffentlicht, Negative nicht immer. Wird z.B. 20x wird die Wirkung eines Medikaments untersucht und einmal wird eine Wirkung festgestellt (erwartet bei $\alpha=0.05$). In der Literatur erscheint es dann so als würde es nur eine Studie geben, die eine Wirkung hätte.

Erinnerungs Bias

- > Wenn man Leute mit Krebs fragt, ob sie früher ungesund gegessen haben, findet man einen deutlichen Zusammenhang.
- > **Warum?**

Selektions Bias

- > „die SVP kann nicht ins Parlament kommen, ich kenne niemanden, der sie wählen würde“. Unsere Freunde sind vermutlich keine repräsentative Stichprobe aus der Bevölkerung

Publikations Bias

- > Positive Studien werden meist veröffentlicht, Negative verschwinden in der Schublade, z.B. 20x wird die Wirkung eines Medikaments untersucht und einmal wird eine Wirkung festgestellt (erwartet bei $\alpha=0.05$). In der Literatur erscheint es dann so als würde es nur eine Studie geben und hätte ein Wirkung.

Erinnerungs Bias

- > Wenn man Leute mit Krebs fragt, ob sie früher ungesund gegessen haben, findet man einen deutlichen Zusammenhang.
- > Der liegt aber darin begründet, dass die Betroffenen im Gegensatz zu den Gesunden in ihrer Erinnerung einen Grund für ihre Krankheit suchen.

Überlebenden Bias

- > Wir können beispielweise beobachten, dass die Noten der Studierenden mit jedem Studienjahr besser werden.
- > **Heisst das, die Benotung wäre nicht fair?**

Überlebenden Bias

- > Wir können beispielweise beobachten, dass die Noten der Studierenden mit jedem Studienjahr besser werden.
- > Nein, aber wer feststellt, dass studieren oder Geographie nicht das Richtige ist, wird das Studium abbrechen, bzw. sich auf den Teil spezialisieren, der ihr/ihm besser liegt.

Überlebenden Bias

- > Wir können beispielweise beobachten, dass die Noten der Studierenden mit jeden Jahr besser werden. Heisst das, die Benotung wäre nicht fair?
- > Nein, aber wer feststellt, dass studieren oder Geographie nicht das richtige ist, wird das Studium abbrechen, bzw. sich auf den Teil spezialisieren, der ihr/ihm besser liegt.

Gesunde Nutzer Bias

- > Wer regelmässig Vitaminpillen nimmt, ist meist gesunder.
- > Aber nicht unbedingt wegen der Pillen oder sogar trotz der Pillen, weil sie/er auch sonst auf gesunde Ernährung, Sport, etc. achtet.

Und das sind noch nicht alle Biases

Sackett (1979): (1) the biases of rhetoric (2) the all's well literature bias (3) one sided reference bias (4) positive results bias (5) hot stuff bias (6) popularity bias (7) centripetal bias (8) referral filter bias (9) diagnostic access bias (10) diagnostic suspicion bias (11) unmasking (detection signal) bias (12) mimicry bias (13) previous opinion bias (14) wrong sample size bias (15) admission rate (Berkson) bias (16) prevalence-incidence (Neyman) bias (17) diagnostic vogue bias (18) diagnostic purity bias (19) procedure selection bias (20) missing clinical data bias (21) non-contemporaneous control bias (22) starting time bias (23) unacceptable disease bias (24) migrator bias (25) membership bias (26) non-respondent bias (27) volunteer bias (28) contamination bias (29) withdrawal bias (30) compliance bias (31) therapeutic personality bias (32) bogus control bias (33) insensitive measure bias (34) underlying cause bias (rumination bias) (35) end-digit preference bias (36) apprehension bias (37) unacceptability bias (38) obsequiousness bias (39) expectation bias (40) substitution game (41) family information bias (42) exposure suspicion bias (43) recall bias (44) attention bias (45) instrument bias (46) post-hoc significance bias (47) data dredging bias (looking for the pony) (48) scale degradation bias (49) tidying-up bias (50) repeated peeks bias (51) mistaken identity bias (52) cognitive dissonance bias (53) magnitude bias (54) significance bias (55) correlation bias (56) under-exhaustion bias

Absolutwerte:

- > Heute ist es 10°C wärmer als gestern (gut vorstellbar)
- > Das Müsli enthält 31 mg Kalium pro 100 g (wir wissen vermutlich nicht, ob das viel oder wenig und gut oder schlecht ist)

Relative Werte [%]:

- > oft sinnvoll für Vergleiche
- > können aber auch verwirrend sein:

- > etwas hat 100 CHF gekostet und wird um 25% reduziert = 75 CHF
- > etwas hat 75 CHF gekostet und wird um 25% im Preis erhöht = **93.75 CHF**

ACHTUNG: Prozent vs. Prozentpunkte. Eine Steuer wird von 3 auf 5 % erhöht:

- > Wie würdet ihr das als verantwortliche politische Partei und wie als die politischen Gegner kommunizieren?

- > ACHTUNG: Prozent vs. Prozentpunkte. Eine Steuer wird von 3 auf 5 % erhöht:
- > Die verantwortliche politische Partei wird sagen „die Steuer wurde um zwei Prozentpunkte erhöht“
- > Die politischen Gegner „die Steuer wurde um 67 Prozent erhöht“
- > Beide Aussagen sind korrekt!

Abhängigkeit von Einheit und Zeit:

Einheit:

- > Hotelzimmer in München kostet 150
- > Hotelzimmer in Zürich kostet 190
- > Klar, dass der Wechselkurs CHF/EUR beim Vergleich berücksichtigt werden muss

Zeit:

- > Hollywood macht das nicht, wenn sie erklären, ein Film hätte mit 500 Mio \$US mehr Geld eingespielt als je ein Film zuvor.
- > Was könnte hier ein Problem sein?

Abhängigkeit von Einheit und Zeit:

Einheit:

- > Hotelzimmer in München kostet 150
- > Hotelzimmer in Zürich kostet 190
- > Klar, dass der Wechselkurs CHF/EUR beim Vergleich berücksichtigt werden muss

Zeit:

- > Hollywood macht das nicht, wenn sie erklären, ein Film hätte mit 500Mio \$US mehr Geld eingespielt als je ein Film zuvor.
- > Nach Korrektur für die **Inflation**, war „Gone with the Wind“ der umsatzstärkste Film mit 100 Mio \$US im Jahr 1939.

Fallen für die die Statistik nichts kann



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Statistische Untersuchungen zeigen einen Zusammenhang zwischen der Nutzung von Mobiltelefonen in Autos und Unfällen.
- > Logische Konsequenz war, die Nutzung zu verbieten, damit die Fahrer weniger abgelenkt sind.
- > Jetzt zeigt sich jedoch, dass die Leute ihre Telefone nur besser verstecken und dadurch noch weniger auf die Strasse schauen. Damit führt das Verbot zu noch mehr Unfällen!

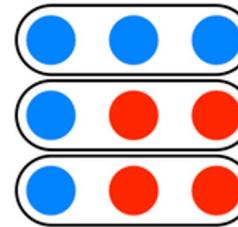
„unbeabsichtigte Folgen“
„unintended consequences“

Statistische Diskriminierung

- > Sollte man Verbrecher nach Statistik jagen? z.B. Personen mit bestimmten äusserlichen Merkmalen öfter auf Drogen untersuchen? Einige werden bestimmt denken, dass ist doch rassistisch und für viele unschuldige Personen extrem ungerecht! Aber falls dies wirklich höhere Aufklärungsquoten und mehr Sicherheit generieren würde? Es gibt Fragen, bei denen Statistik helfen kann, wir aber ethisch/moralisch entscheiden müssen.
- > Sollten Frauen höhere Krankenkassenbeiträge zahlen? Für eine Versicherung vielleicht pure Statistik, aber gesellschaftlich betrachtet Diskriminierung.

Gerrymandering

u^b



b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > oder Wahlkreisschiebung ist die absichtliche Manipulation der Grenzen von Wahlkreisen bei einem Mehrheitswahlsystem. Benannt nach Elbridge Gerry, einem Gouverneur von Massachusetts des frühen 19. Jahrhunderts, dessen Wahlbezirk nach einem Neuzuschnitt – wie ein zeitgenössischer Zeitungskarikaturist bemerkte – einem Salamander glich (Gerry + Salamander)
- Trotz Überzahl der blauen „Stimmen“ entsteht durch den Zuschnitt der drei „Wahlkreise“ eine Mehrheit für die rote „Partei“.



Trash in, trash out

- > Man kann beste Wissenschaft mit simpelster Statistik machen, wenn man sich gute Experimente ausdenkt.
- > Man kann auch beste Wissenschaft mit Daten machen, wenn das Ziel der Studie vor der Datenerhebung noch nicht bekannt war.
- > Aber aus schlechten Daten kann auch die aufwendigste Methode oft nichts sinnvolles herausziehen.

Take-home message

- > Denkt jetzt bitte nicht, jede Statistik sein fehlerhaft
- > Gerade in der Wissenschaft wird versucht, mit Statistik Hypothesen durch ein zusätzliches Argument zu untermauern, indem man die Wahrscheinlichkeit prüft, dieses Ergebnis per Zufall erhalten zu haben
- > Auch andere Quellen wie Bundesämter liefern qualitative hochwertige Statistiken
- > Aber schaut genau auf Begriffe, Zahlen und wie sie erhoben wurden, wenn sie von nicht vertrauenswürdigen/parteiischen Quellen stammen

Prüfung

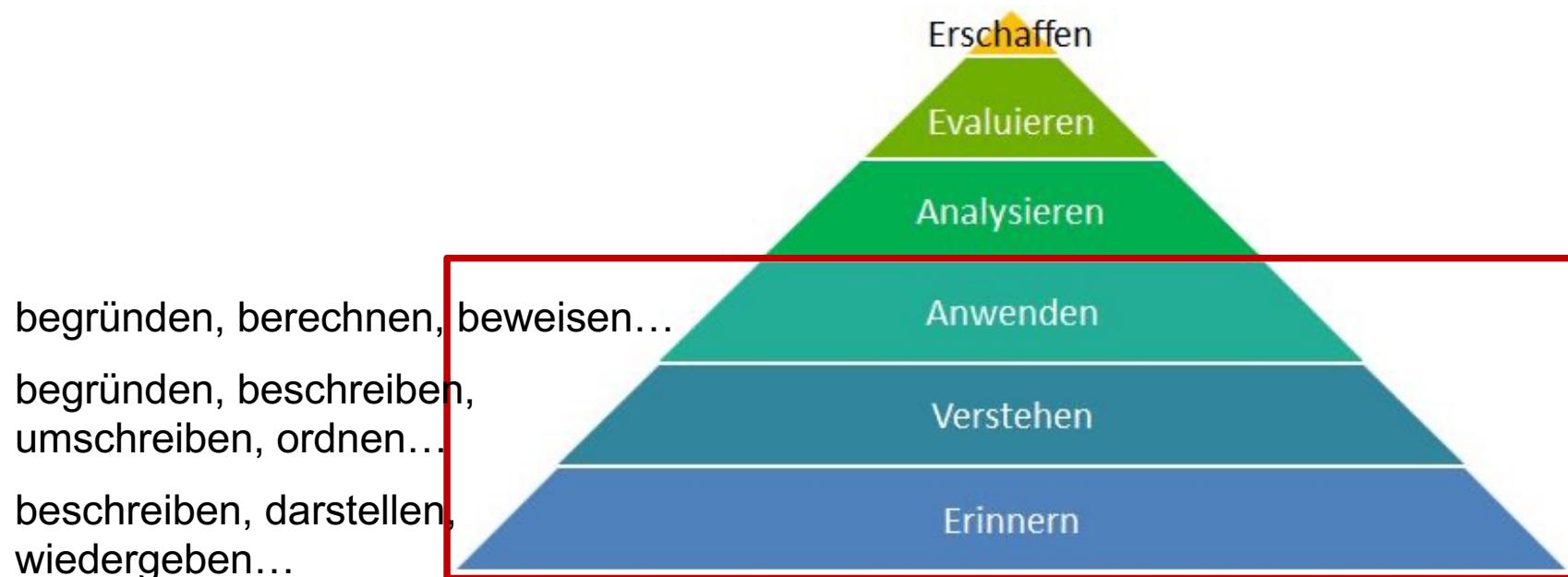
- > IliasExam mit eigenem Computer
- > 60 Minuten, ca. 20 Fragen
- > Keine Minuspunkte für nicht oder falsch beantwortete Fragen
- > Zum Bestehen der Prüfung: 60 % der Maximalpunktzahl
- > Open Book, aber keine Zeit, um alles nachzuschlagen und im Prüfungsraum nicht unbegrenzt Platz ohne andere zu stören
- > siehe Prodädeutikum (Modul 18)
- > **Link zur Prüfung:**
https://exam.unibe.ch/goto.php?target=crs_6394_rcodeB8qQuCR9WW

Prüfungbedingungen

- > Diese Prüfung darf nicht kopiert, fotografiert oder in anderer Weise gespeichert werden. Sie darf auch nicht verbreitet werden.
- > Auf Ihrem Gerät dürfen Sie die Prüfung in nur einem Browserfenster bzw. -tab geöffnet haben.
- > Es ist untersagt, Geräte (z.B. Mobiltelefon/SmartWatch) und Dienste (soziale Medien/Messenger) zu verwenden, die eine Kommunikation mit anderen oder die Übertragung von Texten oder Daten an andere ermöglichen.
- > Antworten können auf Papier oder in einem Worddokument gesichert werden.

Prüfungsaufgaben

- > single choice, multiple choice, kPrim
- > Rechenaufgabe
- > Abbildung interpretieren
- > R-Output interpretieren



Beispiel Prüfungsfragen

Psychologen messen den Einfluss von Schule und Geschlecht auf den Intelligenzquotienten an privaten und öffentlichen. In ihrem Bericht schreiben sie: *“Der mittlere Intelligenzquotient aller Schüler/innen war 108.27. Es gab keinen signifikanten Unterschied zwischen Mädchen und Jungen ($N = 532, t = 0.521, p = 0.631$). Aber wir haben einen statistisch signifikanten Unterschied zwischen privaten und öffentlichen Schulen identifiziert ($N = 532, t = 2.172, p = 0.032$).”*

Warum sind diese Resultate schwer zu beurteilen?

- Es fehlt die Information, ob der Intelligenzquotient an privaten oder öffentlichen Schulen höher ist. RICHTIG / FALSCH
- Bei so einer derart Stichprobe ($N=532$), zeigt die statische Signifikanz klar auf, dass es einen in der Praxis relevanten Unterschied gibt. RICHTIG / FALSCH
- Es fehlt die Information, wie viele Mädchen und wie viele Jungen in der Studie sind. RICHTIG / FALSCH
- Es fehlt die Information, wie gross der Unterschied des Intelligenzquotienten zwischen privaten und öffentlichen Schulen ist. RICHTIG / FALSCH

Beispiel Prüfungsfragen

Interpretiert die R Ausgabe eines linearen Regressionsmodells:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 348.356477 12.327774  28.26 <2e-16 ***
mort$Year   -0.174834  0.006324 -27.65 <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Residual standard error: 3.253 on 145 degrees of freedom
Multiple R-squared:  0.8406,    Adjusted R-squared:  0.8395
F-statistic: 764.4 on 1 and 145 DF,  p-value: < 2.2e-16
```

Weicht die Steigung der Regressionsgerade signifikant ($\alpha < 0.001$) von Null ab?
JA / NEIN

Ist die Steigung der Regressionsgerade positiv oder negativ
JA / NEIN

Beispiel Prüfungsfragen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Hier seht ihr jährliche Erhebungen der Kindersterblichkeit in der Schweiz. Schätzt den Regressionskoeffizienten des linearen Trends über die Zeit. Gebt maximal zwei Dezimalstellen an.

