Deskriptive Statistik und Wahrscheinlichkeitsrechnung

Vorlesung an der Universität des Saarlandes

PD Dr. Martin Becker

Sommersemester 2020

Organisatorisches I

- Vorlesung (nach Wiederaufnahme des Präsenzbetriebs): Freitag, 12-14 Uhr, Gebäude B4 1, Audimax (HS 0.01)
- Übungen (nach Wiederaufnahme des Präsenzbetriebs): nach gesonderter Ankündigung (siehe Homepage)
- Prüfung: *voraussichtlich* 2-stündige Klausur nach Semesterende (1. Prüfungszeitraum)
  Anmeldung und Informationen zum Termin im ViPa
- Hilfsmittel für Klausur
  - „Moderat" programmierbarer Taschenrechner, auch mit Grafikfähigkeit
  - 2 *beliebig gestaltete* DIN A 4–Blätter (bzw. 4, falls nur einseitig)
  - Benötigte Tabellen werden gestellt, aber **keine weitere Formelsammlung!**
- Durchgefallen — was dann?
  - „Wiederholungskurs" im kommenden (Winter-)Semester
  - „Nachprüfung" (voraussichtlich) erst März/April 2021 (2. Prüfungszeitraum)
  - „Reguläre" Vorlesung/Übungen wieder im Sommersemester 2021

Organisatorisches II

- Informationen und Materialien unter

  `https://www.lehrstab-statistik.de`

  bzw. spezieller

  `https://www.lehrstab-statistik.de/deskrwrss2020.html`

  (bei Problemen `https://www2.lehrstab-statistik.de` versuchen!)
- Im Präsenzbetrieb:

  Kontakt: PD Dr. Martin Becker
  Geb. C3 1, 2. OG, Zi. 2.17
  e-Mail: `martin.becker@mx.uni-saarland.de`
- Sprechstunde (nach Wiederaufnahme des Präsenzbetriebs) nach Vereinbarung (Terminabstimmung per e-Mail)
- Vorlesungsunterlagen
  - Vorlesungsfolien
  - Zusätzliche digitale Lehrmaterialien (je nach Dauer des Notbetriebs):
    - Online-Skript (wird nach und nach ausgebaut)
    - eventuell weitere digitale Lehrmaterialien

Organisatorisches III

- Übungsunterlagen
  - Übungsblätter (im Präsenzbetrieb wöchentlich, vorher unregelmäßiger)
  - Ergebnisse *(keine Musterlösungen!)* zu den meisten Aufgaben
  - **Im Präsenzbetrieb:** Besprechung der Übungsblätter mit ausführlicheren Lösungsvorschlägen in den Übungsgruppen der folgenden Woche
  - **Im Notbetrieb:** Lösungen (ca. eine Woche nach Übungsblättern) online verfügbar.
  - **Übungsaufgaben sollten – auch im Notbetrieb – unbedingt (vor dem Studieren der Lösungen) selbst bearbeitet werden!**
- Alte Klausuren
  - Aktuelle Klausuren inklusive der meisten Ergebnisse unter „Klausuren" auf Homepage des Lehrstabs verfügbar
  - Prüfungsrelevant sind (natürlich) alle in Vorlesung und Übungsgruppen behandelten Inhalte, nicht nur die Inhalte der Altklausuren!

Was ist eigentlich „Statistik“?

- Der Begriff „Statistik“ hat verschiedene Bedeutungen, insbesondere:
  - Oberbegriff für die Gesamtheit der Methoden, die für die Erhebung und Verarbeitung empirischer Informationen relevant sind ($\rightarrow$ statistische Methodenlehre)
  - (Konkrete) Tabellarische oder grafische Darstellung von Daten
  - (Konkrete) Abbildungsvorschrift, die in Daten enthaltene Informationen auf eine „Kennzahl“ ($\rightarrow$ Teststatistik) verdichtet
- Grundlegende Teilgebiete der Statistik:
  - Deskriptive Statistik (auch: beschreibende Statistik, explorative Statistik)
  - Schließende Statistik (auch: inferenzielle Statistik, induktive Statistik)
- Typischer Einsatz von Statistik:

  Verarbeitung — insbesondere Aggregation — von (eventuell noch zu erhebenden) Daten mit dem Ziel, (informelle) Erkenntnisgewinne zu erhalten bzw. (formal) Schlüsse zu ziehen.

$\rightsquigarrow$ Bestimmte Informationen „ausblenden“, um neue Informationen zu erkennen

Vorurteile gegenüber Statistik

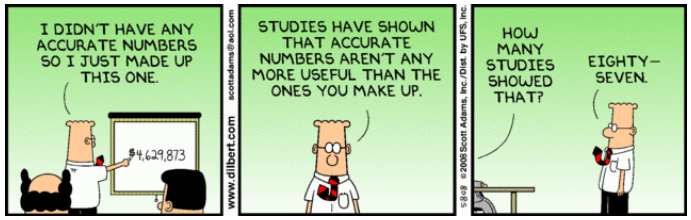
- Einige Zitate oder „Volksweisheiten“:
  - „Statistik ist pure Mathematik, und in Mathe war ich immer schlecht...“
  - „Mit Statistik kann man alles beweisen!“
  - „Ich glaube nur der Statistik, die ich selbst gefälscht habe.“ (häufig Winston Churchill zugeschrieben, aber eher Churchill von Goebbels' Propagandaministerium „in den Mund gelegt“)
  - „There are three kinds of lies: lies, damned lies, and statistics.“ (häufig Benjamin Disraeli zugeschrieben)

  $\rightsquigarrow$ negative Vorurteile gegenüber der Disziplin „Statistik“
- Tatsächlich aber
  - benötigt man für viele statistische Methoden nur die vier Grundrechenarten.
  - ist „gesunder Menschenverstand“ viel wichtiger als mathematisches Know-How.
  - sind nicht die statistischen Methoden an sich schlecht oder gar falsch, sondern die korrekte Auswahl und Anwendung der Methoden zu hinterfragen.
  - werden viele (korrekte) Ergebnisse statistischer Untersuchungen lediglich falsch interpretiert.

Kann man mit Statistik lügen? I

Und falls ja, wie (schützt man sich dagegen)?

- Natürlich kann man mit Statistik „lügen“ bzw. täuschen!
- „Anleitung“ von Prof. Dr. Walter Krämer (TU Dortmund): *So lügt man mit Statistik, Campus, 2015*
- Offensichtliche Möglichkeit: Daten (vorsätzlich) manipulieren/fälschen:



Kann man mit Statistik lügen? II

Und falls ja, wie (schützt man sich dagegen)?

- Weitere Möglichkeiten zur Täuschung
  - Irreführende Grafiken
  - (Bewusstes) Weglassen relevanter Information
  - (Bewusste) Auswahl ungeeigneter statistischer Methoden
- Häufiges Problem (vor allem in den Medien): *Suggestion von Sicherheit durch hohe Genauigkeit angegebener Werte*

  $\rightsquigarrow$ zusätzlich: Ablenkung vom „Adäquationsproblem“ (misst der angegebene Wert überhaupt das „Richtige“?)
- Schutz vor Täuschung:
  - Mitdenken!
  - „Gesunden Menschenverstand“ einschalten!
  - Gute Grundkenntnisse in Statistik!

Beispiel (Adäquationsproblem) I

vgl. Walter Krämer: So lügt man mit Statistik, Piper, München, 2009

- Frage: Was ist *im Durchschnitt* sicherer, Reisen mit Bahn oder Flugzeug?
- Statistik 1:

Bahn	9 Verkehrstote pro 10 Milliarden Passagierkilometer
Flugzeug	3 Verkehrstote pro 10 Milliarden Passagierkilometer

  $\rightsquigarrow$ Fliegen sicherer als Bahnfahren!
- Statistik 2:

Bahn	7 Verkehrstote pro 100 Millionen Passagierstunden
Flugzeug	24 Verkehrstote pro 100 Millionen Passagierstunden

  $\rightsquigarrow$ Bahnfahren sicherer als Fliegen!
- Widerspruch? Fehler?

Beispiel (Adäquationsproblem) II

vgl. Walter Krämer: So lügt man mit Statistik, Piper, München, 2009

- Nein, Unterschied erklärt sich durch höhere Durchschnittsgeschwindigkeit in Flugzeugen (Annahme: ca. 800 km/h vs. ca. 80 km/h)
- Wie wird „Sicherheit“ gemessen? Welcher „Durchschnitt“ ist geeigneter?
  $\rightsquigarrow$ Interpretation abhängig von der Fragestellung! Hier:
  - Steht man vor der Wahl, eine gegebene Strecke per Bahn oder Flugzeug zurückzulegen, so ist Fliegen sicherer.
  - Vor einem vierstündigen Flug ist dennoch eine größere „Todesangst“ angemessen als vor einer vierstündigen Bahnfahrt.

Beispiel („Schlechte“ Statistik) I

- Studie/Pressemitteilung des ACE Auto Club Europa *anläßlich des Frauentags am 8. März 2010*: „Autofahrerinnen im Osten am besten“ (siehe `https://www.ace.de/fileadmin/user_uploads/Der_Club/Dokumente/Verkehrspolitik/Handout-Booklet-ACE-Studien.pdf`, S. 88–90)
- Untersuchungsgegenstand:
  - Regionale Unterschiede bei Unfallhäufigkeit mit Frauen als Hauptverursacher
  - Vergleich Unfallhäufigkeit mit Frau bzw. Mann als Hauptverursacher
- Wesentliche Datengrundlage ist eine Publikation des Statistischen Bundesamts (Destatis): „Unfälle im Straßenverkehr nach Geschlecht 2008“

Beispiel („Schlechte“ Statistik) II

- Beginn der Pressemitteilung des ACE:

  „Von wegen schwaches Geschlecht: Hinterm Steuer sind Frauen besonders stark.“

  Weiter heißt es:

  “Auch die durch Autofahrerinnen verursachten Unfälle mit Personenschaden liegen wesentlich hinter den von Männern verursachten gleichartigen Karambolagen zurück.“

  und in einer Zwischenüberschrift

  „Schlechtere Autofahrerinnen sind immer noch besser als Männer“

Beispiel („Schlechte“ Statistik) III

- „Statistische“ Argumentation: Laut Destatis-Quelle sind (**angeblich!**)
  - mehr als 2/3 aller Unfälle mit Personenschaden 2008 (genauer: 217 843 von etwas über 320 000 Unfällen) durch PKW-fahrende Männer verursacht worden,
  - nur 37% aller Unfälle mit Personenschaden 2008 durch PKW-fahrende Frauen verursacht worden.
- Erste Auffälligkeit: $66.6\% + 37\% = 103.6\%$ **(???)**
- Lösung: **Ablesefehler** (217 843 aller 320 614 Unfälle mit Personenschaden (67.9%) wurden mit **PKW-Fahrer** (geschlechtsunabhängig) als Hauptverursacher registriert)

Beispiel („Schlechte“ Statistik) IV

- Korrekte Werte:
  - Bei 210 905 der 217 843 Hauptunfallverursacher als PKW-Fahrzeugführer wurde Geschlecht registriert.
  - 132 757 waren männlich (62.95%), 78 148 weiblich (37.05%)
- **Also:** immer noch deutlich mehr Unfälle mit PKW-fahrenden Männern als Hauptverursacher im Vergleich zu PKW-Fahrerinnen.
- **Aber:** Absolute Anzahl von Unfällen geeignetes Kriterium für Fahrsicherheit?

Beispiel („Schlechte“ Statistik) V

- Modellrechnung des DIW aus dem Jahr 2004 schätzt
  - Anzahl Männer mit PKW-Führerschein auf 28.556 Millionen,
  - Anzahl Frauen mit PKW-Führerschein auf 24.573 Millionen.
- Weitere ältere Studie (von 2002) schätzt
  - durchschnittliche Fahrleistung von Männern mit PKW-Führerschein auf 30 km/Tag,
  - durchschnittliche Fahrleistung von Frauen mit PKW-Führerschein auf 12 km/Tag.
- Damit stehen also
  - bei Männern 132 757 verursachte Unfälle geschätzten $30 \cdot 365 \cdot 28.556 = 312688.2$ Millionen gefahrenen Kilometern,
  - bei Frauen 78 148 verursachte Unfälle geschätzten $12 \cdot 365 \cdot 24.573 = 107629.74$ Millionen gefahrenen Kilometern

  gegenüber.

Beispiel („Schlechte“ Statistik) VI

- Dies führt im Durchschnitt
  - bei Männern zu 0.425 verursachten Unfällen mit Personenschaden pro eine Million gefahrenen Kilometern,
  - bei Frauen zu 0.726 verursachten Unfällen mit Personenschaden pro eine Million gefahrenen Kilometern.
- Pro gefahrenem Kilometer verursachen (schätzungsweise) weibliche PKW-Fahrer also durchschnittlich ca. **71% mehr** Unfälle als männliche!
- Anstatt dies zu konkretisieren, räumt die Studie lediglich weit am Ende ein entsprechendes Ungleichgewicht bei der jährlichen Fahrleistung ein.

Beispiel („Schlechte“ Statistik) VII

- Welt Online (siehe `http://www.welt.de/vermischtes/article6674754/Frauen-sind-bessere-Autofahrer-als-Maenner.html`) beruft sich auf die ACE-Studie in einem Artikel mit der Überschrift

  „Frauen sind bessere Autofahrer als Männer“

  und der prägnanten Bildunterschrift

  „Männer glauben bloß, sie seien die besseren Autofahrer. Eine Unfall-Statistik beweist das Gegenteil.“

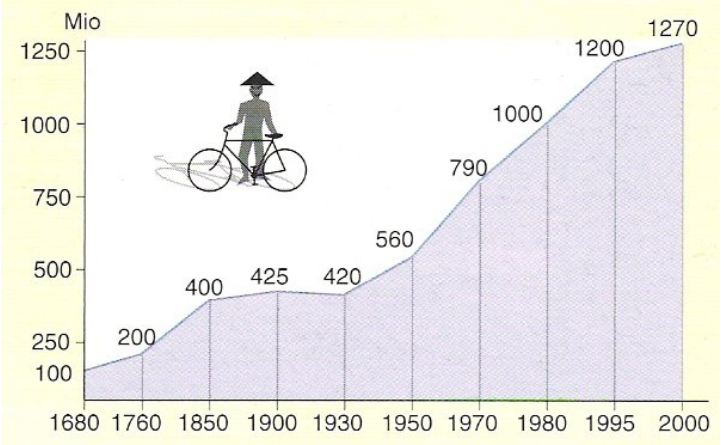
  Erst am Ende wird einschränkend erwähnt:

  „Fairerweise muss man erwähnen, dass Männer täglich deutlich mehr Kilometer zurücklegen. Und: Während 93 Prozent von ihnen einen Führerschein besitzen, sind es bei den Frauen lediglich 82 Prozent.“

Beispiel (Irreführende Grafik) I

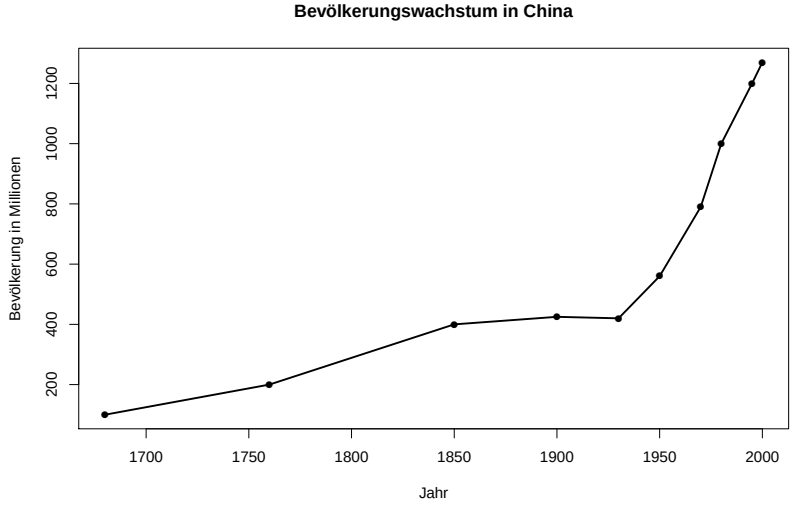
vgl. `http://www.klein-singen.de/statistik/h/Wissenschaft/Bevoelkerungswachstum.html`

Bevölkerungswachstum in China



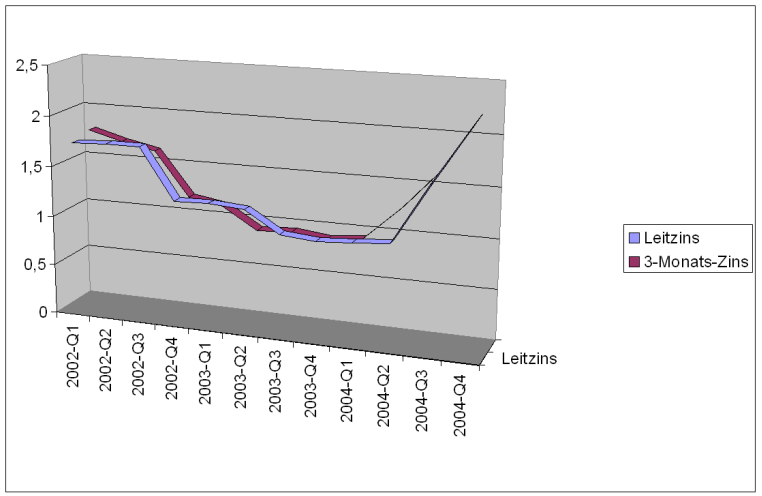
Beispiel (Irreführende Grafik) II

identischer Datensatz, angemessene Skala



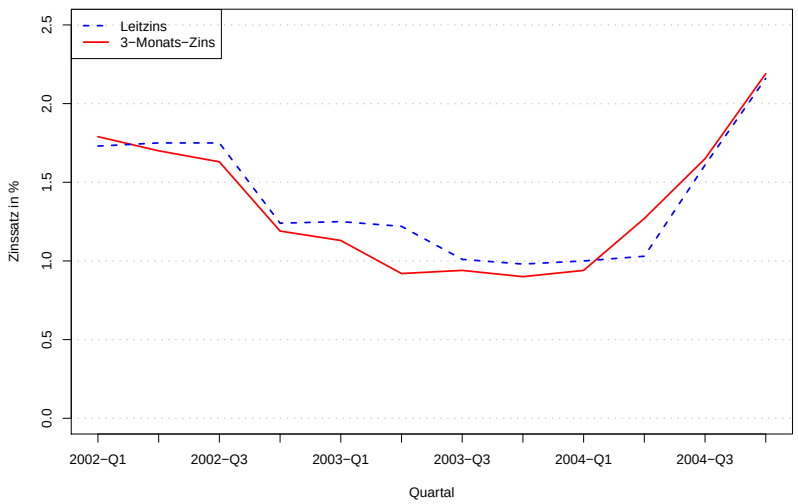
Beispiel (Chartjunk)

Microsoft Excel mit Standardeinstellung für 3D-Liniendiagramme



Beispiel (Grafik ohne Chartjunk)

Statistik-Software **R**, identischer Datensatz



Kann Statistik auch nützlich sein?

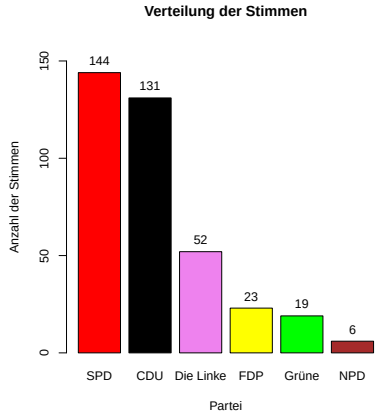
Welche Partei erhält wie viele Stimmen im Wahlbezirk 1.206 der Gemeinde Losheim am See bei den Erststimmen zur Bundestagswahl 2009? Stimmen:

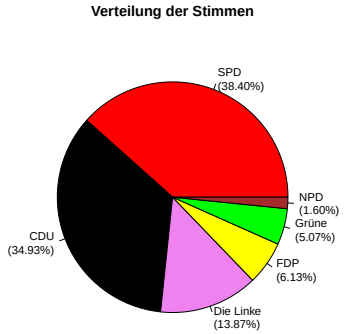
Die Linke, SPD, CDU, Die Linke, SPD, SPD, Die Linke, CDU, FDP, Grüne, Die Linke, SPD, Die Linke, CDU, SPD, CDU, CDU, SPD, SPD, FDP, CDU, FDP, Die Linke, Die Linke, Grüne, CDU, CDU, CDU, CDU, Die Linke, CDU, CDU, CDU, SPD, CDU, SPD, SPD, CDU, FDP, FDP, SPD, CDU, CDU, CDU, CDU, SPD, SPD, SPD, CDU, NPD, SPD, Die Linke, CDU, CDU, FDP, Grüne, SPD, FDP, CDU, CDU, CDU, SPD, SPD, SPD, CDU, Die Linke, CDU, Die Linke, SPD, FDP, CDU, SPD, CDU, CDU, CDU, SPD, Die Linke, CDU, Die Linke, NPD, SPD, Grüne, FDP, SPD, FDP, SPD, CDU, SPD, CDU, SPD, SPD, SPD, SPD, CDU, CDU, Die Linke, CDU, CDU, SPD, CDU, CDU, Die Linke, CDU, SPD, SPD, SPD, SPD, SPD, SPD, Die Linke, Die Linke, Die Linke, CDU, Die Linke, CDU, Grüne, CDU, CDU, SPD, CDU, SPD, CDU, CDU, SPD, SPD, CDU, FDP, CDU, SPD, SPD, SPD, CDU, CDU, Die Linke, CDU, CDU, CDU, SPD, FDP, SPD, SPD, Die Linke, SPD, Grüne, SPD, Grüne, FDP, SPD, CDU, Die Linke, FDP, SPD, CDU, SPD, SPD, SPD, SPD, Die Linke, SPD, SPD, CDU, SPD, CDU, Die Linke, SPD, CDU, CDU, CDU, SPD, SPD, SPD, Die Linke, FDP, Grüne, CDU, SPD, CDU, SPD, SPD, Die Linke, SPD, CDU, CDU, CDU, SPD, SPD, SPD, Die Linke, SPD, SPD, SPD, SPD, Die Linke, CDU, CDU, Die Linke, CDU, CDU, SPD, SPD, CDU, CDU, SPD, SPD, CDU, CDU, NPD, SPD, SPD, CDU, SPD, SPD, Grüne, CDU, SPD, SPD, Die Linke, FDP, Die Linke, CDU, SPD, Grüne, SPD, CDU, SPD, Die Linke, Die Linke, SPD, CDU, Die Linke, SPD, SPD, SPD, Die Linke, Die Linke, SPD, SPD, FDP, CDU, CDU, SPD, SPD, CDU, SPD, CDU, SPD, SPD, CDU, SPD, CDU, CDU, SPD, Grüne, SPD, SPD, SPD, CDU, CDU, SPD, SPD, SPD, FDP, Die Linke, CDU, FDP, CDU, Die Linke, SPD, CDU, CDU, CDU, CDU, Grüne, CDU, CDU, CDU, SPD, CDU, SPD, Die Linke, CDU, Die Linke, SPD, Die Linke, NPD, CDU, Grüne, Die Linke, CDU, CDU, Die Linke, Die Linke, SPD, SPD, CDU, Grüne, SPD, Die Linke, SPD, SPD, SPD, CDU, Die Linke, SPD, SPD, SPD, NPD, SPD, CDU, SPD, SPD, SPD, Grüne, CDU, SPD, SPD, SPD, FDP, Grüne, SPD, Die Linke, CDU, SPD, SPD, CDU, SPD, SPD, Die Linke, Die Linke, CDU, FDP, CDU, SPD, Die Linke, SPD, CDU, CDU, SPD, SPD, SPD, CDU, CDU, Grüne, CDU, CDU, CDU, FDP, Die Linke, SPD, CDU, Die Linke, CDU, SPD, CDU, FDP, SPD, SPD, CDU, SPD, CDU, CDU, CDU, CDU, NPD, CDU, Grüne, SPD, SPD, CDU, Grüne, CDU, SPD, CDU, SPD

- Mit etwas (deskriptiver) Statistik in tabellarischer Form:

	SPD	CDU	Die Linke	FDP	Grüne	NPD	Summe
Anzahl der Stimmen	144	131	52	23	19	6	375
Stimmenanteil in %	38.40	34.93	13.87	6.13	5.07	1.60	100.00

- Grafisch aufbereitete Varianten:





Organisation der Statistik-Veranstaltungen