

# Vorlesung: Statistik I für Studierende der Statistik, Mathematik & Informatik

---

Dozent: Fabian Scheipl

Material: H. Küchenhoff

**LMU München**

## **Homepage:**

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

## **Vorlesung:** *Dr. Fabian Scheipl*

Mo. 12-14h, Hauptgebäude D 209 (nicht jede Woche!)

Mi. 12-14h, Hauptgebäude E 004

## **Übung:** *Elizabeth Heller, M.A., M.Sc.*

Do. 10-12h (Gruppe 1) / 12-14h (Gruppe 2), Hauptgebäude D 209  
(ca. 14-tägig)

## **Tutorium:** *Johanna Völkl, B.Sc.*

Mi. 8-10h, Hauptgebäude A 014

(ca. 14-tägig)

Schwesterveranstaltung: **Einführung in die statistische Software**

*Dr. Cornelia Oberhauser, Dipl. Stat. Micha Schneider*

Mo. 12-14h, Hauptgebäude D 209

am 31.10., 21.11., 28.11., 5.12.

Übung 12.-18.12.

Hausübung 19.12. - 20.1.

(verpflichtend für Studierende der Statistik)

## Programmpaket R

- Frei verfügbar, *open source*: [r-project.org](http://r-project.org)
- Kurse im Rahmen der Veranstaltung “**Statistische Software**”
- *lingua franca* der Statistik
  - (fast) jede denkbare statistische Methode implementiert
  - aktuelle Forschung geht in Form von Zusatzpaketen (“packages”) ein
  - immer wichtiger / etablierter auch in Wirtschaft, Verwaltung, ...
- Obacht, steile Lernkurve: befehlsbasiert, keine GUI!
- Editor **RStudio** ([rstudio.com](http://rstudio.com), frei verfügbar)
- Interaktives R-Kursprogramm: ([swirl](#), auf englisch)

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

**Statistik - Der Weg zur Datenanalyse**

Springer-Verlag, 7. Auflage, 2009

**H.Toutenburg, C.Heumann:**

**Deskriptive Statistik - Eine Einführung in Methoden und  
Anwendungen mit R und SPSS**

Springer-Verlag, 2009

(s.a. Moodle-Seite)

- 6 ECTS
- Prüfungsleistung **je nach HF**:
  - **Statistik**: Klausur + \*-Übungsaufgaben (> 50%) + Hausübung bestanden
  - **Informatik**: Klausur + \*-Übungsaufgaben (> 50%)
  - **Mathematik**: Klausur
- Klausur am Ende des Semesters (8.2.2017 14:00h, 90 min):
  - Taschenrechner
  - *open book*: alle Unterlagen, Bücher, Notizen, ... zugelassen
- Note = Klausurnote

1. Einführung, Beispiele, Geschichte
2. Grundlagen der Datenerhebung: Messung, Skalenniveaus
3. Typen von Studien, Auswahlverfahren
4. Univariate deskriptive Statistik 1: Häufigkeiten und graphische Darstellungen, kumulierte Verteilung, Lage- und Streuungsparameter
5. Univariate deskriptive Statistik 2: Boxplots, Schiefe und Wölbung, Kerndichteschätzung
6. Kontingenztafeln, Zusammenhangsmaße für nominalskalierte Merkmale, Plots
7. Rangkorrelationskoeffizient, Zusammenhänge bei quantitativen Merkmalen, Bravais-Pearson-Korrelationskoeffizient
8. Kendall's Tau, Invarianzeigenschaften
9. Grafische Darstellung von Zusammenhängen
10. Regression (lineare Einfachregression, Streuungserlegung)
11. Multiple Regression, Scheinkorrelation
12. Interaktive Grafik

Einführung: Was ist Statistik?

Datenerhebung und Messung

# Einführung: Was ist Statistik?

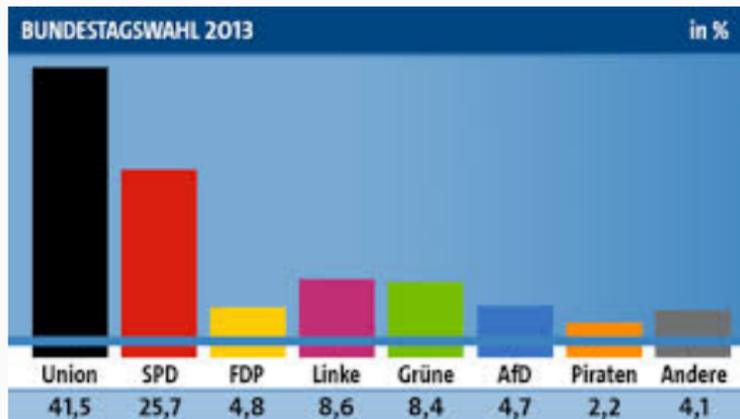
---

# Beispiel 1: Bundestagswahl 2013

## Prognose 18:00 Infratest Dimap (ARD):

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

## Ergebnis:



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

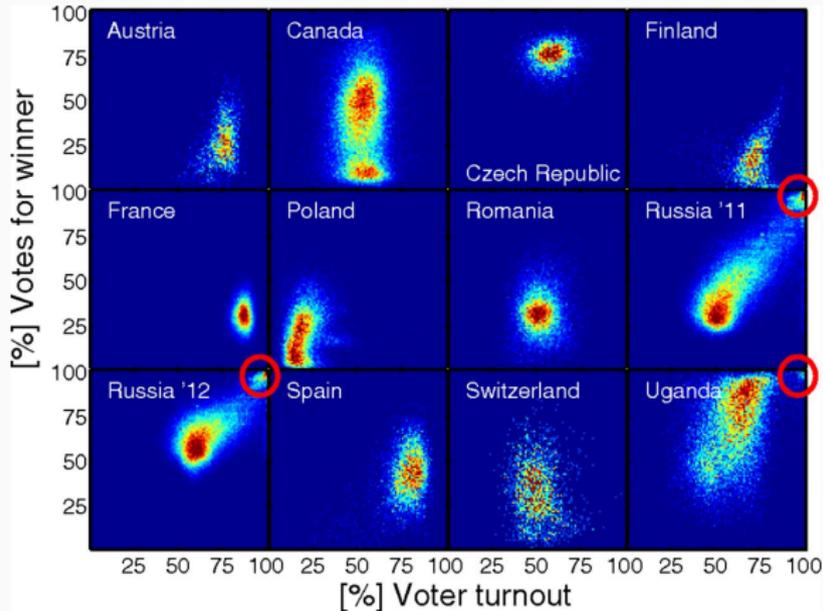
Basis: Nachwahlbefragung 100 000 Wahlberechtigte

Ziele:

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

## Beispiel 2: Wahlfälschung

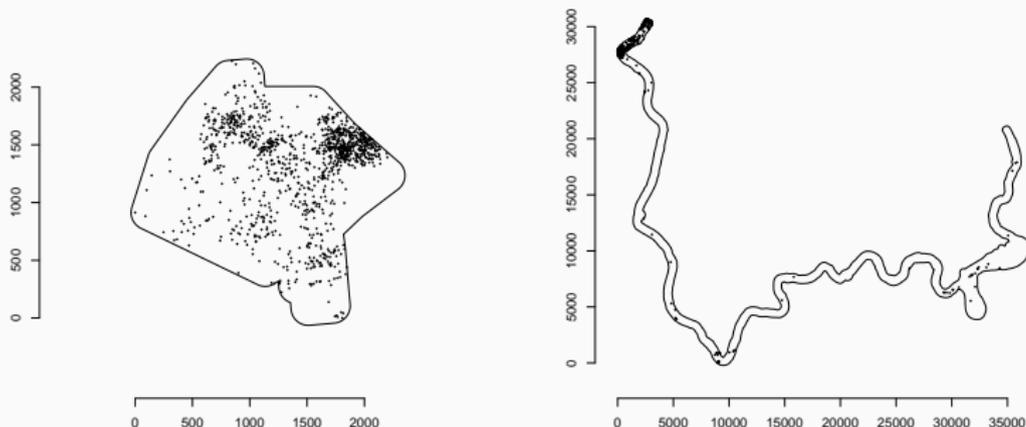
Einfache Idee: Untersuche Zusammenhang zwischen Wahlergebnis (Stimmenanteil des Siegers) gegen die Wahlbeteiligung.



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

# Beispiel 3: Analyse von Daten zu Bombentrümmern

StaBLab-Projekt mit H. Küchenhoff, M. Höhle und M. Mahling

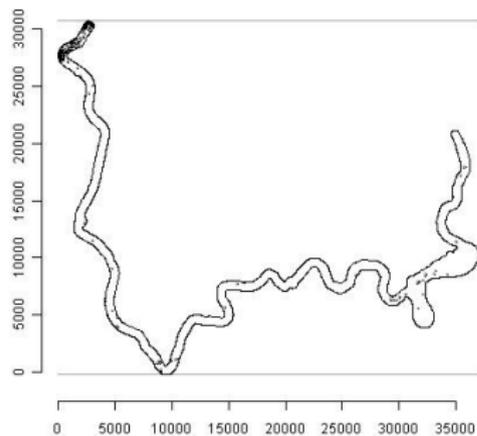
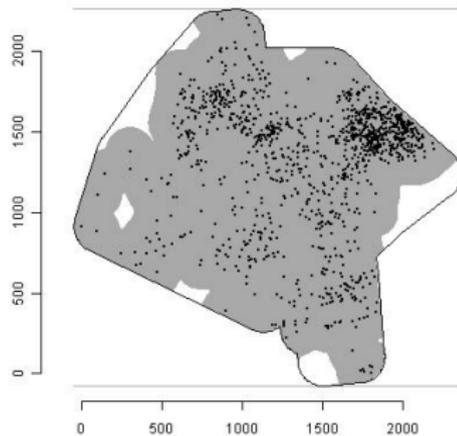


M. Mahling, M. Hoehle, H. Küchenhoff (2013). Determining high-risk zones for unexploded World War II bombs by using point process methodology. *Journal of the Royal Statistical Society Series C-Applied Statistics* 62(2):181-199.

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

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

Sicherheitszonen:



### Gibt es eine Midlife Crisis?

- Analysen von Panel-Daten zur subjektiven Lebenszufriedenheit mit semiparametrischen Regressionsmodellen.
- Forschungsarbeit von Andrea Wiencierz mit Helmut Küchenhoff, Sonja Greven, Christoph Wunder

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

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

# Datengrundlage

- Daten stammen aus den Haushaltsstichproben A (Westdeutsche) und C (Ostdeutsche) des Sozio-ökonomischen Panels (SOEP)
- für die ausgewählten Modellvariablen liegen Beobachtungen aus den Jahren 1992, 1994 bis 2006 vor
- durchschnittliche Anzahl von Beobachtungen pro Person: 7.8
- in die Modellberechnungen gingen 102 708 vollständige Beobachtungen von 13 224 Individuen ein
- Anzahl Beobachtungen pro Jahr:

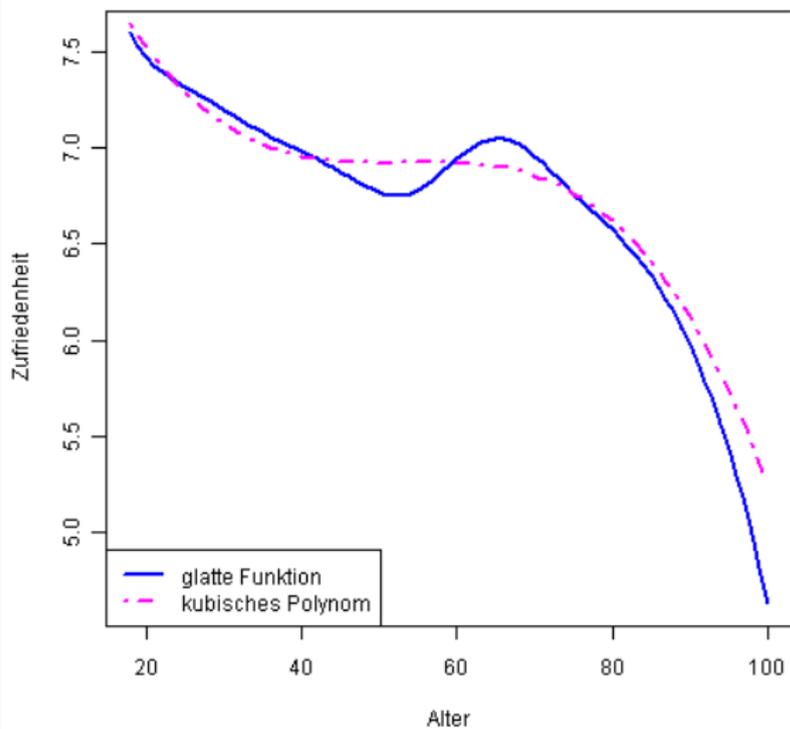
<b>1992</b>	<b>1994</b>	<b>1995</b>	<b>1996</b>	<b>1997</b>	<b>1998</b>	<b>1999</b>
8 145	7 720	7 943	7 606	8 052	7 550	7 403

<b>2000</b>	<b>2001</b>	<b>2002</b>	<b>2003</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>
7 628	7 092	7 068	7 000	6 876	6 543	6 082

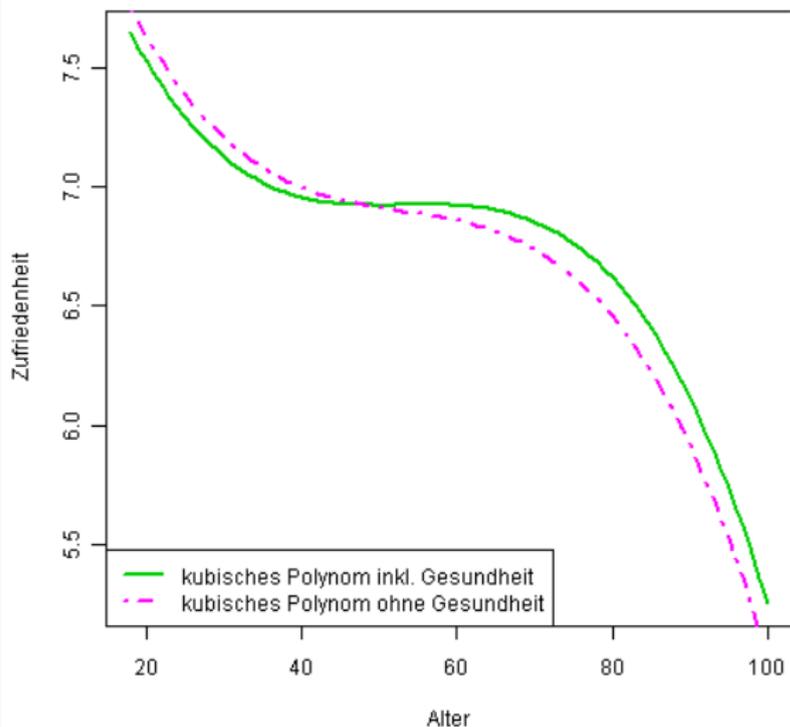
# Ergebnis für Alterseffekt

geschätzte Funktion inkl. AR(1) für \*Durchschnittsmensch\*



# Ergebnisse ohne Gesundheitsvariable

geschätzte Funktion für \*Durchschnittsmensch\* inkl. AR(1)



- Zusammenhänge analysieren
- Komplexe Einflüsse
- Flexibles Modell

## Beispiel 5: Mineralwasserstudie

Studie in Zusammenarbeit mit Prof. Adam (LMU)

**Fragestellung:** Schmeckt mit Sauerstoff angereichertes Mineralwasser besser als gewöhnliches Mineralwasser ?

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

Ergebnis (Clausnitzer et al., 2004) :

- Placebo: 76% gaben an, dass das zweite Wasser anders schmeckt
- Verum : 89 % gaben an, dass das zweite Wasser anders schmeckt

Signifikanter Effekt → Zulassung

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

## Beispiel 6: Umweltzone und Feinstaubbelastung

Wirkt die Umweltzone?

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

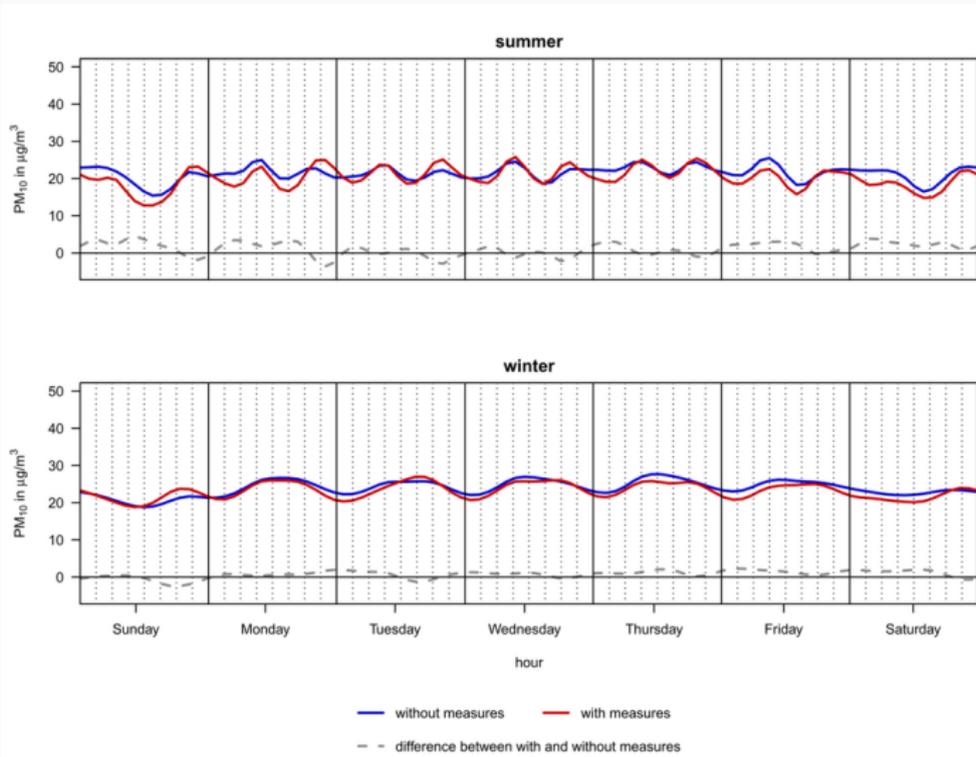
Probleme:

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

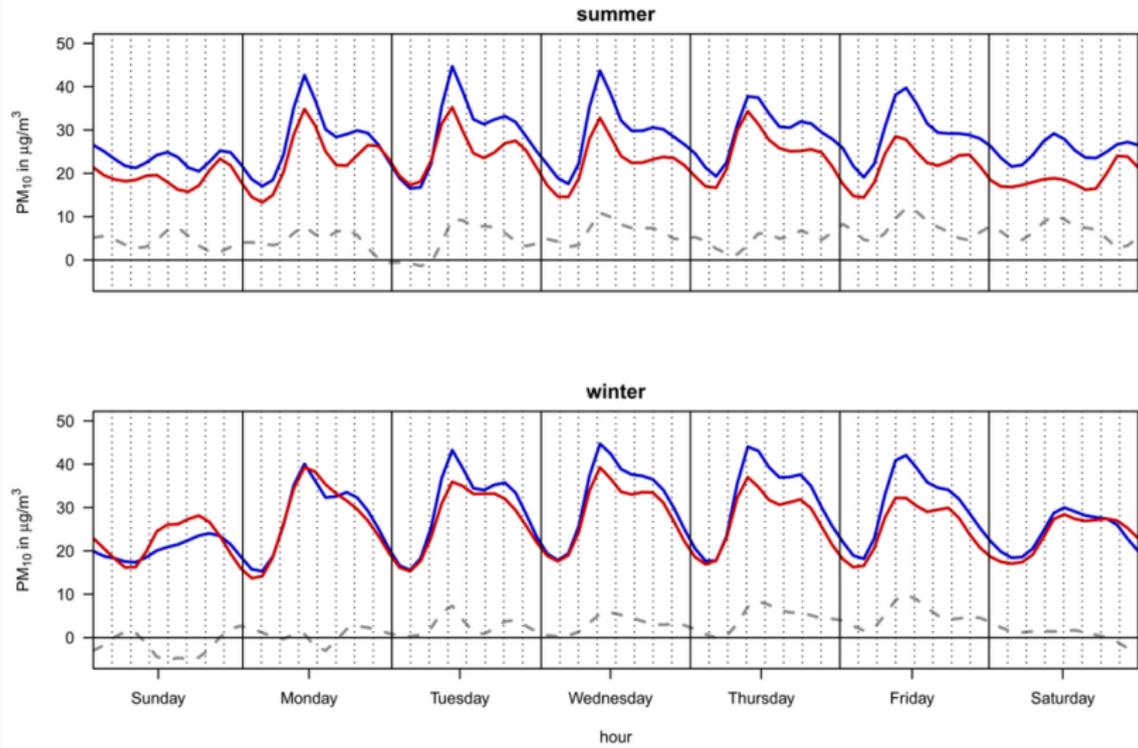
Daher: Regressionsmodell mit Referenzstation, Wetter, Tagesverlauf

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

# Wirkung der Umweltzone: Lothstrasse



# Wirkung der Umweltzone: Prinzregentenstrasse



# Weitere Beispiele

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

- A-B-Testing von Websitedesigns, User Interfaces
- Zugriffsstatistiken auf Webserver
- Ressourcenplanung von Netzwerken und Servern

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

Teilgebiete:

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

- *Statistics is the grammar of science.*  
Karl Pearson
- *Statistics: the mathematical theory of ignorance.*  
Morris Kline
- *Math is a language that you use to describe statistics, but really it's about collecting information and putting it in an order that makes sense.*  
Lauren Stamile
- *Statistics is the science of learning from experience.*  
Bradley Efron
  
- ⇒ **Statistische Methodik als unersetzliches Werkzeug empirischer Wissenschaft.**

## Zitate:

- *Statistics is a body of methods for making wise decisions in the face of uncertainty.*  
W. Allen Wallis
- *Cognitive psychology tells us that the unaided human mind is vulnerable to many fallacies and illusions because of its reliance on its memory for vivid anecdotes rather than systematic statistics.*  
Steven Pinker
- *It is the mark of a truly intelligent person to be moved by statistics.*  
George Bernard Shaw
- *Statistik ist für mich das Informationsmittel der Mündigen. Wer mit ihr umgehen kann, ist weniger leicht zu manipulieren. Der Satz "Mit Statistik kann man alles beweisen" gilt nur für die Bequemen, die keine Lust haben, genau hinzusehen.*  
Elisabeth Noelle-Neumann
- ⇒ **Statistisches Denken als Mittel um Irrationalität zu verringern und mit Unsicherheit umzugehen.**

## Zitate:

- *All models are wrong, but some are useful.*  
George E. P. Box
- *Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.*  
Aaron Levenstein
- *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*  
John Tukey
- *If you torture the data enough, nature will always confess.*  
Ronald Coase
- *There are no routine statistical questions, only questionable statistical routines.*  
David R. Cox
- ⇒ **Statistik ist komplex und ihre Ergebnisse werden oft missbraucht oder missinterpretiert.**

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

Dennis Prager

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

Hilaire Belloc

- *The best thing about being a statistician is that you get to play in everyone's backyard.*

John Tukey

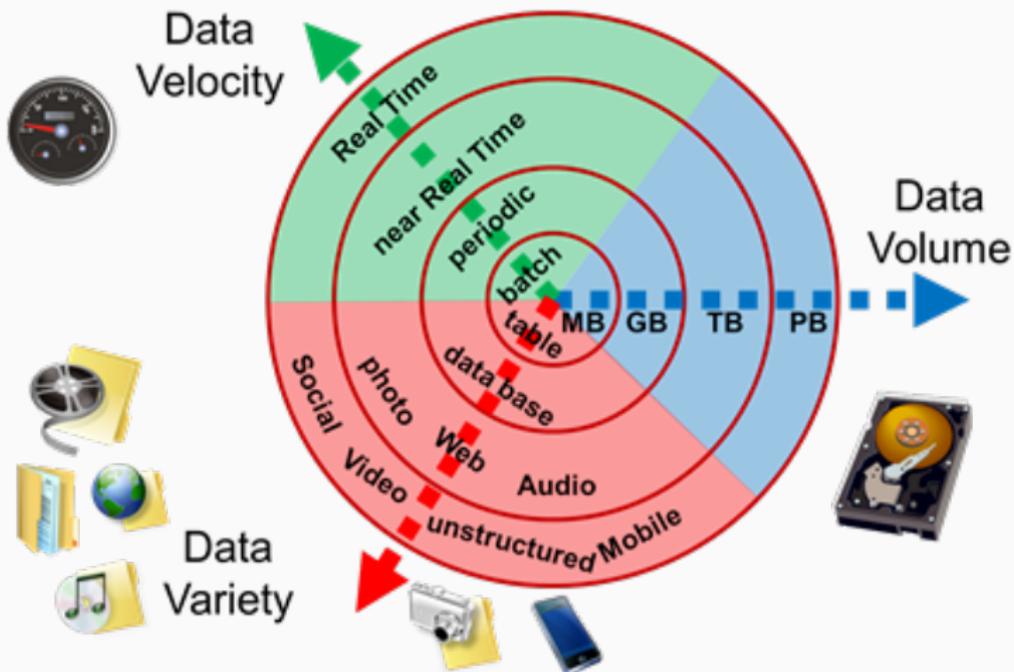
- *I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding.*

Hal Varian

# “Big Data”

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

# “Big Data”



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

**John Graunt (1620 - 1674):**

Sterbetafeln der Stadt London nach Ausbruch der Pest (Demographie)

**Sir William Petty (1623 - 1687):**

“Politische Arithmetik”

Daten für die Verwaltungsreform in Irland

- **Gottfried Achenwall (1719 - 1722):**  
Begriff “‘Statistik’” (= den Staat betreffend)  
“Fach, das sich mit allerlei Staatsmerkwürdigkeiten beschäftigt.”

**Sir Thomas Bayes (1702 - 1761):**

Wichtige theoretische Grundlagen



**Pierre-Simon Laplace (1749 - 1822):**

Wahrscheinlichkeitsrechnung



## John Snow (1813-1858)

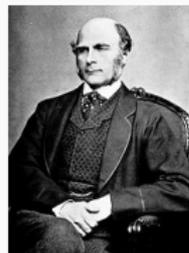


## Cholera in London 1854



**Francis Galton (1822 - 1911):**

Grundlagen der Regression



**Karl Pearson (1857 - 1936):**

Theorie der statistischen Tests

(gemeinsam mit J.Neyman)

**R.A. Fisher (1890 - 1962):**

Grundkonzepte der statistischen Inferenz

Maximum-Likelihood-Prinzip

Statistische Tests



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

# Neuere Entwicklungen

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

## Emil Julius Gumbel (München 1891- New York 1966)



Wichtige Beiträge zur Extremwertstatistik (1958)

## **Generalized Linear Models**

Insgesamt 5299 mal zitiert (Stand Sep '16)

## **Spatial interaction and the statistical analysis of lattice systems**

- Insgesamt 5693 mal zitiert (Stand Sep '16)
- Grundidee der räumlichen Statistik

## Buch : **Generalized Additive Models**

- Insgesamt 12242 mal zitiert (Stand Sep '16)
- Grundidee der flexiblen nichtlinearen Regression

Buch : **An Introduction to the Bootstrap**

- Insgesamt 32202 mal zitiert (Stand Sep '16)
- Bootstrap: Quantifikation von Unsicherheit ohne viele Annahmen

Buch: **Markov Chain Monte Carlo in Practice**

- Popularisierung der Bayesianischen Statistik

Buch: **The Elements of Statistical Learning**

- Machine Learning, Data Mining, Künstliche Intelligenz, etc. aus Sicht der Statistik

# Datenerhebung und Messung

---

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

“Measurement is the contact of reason with nature.”

Henry Margenau (1959)

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

Stanley S. Stevens (1951)

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

- Physikalische Messungen: Gewicht, Blutdruck, Fettaufnahme, ...
- Psychometrische Beispiele: Intelligenz, Gewaltbereitschaft, ...
- Wirtschaftswissenschaftliche Beispiele: Inflation, Bruttosozialprodukt, ...

# Definition

Marco  $\longrightarrow$  1.84

Anna  $\longrightarrow$  1.61

Laura  $\longrightarrow$  1.72

Merkmal definiert Relation (Struktur) zwischen den Objekten.

Messung: strukturerhaltende Abbildung (Homomorphismus)

Bsp: Anna ist kleiner als Laura  $\Leftrightarrow 1.61 < 1.72$

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

Homomorphe Abbildung

empirisches Relativ  $\Rightarrow$  numerisches Relativ

## **Existenz:**

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

$\Rightarrow$  Axiome von Repräsentationstheoremen müssen erfüllt sein (z.B. Transitivität o.ä.)

## **Eindeutigkeit:**

Gibt es mehrere zulässige Skalen?

(z.B. Länge in cm, m; Temperatur in C, F oder K)

$\Rightarrow$  Zulässige (strukturerhaltende) Transformationen

## Skalentypen (Messniveaus)

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

- Existenz ist nicht immer gegeben
- Eindimensionalität könnte verletzt sein

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

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

# Ordinal- oder Rangskala

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

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

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

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

## Intervallskala mit (natürlichem) Nullpunkt

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

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

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

# Skalenniveau

Beachte:

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

Sinnvoll interpretierbare Berechnungen:

Skalenart	sinnvoll interpretierbare Berechnungen			
	auszählen	ordnen	Differenzen bilden	Quotienten bilden
Nominal	✓	X	X	X
Ordinal	✓	✓	X	X
Intervall	✓	✓	✓	X
Verhältnis	✓	✓	✓	✓

Beispiel: Konzentration von Bakterien

$$\begin{array}{ll} 3 \cdot 10^{-3} & \log(3 \cdot 10^{-3}) \approx -2.5 \\ 2 \cdot 10^{-4} & \text{oder } \log(2 \cdot 10^{-4}) \approx -3.6 \\ 2.2 \cdot 10^{-5} & \log(2.2 \cdot 10^{-5}) \approx -10.7 \end{array}$$

Skalenwahl  $\Leftrightarrow$  Interpretation der Differenz

Bei log-Skala: Differenz =  $\log(\text{Faktor der Veränderung})$

Verwende  $\log$  zur Basis 10

# Gütekriterien: Genauigkeit

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

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

$U$  im Mittel = 0 (unsystematischer Fehler, unverzerrte Vermessung)

⇒ klassischer additiver Messfehler

Modell der klassischen (psychologischen) Testtheorie

Aspekte:

- Validität = Gültigkeit
- Reliabilität = Zuverlässigkeit

Frage: Wird das gemessen, was gemessen werden soll?

In der emp. Sozialforschung:

- Inhaltsvalidität
- Kriteriumsvalidität
- Konstruktvalidität

Statistik: systematischer Messfehler?  $U$  im Mittel = 0

Ist die Messung zuverlässig?

Erhält man bei Wiederholung den gleichen Wert?

Erhält man unter verschiedenen Bedingungen den gleichen Wert? →  
Interrater-Reliabilität

Abhängig von:

Verhältnis von Streuung des Messfehlers zur Gesamtstreuung

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

$\sigma_X^2, \sigma_{X^*}^2$  Varianz von  $X$  bzw.  $X^*$

→ Interne Konsistenz: Cronbachs Alpha (später)

Zusammenfassung von Einzelindikatoren zu einer neuen Variablen

Häufig: Bildung von (gewichteten) Summen von einzelnen Variablen

Beispiel:

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

Indexbildung folgt nur theoretischen Vorgaben und fachspezifischen Überlegungen

Fragen der Statistik:

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

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

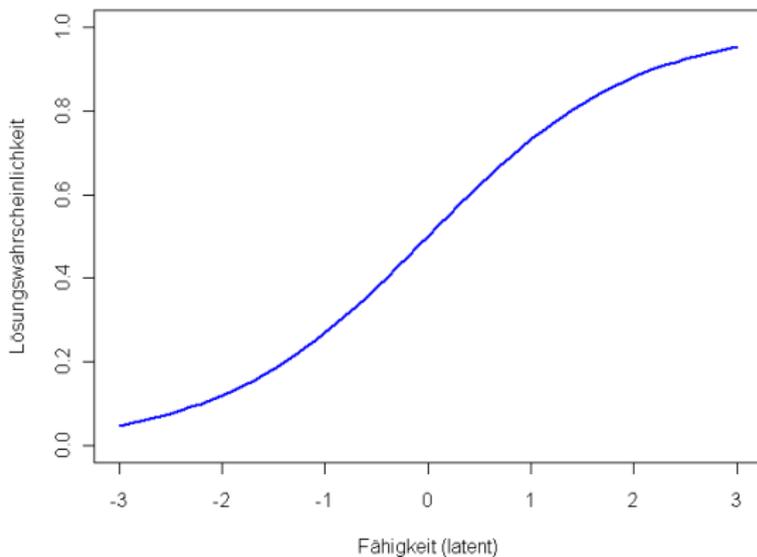
Wichtigste Beispiele :

- Likert Skalen
- Faktorenmodell
- Rasch - Modell

# Itemcharakteristik

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

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



# Ausblick: Statistische Verfahren zur Überprüfung der Messung (Skalierung)

- Itemanalyse
- Rasch-Modell
- Faktorenanalyse
- Analyse von Wiederholungs- und Mehrfachmessungen
- Messfähigkeitsanalyse in der Qualitätskontrolle

Weiter: Verfahren zur **Berücksichtigung** von Messfehlern

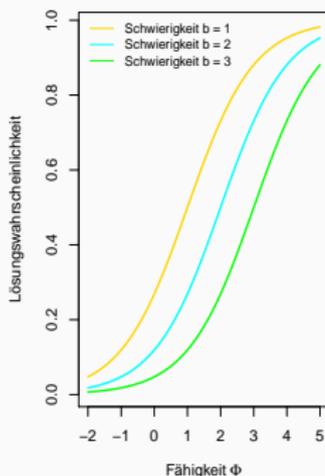
Items haben verschiedene Schwierigkeiten

Lösungswahrscheinlichkeiten lassen sich durch Personenfähigkeit und Itemschwierigkeit beschreiben:

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

$\Phi_i$ : Fähigkeit von Person  $i$

$b_j$ : Schwierigkeit von Item  $j$



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

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

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

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

Induktive Statistik in der Regel nur mit zufälliger Stichprobe möglich!

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

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

Es werden in der Regel verschiedene “Behandlungen” verglichen

Gemeinsamkeit: Experimenteller Eingriff ins Geschehen

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