

Datenerhebung und Statistik

Wirtschaftspsychologie – SoSe 2018

FOM

Inhaltsverzeichnis

- 1 Organisatorisches
- 2 Wissenschaftliche Grundlagen
- 3 Grundlagen Quantitativer Datenanalyse
- 4 Einführung R
- 5 Explorative Datenanalyse
- 6 Einführung Inferenz
- 7 Normalverteilung
- 8 Inferenz kategorialer Daten
- 9 Inferenz numerischer Daten
- 10 Lineare Regression
- 11 Datenhandling
- 12 Organisatorisches

1 Organisatorisches

1. Organisatorisches Literatur (Auswahl)

- ▶ David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel (2014): *Introductory Statistics with Randomization and Simulation*,
https://www.openintro.org/stat/textbook.php?stat_book=isrs
- ▶ Nicholas J. Horton, Randall Pruim, Daniel T. Kaplan (2015): Project MOSAIC Little Books *A Student's Guide to R*, <https://github.com/ProjectMOSAIC/LittleBooks/raw/master/StudentGuide/MOSAIC-StudentGuide.pdf>
- ▶ Chester Ismay, Albert Y. Kim (2017): *ModernDive – An Introduction to Statistical and Data Sciences via R*, <http://moderndive.com/>
- ▶ Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning – with Applications in R*,
[<http://www-bcf.usc.edu/~gareth/ISL/>] (<http://www-bcf.usc.edu/~gareth/ISL/>)

1. Organisatorisches

Lizenz / Version

Diese Folien wurden von Karsten Lübke zusammen mit Kolleg*innen von der FOM
<https://www.fom.de/> entwickelt und stehen unter der Lizenz CC-BY-SA-NC 3.0 de:
<https://creativecommons.org/licenses/by-nc-sa/3.0/de/>

Der verwendete Code sowie das Beamer Template aus dem [NPBT-Projekt](#) von Norman Markgraf stehen unter der Lizenz [GNU General Public License v3.0](#).

- ▶ Datum erstellt: 2018-02-09
- ▶ R Version: 3.4.3
- ▶ mosaic Version: 1.1.1

Bitte melden Sie Fehler und Verbesserungsvorschläge: karsten.luebke@fom.de

Mitarbeit und Hinweise von Thomas Christiaans, Oliver Gansser, Matthias Gehrke, Jörg Horst, Bianca Krol, Norman Markgraf, Sebastian Sauer, Daniel Ziggel. **Vielen Dank!**

Die Studierenden können nach erfolgreichem Abschluss des Moduls:

- ▶ den Prozess der Erkenntnisgewinnung in der Psychologie nutzen,
- ▶ Daten gewinnen, zusammenfassen, analysieren und graphisch darstellen,
- ▶ statistische Aussagen über Zusammenhänge und Prognosen machen,
- ▶ die Gültigkeit der gefundenen Schlussfolgerungen abschätzen,
- ▶ wichtige Methoden der deskriptiven und Inferenzstatistik passgenau auswählen und anwenden,
- ▶ Auswertungen mit R durchführen,
- ▶ die Anwendung statistischer Auswertungen in Fachveröffentlichungen verstehen und einordnen,
- ▶ in der Struktur psychologischer Veröffentlichungen ihre Ergebnisse berichten,
- ▶ vorbereitend für die Projektarbeiten und ihre Abschlussarbeit angemessene empirische Methoden einsetzen.

- ▶ Datenanalyse (ca. 1500 Wörter)
- ▶ Klausur 90 Minuten

Seminararbeit und Klausur gehen jeweils zu 50 % in die Modulnote ein, beide Prüfungsleistungen müssen mit mindestens 4,0 bewertet werden.

Beachten Sie die im OC hinterlegten Fristen!

Workload:

- ▶ Präsenzstunden: 60,0 UE
- ▶ Strukturiertes Eigenstudium 130,00 ZStd
- ▶ Workload gesamt: 175,0 ZStd
- ▶ ECTS-Credit Punkte: 7

Bitte

- ▶ Stellen Sie ein Namensschild auf.
- ▶ Seien Sie offen für das Thema.
- ▶ Haben Sie Respekt aber keine Angst vor der Schwierigkeit des Themas.
- ▶ Stellen Sie Fragen!
- ▶ Sie können Sich gerne während der Übungen unterhalten, aber nicht wenn ich etwas erkläre – das ist u. a. auch unfair!
- ▶ Versuchen Sie die Übungen selbst zu lösen - der Lernerfolg ist ungleich größer, als wenn Sie die Lösung "abnicken". Für die Quizze wird ein Live-Feedback-System eingesetzt: <https://tweedback.de/> **Nehmen Sie daran teil!**¹
- ▶ Versuchen Sie Ablenkungen (Facebook, WhatsApp etc.) zu vermeiden.²
- ▶ Wenn möglich bringen Sie einen Laptop mit R³ zu den Vorlesungsterminen mit.

¹Siehe z. B. Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: the good, the bad, and the ugly. *Teaching of Psychology*, 42(1), 87-92.

²Siehe z. B. Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, 62, 24-31.

³Installationsanleitung [Hier](#)

Bleiben Sie dran!

Die Inhalte bauen aufeinander auf, d. h., arbeiten Sie nach. Die angegebene Literatur ist frei verfügbar.

Ich kann versuchen, es Ihnen zu erklären, ich kann es *nicht* für Sie verstehen.

Tipps von (fiktiven) Studierenden:

- ▶ Ich besuche die Vorlesung nicht, ich gucke Videos.
- ▶ Kontinuierlich nacharbeiten? Quatsch, ich lerne eine Woche vor der Klausur intensiv.
- ▶ Es reicht wenn ich mit einem halben Ohr zuhöre – ich spiele, chatte, surfe während der Vorlesung.⁴
- ▶ Mitschreiben? Ich mache, wenn überhaupt, ein Foto.⁵
- ▶ Selbstlernunterlagen und Literatur habe ich nicht nötig.
- ▶ Mir genügt die Übungsklausur zur Klausurvorbereitung - falls eine angeboten wird.
- ▶ Übungen selber lösen? – Es reicht, wenn ich die Lösung kenne.

Manche dieser Studierenden haben es leider **nicht** geschafft.

Sie können es **besser!**⁶

⁴vgl. <https://youtu.be/vJG698U2Mvo>

⁵Hier **nicht** erlaubt.

⁶Siehe z. B. echte Tipps unter Putnam, A. L., Sungkhasettee, V. W., & Roediger III, H. L. (2016).

Optimizing learning in college: tips from cognitive psychology. Perspectives on Psychological Science, 11(5), 652-660.

In God we trust; all others bring data.

– W. Edwards Deming (zugeschrieben)

2 Wissenschaftliche Grundlagen

Science is a particular way of knowing about the world. In science, explanations are limited to those based on observations and experiments that can be substantiated by other scientists. Explanations that cannot be based on empirical evidence are not part of science.

Fact: In science, an observation that has been repeatedly confirmed and for all practical purposes is accepted as "true." Truth in science, however, is never final, and what is accepted as a fact today may be modified or even discarded tomorrow.

Hypothesis: A tentative statement about the natural world leading to deductions that can be tested. If the deductions are verified, it becomes more probable that the hypothesis is correct. If the deductions are incorrect, the original hypothesis can be abandoned or modified. Hypotheses can be used to build more complex inferences and explanations.

Law: A descriptive generalization about how some aspect of the natural world behaves under stated circumstances.

Theory: In science, a well-substantiated explanation of some aspect of the natural world that can incorporate facts, laws, inferences, and tested hypotheses.

Quelle: Science and Creationism: A View from the National Academy of Sciences, Second Edition (1999)⁷

⁷<https://doi.org/10.17226/6024>

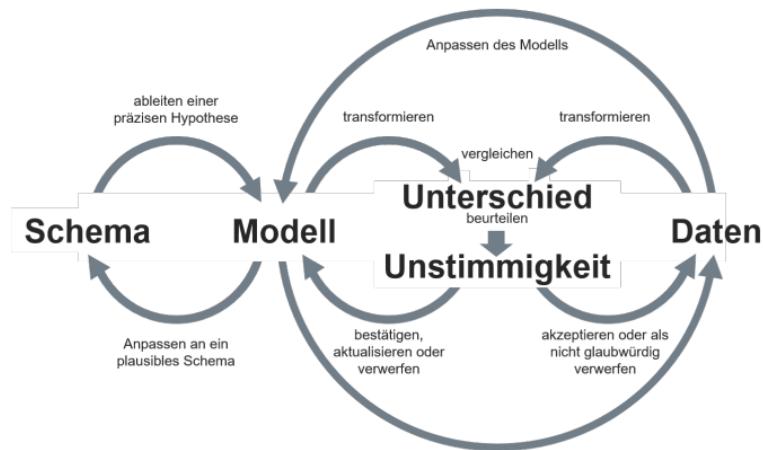


Abbildung nach Garrett Grolemund, Hadley Wickham. A Cognitive Interpretation of Data Analysis.
International Journal of Statistics, vol. 82, no. 2, pp. 184–204, 2014.

- Schema: Ein mentales Modell, dass die ganze Bandbreite der Information über ein Thema beinhaltet
- Modell: Repräsentation eines Ausschnitts innerhalb des Themas
- Daten: Messungen der Realität⁸

⁸<http://vita.had.co.nz/papers/sensemaking.html>

Übung 1: Daten

Stimmt die Aussage: Daten und deren Analyse sind ein zentraler Bestandteil wissenschaftlicher Argumentation?

- ▶ Ja.
- ▶ Nein.

- ▶ Laut dem **wissenschaftlichen Realismus** existiert eine reale Welt unabhängig von der Sicht des Betrachtenden.
- ▶ Im **Konstruktivismus** wird angenommen, dass Wissen über die Wirklichkeit erst durch Wahrnehmung erschaffen wird.
- ▶ Die Realität ist komplex, teilweise verdeckt und dynamisch (siehe auch Chaosforschung).

- ▶ Eine Theorie ist eine strukturierte Sammlung von Hypothesen.
- ▶ Sie schlägt eine vorläufige Antwort auf eine offene Frage vor.
- ▶ Sie lässt sich kaum in ihrem vollen Umfang (auf einmal) prüfen.
- ▶ Häufig sind Theorien zumeist an kausalen Beziehungen interessiert.
- ▶ Gute Theorien ermöglichen Vorhersagen, Erklärungen, Nutzen.

Hypothese

Eine Hypothese ist eine aus der Theorie abgeleitete Aussage.

- ▶ Sie sind weniger umfangreich als Theorien.
- ▶ Sie stellen Vermutungen über einen Sachverhalt an.
- ▶ Sie ist eine provisorische Antwort auf ein wissenschaftliches Problem.
- ▶ Sie lassen sich überprüfen (sind also potentiell "falsifizierbar", man kann zeigen, dass sie falsch sind). Hypothesen sind (nach Karl Popper) nie beweisbar/bestätigbar, man kann höchstens zeigen, dass sie falsch sind.

Kennzeichen einer wissenschaftlichen Hypothese:

- ▶ Sie ist eine allgemeingültige über den Einzelfall hinausgehende Behauptungen.
- ▶ Formalstruktur eines Konditionalsatzes
 - ▶ (wenn-dann-Satz, je-desto-Satz)
 - ▶ (Wenn-Teil \approx Antezedenz \approx h. V., Dann-Teil \approx Konsequenz \approx h. V.)

Modelle sind ganz allgemein vereinfachte Darstellungen relevanter Teile der Realität. Die Repräsentation der Realität durch Modelle ermöglicht eine einfachere Analyse.

- ▶ Darstellung von Modellen
 - ▶ graphisch (Pfadiagramme)⁹
 - ▶ verbal
 - ▶ In der Wissenschaft verwendet man häufig: **mathematisch-statistische Modelle** zur Beschreibung der Realität.
 - ▶ Das Instrumentarium der Mathematik kann eingesetzt werden, um zu optimieren.
 - ▶ Mathematisch formulierte Modelle lassen sich gut statistisch überprüfen und ermöglichen Prognosen.
 - ▶ Statistische Modelle sind mathematische Modelle die mit Hilfe von Daten gewonnen werden.

⁹Beispiel: Kundenzufriedenheit

<http://www.theacsi.org/about-acsi/the-science-of-customer-satisfaction>

Offene Übung 2: Hypothese

- ▶ Folgende Zahlen folgen einer Regel: 2, 4, 6.
- ▶ Wie lautet die Regel?
- ▶ Sie können Zahlentripel vorschlagen, um die Regel zu prüfen!¹⁰

¹⁰Wason, P. C. (1960): *On the failure to eliminate hypotheses in a conceptual task*. Quarterly journal of experimental psychology, 12(3), 129-140. <http://dx.doi.org/10.1080/17470216008416717>

Induktion:

Generalisierung von in der Realität beobachteten Regelmäßigkeiten zu einer allgemeineren Vermutungen.

Häufig: Hypothesenbildung.

Beispiele: Pawlow, Freud

Deduktion:

Ableitung von Aussagen aus anderen (allgemeineren) Aussagen mit Hilfe logischer Regeln.

Häufig: Hypothesenprüfung.

Abduktion:

Verknüpfung von Einzelbeobachtungen und erkennen von Regeln.

- ▶ *Induktion*: Hm ich habe schon 30 Bohnen aus dem Sack gezogen... Alle weiß. Noch 30 Bohnen... schon wieder alle weiß. Ich hab's: Die Bohnen müssen alle weiß sein!
- ▶ *Deduktion*: Ich habe die Bohnen in den Sack gefüllt. Sie waren alle weiß. Jetzt nehme ich eine Bohne aus dem Sack: sie ist weiß!
- ▶ *Abduktion*: Vor mir steht ein Sack; ich sehe, dass Bohnen darin sind. Ich finde eine weiße Bohne irgendwo im Raum auf dem Boden. Daraus schließe ich: Die Bohne muss aus dem Sack sein!

Quelle: Walach, H. (2013): *Psychologie: Wissenschaftstheorie, philosophische Grundlagen und Geschichte; ein Lehrbuch*. Stuttgart: Kohlhammer.

Eine Dozentin hat mehrfach beobachtet, dass ihre Studierenden interessiert am Fach Statistik sind. Nun schließt sie, dass alle Studierenden interessiert am Fach Statistik sind. Welche Schlussart liegt vor?

- A. Induktion.
- B. Deduktion.

Falsifikationsprinzip

- ▶ **Induktionsproblematik:** Kann durch Induktion von Einzelbeobachtungen *sicher* auf ein allgemeines Gesetz – auch in Zukunft – geschlossen werden?
- ▶ **Falsifikationsprinzip:** Obwohl es nicht möglich ist, die Richtigkeit einer wissenschaftlichen Theorie auf Basis einer begrenzten Menge von Daten zu beweisen, ist es möglich zu beweisen, dass eine Theorie falsch ist. Wissenschaftliche Aussagen sollen daher nach Karl Popper empirisch widerlegbar sein. Beispiel: Kann man die Hypothese beweisen, dass alle Schwäne weiß sind?
 - ▶ Das ist kaum/nicht möglich, man müsste die ganze Welt bereisen (und die Zukunft und die Vergangenheit). Und selbst dann: wer weiß, vielleicht habe ich einen übersehen?
 - ▶ Zu zeigen, dass die Hypothese falsch ist, ist einfach: Ein schwarzer Schwan reicht, um zu zeigen, dass die Behauptung, alle Schwäne seien weiß, falsch ist!

Übung 4: Beweis

Hat sie die Aussage “alle Studierenden sind interessiert an Statistik” endgültig bewiesen?

- A. Ja.
- B. Nein.
- C. Vielleicht.

Forschungsprozess

- ▶ **Planung:** Formulierung der Forschungsfrage: Forschungsidee, Informationssammlung, Forschungsfrage und -hypothesen.
- ▶ **Datenerhebung:** Primär- oder Sekundärerhebung¹¹, Operationalisierung, Versuchsplanung und -durchführung, Datenerhebung, Datenbeschaffung.
- ▶ **Datenaufbereitung** und **Datenanalyse:** Fehlende Werte, Ausreißer, Explorative Datenanalyse, Inferenzstatistik, Modellierung.
- ▶ **Interpretation:** Was sagt das Ergebnis aus? Schlussfolgerung, Mitteilung der Ergebnisse.

¹¹Digitalisierung!

Übung 5: Literatur

An welcher Stelle des Forschungsprozesses ist Literaturrecherche besonders zentral?

- A. Zu Beginn.
- B. Bei der Datenerhebung.
- C. Bei der Datenauswertung.
- D. Zum Ende.

Gütekriterien für Forschung:

- ▶ **Ethische Aspekte:** Können negative Folgen, z. B. bei befragten / untersuchten Personen auftreten?
- ▶ **Transparenz:** Das Vorgehen ist klar dokumentiert und nachprüfbar (und damit prinzipiell **reproduzierbar**)?
- ▶ **Objektivität:** Sind die Ergebnisse unabhängig von der Person? Kommen andere zum selben Ergebnis?
- ▶ **Interne Validität:** Keine anderen Erklärungen für die Ergebnisse? Ist der behauptete Zusammenhang richtig?
- ▶ **Externe Validität:** Übertragbarkeit der Ergebnisse? Zeigt sich der behauptete Zusammenhang auch in anderen Situationen?

Übung 6: Wissenschaftliches Arbeiten

Was ist ein unabdingbares Merkmal wissenschaftlichen Arbeitens?

- A. Spektakuläre Ergebnisse.
- B. Bestätigung der Forschungshypothese.
- C. Kein Praxisbezug.
- D. Keines der in A-C genannten.

Übung 7: Reproduzierbarkeit

Das Forschungsergebnis der Dozentin (“alle Studierenden sind interessiert an Statistik”) kann nicht reproduziert werden. Welcher Aspekt wissenschaftlichen Arbeitens könnte verletzt sein?

- A. Nur Objektivität.
- B. Nur interne Validität.
- C. Nur externe Validität.
- D. Alle in A–C genannten.

► Quantitative Methoden

- ▶ Messung und numerische Beschreibung der Wirklichkeit.
- ▶ Allgemeingültige Gesetze für die Grundgesamtheit.
- ▶ Ein Ausschnitt der beobachteten sozialen Vielfalt wird auf Skalen abgebildet, und es wird mit Häufigkeiten, Mittelwerten, Wahrscheinlichkeiten des Auftretens von Merkmalsausprägungen operiert.

► Qualitative Methoden

- ▶ Verbalisierung der Erfahrungswirklichkeit.
- ▶ Wirklichkeitsinterpretationen sind durch spezifische soziale Handlungsweisen geprägt und strukturieren gleichzeitig das soziale Handeln der Einzelperson vor.
- ▶ Untersuchungsgegenstand soll möglichst in seinem natürlichen Umfeld detailliert, ganzheitlich und umfassend erfasst werden.

3 Grundlagen Quantitativer Datenanalyse

- ▶ Beim Messen wird einer Eigenschaft eines Objektes ein Wert zugewiesen. Dabei sollte die Beziehung der Werte der Beziehung der Eigenschaften der Objekte entsprechen.
- ▶ **Latente Variablen / Konstrukte** können nicht direkt gemessen werden, sie müssen erst **operationalisiert** werden, z. B. Intelligenz.
- ▶ **Manifeste Variablen** können direkt gemessen werden, z. B. Größe.

Bsp. Schwierigkeit von Statistik. **Multi-Item Likert-Skala:** Auf einer Skala von 1 (trifft überhaupt nicht zu) über 4 (weder zutreffend, noch unzutreffend) bis 7 (trifft voll und ganz zu) werden folgende Aussagen bewertet:¹²

- ▶ Statistische Formeln sind leicht zu verstehen.
- ▶ Statistik ist ein kompliziertes Fach.*
- ▶ Statistik ist ein Fach, das die meisten Menschen schnell lernen.
- ▶ Das Lernen von Statistik erfordert sehr viel Disziplin.*
- ▶ Statistik beinhaltet sehr umfangreiche Rechnungen.*
- ▶ Statistik ist eine sehr technische Materie.*
- ▶ Die meisten Menschen müssen lernen anders zu denken, um Statistik anwenden zu können.*

Die Items mit Sternchen * sind sogenannte inverse Items, bei denen die Zustimmung eine höhere Schwierigkeit im Umgang mit Statistik bedeutet.

¹²Candace Schau: Survey of Attitudes Toward Statistics, SATS-36

Übung 8: Messung

Stimmt die Aussage: Das “Interesse der Studierenden” ist eine latente Variable?

- ▶ Ja.
- ▶ Nein.

Offene Frage: Was folgt daraus?

Übung 9: Messung Relation

An einem Ort sei die Durchschnittstemperatur im Sommer 20°C , im Winter 10°C .
Stimmt die Aussage: Im Sommer ist es durchschnittlich doppelt so warm wie im Winter?

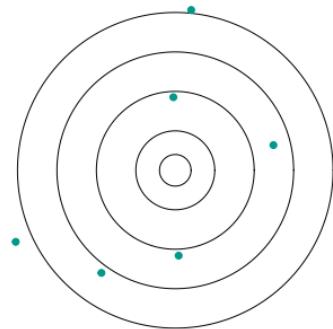
- ▶ Ja.
- ▶ Nein.

Gütekriterien einer Messung

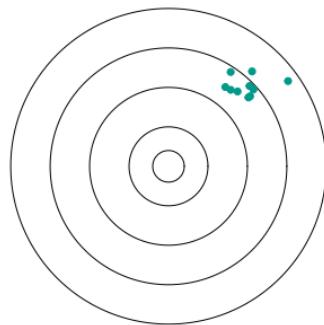
- ▶ **Genauigkeit**, d. h. Exaktheit einer Messung, z. B. "Umsatz hoch / niedrig" oder in Euro.
- ▶ **Objektivität**, d. h. Messung unabhängig vom Messenden, z. B. Kreditrating verschiedener Agenturen.
- ▶ **Reliabilität**, d. h. Zuverlässigkeit einer Messung, z. B. bei wiederholter / anderer Messung dasselbe Ergebnis bzgl. Kundenzufriedenheit.
- ▶ **Validität**, d. h., es wird das gemessen, was gemessen werden soll, z. B. Unternehmenserfolg oder Bilanz-Kniffe.

Messung: Varianz und Verzerrung

Varianz



Verzerrung



- ▶ hohe Varianz: geringe Reliabilität
- ▶ Verzerrung / Bias: geringe Validität

Welches Kriterium ist verletzt, wenn die Dozentin statt "Interesse der Studierenden" "Angst vor der Klausur" gemessen hat?

- A. Genauigkeit.
- B. Objektivität.
- C. Reliabilität.
- D. Validität.

Kategoriale Skala, qualitativ

- ▶ **Nominal:** Merkmalsausprägungen können unterschieden werden, bspw. Geschlecht.
- ▶ **Ordinal:** Merkmalsausprägungen können unterschieden und in eine Reihenfolge gebracht werden, bspw. Bildungsabschlüsse. Die Abstände zwischen den Werten können nicht direkt verglichen oder interpretiert werden.

Numerisch / metrische Skala, quantitativ, kardinal

Merkmalsausprägungen können unterschieden und in eine Reihenfolge gebracht werden, die Abstände sind vergleichbar.

- ▶ **Verhältnisskala:** Nullpunkt gegeben, bspw. Gewicht
- ▶ **Intervallskala:** Nullpunkt gesetzt, bspw. Zeitrechnung (Jahr 0)

Weitere Unterscheidung:

- ▶ **stetig:** beliebige Zwischenwerte im Intervall sind möglich, bspw. Größe
- ▶ **diskret:** höchstens abzählbar viele Werte sind möglich, bspw. Anzahl Kinder

Welches Skalenniveau hat die Variable Stundenlohn?

- A. Kategorial nominal.
- B. Kategorial ordinal.
- C. Metrisch stetig.
- D. Metrisch diskret.

Welches Skalenniveau hat die Variable Telefonvorwahl?

- A. Kategorial nominal.
- B. Kategorial ordinal.
- C. Metrisch stetig.
- D. Metrisch diskret.

Aufbau eines Datensatzes:

Name	Geschlecht	Größe
Ahmet	m	180
Gabi	w	170
Max	m	186
Susi	w	172

- ▶ Zeilen: Beobachtungen
- ▶ Spalten: Variablen

Messwerte einer Variable variieren / streuen, u. a.

- ▶ zufällig,
- ▶ aufgrund der Messung,
- ▶ aufgrund der Stichprobe,
- ▶ systematisch – kann evtl. modelliert werden.

Bsp.: Punkte einer Klausur variieren. *Warum?*

- ▶ **Unabhängige Variable** (exogen, erklärend, UV): Wert hängt von keiner anderen Variable ab (" x ").
- ▶ **Abhängige Variable** (endogen, erklärt, AV): Wert hängt von der / den unabhängigen(n) Variable ab (" y ").
- ▶ **Kovariablen**/ Störvariablen: Variablen, deren Wert ebenfalls auf die abhängige Variable einwirkt und / oder den Zusammenhang zwischen unabhängigen und abhängigen Variablen beeinflusst (" z ").¹³

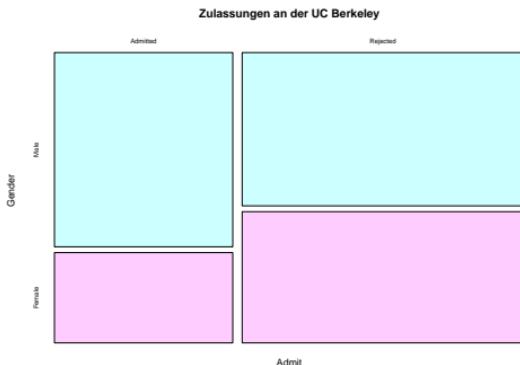
Hinweis: x steht in Zusammenhang mit y heißt nicht zwangsläufig, dass x kausal (ursächlich) für y sein muss!

- ▶ in der Mathematik: $y = f(x)$
- ▶ in der Statistik / in R: $y \sim x$

¹³Video <https://www.causeweb.org>: McLellan M © Confounding Variables

Simpson-Paradoxon (I / II)

Das ignorieren von Kovariablen kann zu verzerrten Ergebnissen führen.¹⁴



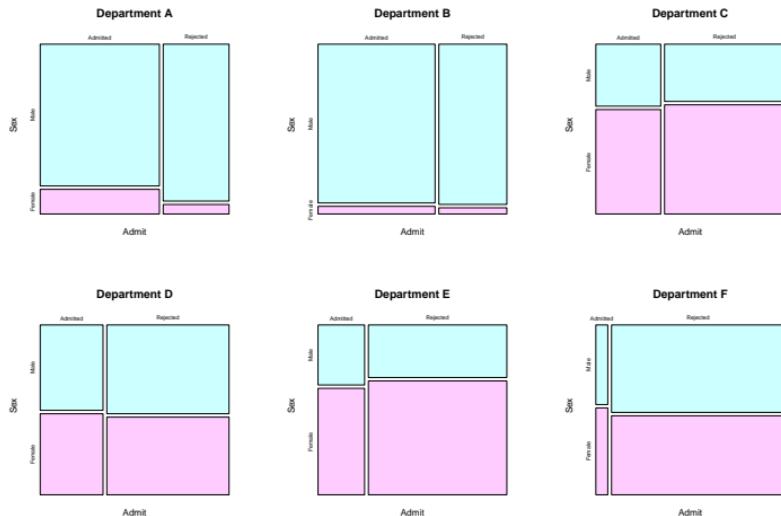
Höherer Frauenanteil bei den Nicht-Zugelassenen als bei den Zugelassenen:
Diskriminierung?

¹⁴Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187, 398–403. <https://doi.org/10.1126/science.187.4175.398>

3. Grundlagen Quantitativer Datenanalyse

Simpson-Paradoxon (II / II)

Zulassungen an der UC Berkeley



Je nach Department mal mehr mal weniger Frauen bei den Nicht-Zugelassenen als bei den Zugelassenen. Aber Frauen haben sich mehr für Fächer beworben, in denen der Anteil der Zugelassenen geringer war.

Die Dozentin stellt fest, dass die Motivation der Studierenden mit der Uhrzeit zusammenhängt, und zwar unterschiedlich für Frauen und Männer. Welche Aussage stimmt?

- A. Es gibt eine abhängige Variable (Motivation), eine unabhängige Variable (Uhrzeit) und eine Kovariable (Geschlecht).
- B. Es gibt eine abhängige Variable (Uhrzeit), eine unabhängige Variable (Motivation) und eine Kovariable (Geschlecht).
- C. Es gibt zwei abhängige Variable (Motivation und Geschlecht) und eine unabhängige Variable (Uhrzeit).
- D. Es gibt eine abhängige Variable (Geschlecht) und zwei unabhängige Variablen (Motivation und Uhrzeit).

Stichproben

- ▶ **Stichproben** sind eine Teilmenge der **Population** / Grundgesamtheit, die Beobachtungen / Daten.
- ▶ In der Regel ist man daran interessiert, das Ergebnis einer Stichprobe zu verallgemeinern, zu generalisieren: vom Geschmack des Suppenlöffels auf die ganze Suppe.¹⁵



¹⁵ hier: Kartoffelcremesuppe, Foto: Johann Hauke

Begriffe: Stichproben (I / II)

- ▶ **Population:** die Menge, über die eine Aussage getroffen werden soll: die ganze Suppe im Suppentopf.
- ▶ **Stichprobe:** Teilmenge der Population, die zur Analyse ausgewählt wurde: der Löffel voll Suppe.
- ▶ **Stichprobenverfahren:** der Prozess, mit dem die Teilmenge ausgewählt wurde. Z. B. **zufällig:** der Auswahlprozess, wo und wie der Löffel aus dem Suppentopf gefüllt wurde.
- ▶ **Repräsentative** Stichprobe: Ist die Verteilung der Eigenschaften der Stichprobe ähnlich der in der Population? Wenn der Löffel anders schmeckt als die Suppe, war der Löffel nicht repräsentativ.
- ▶ **Bias / Verzerrung:** Ein Teil der Population wird bevorzugt: nur Fleischbällchen auf dem Löffel.
- ▶ **Generalisierbarkeit:** Inwieweit kann von der Stichprobe auf die Grundgesamtheit geschlossen werden? Wenn wir gut umgerührt haben, sollten die Verteilung der Gewürze etc. auf dem Löffel ähnlich der im Topf sein und wir können vom Löffel auf den Topf schließen.

- ▶ **Parameter:** Wert der Grundgesamtheit, an dem wir interessiert sind: z. B. Temperatur der Suppe insgesamt.¹⁶
- ▶ **Statistik:** Wert, der auf Basis der Stichprobe berechnet wird: z. B. Temperatur der Suppe auf dem Löffel.¹⁷

¹⁶symbolisiert häufig durch griechische Buchstaben: μ, \dots

¹⁷symbolisiert häufig durch lateinische Buchstaben: \bar{x}, \dots

Übung 14: Stichprobe

Kann die Dozentin von den Studierenden, die die Vorlesung besuchen¹⁸, unverzerrt auf das Interesse aller Studierenden schließen, die für die Vorlesung angemeldet sind¹⁹?

- ▶ Ja.
- ▶ Nein.

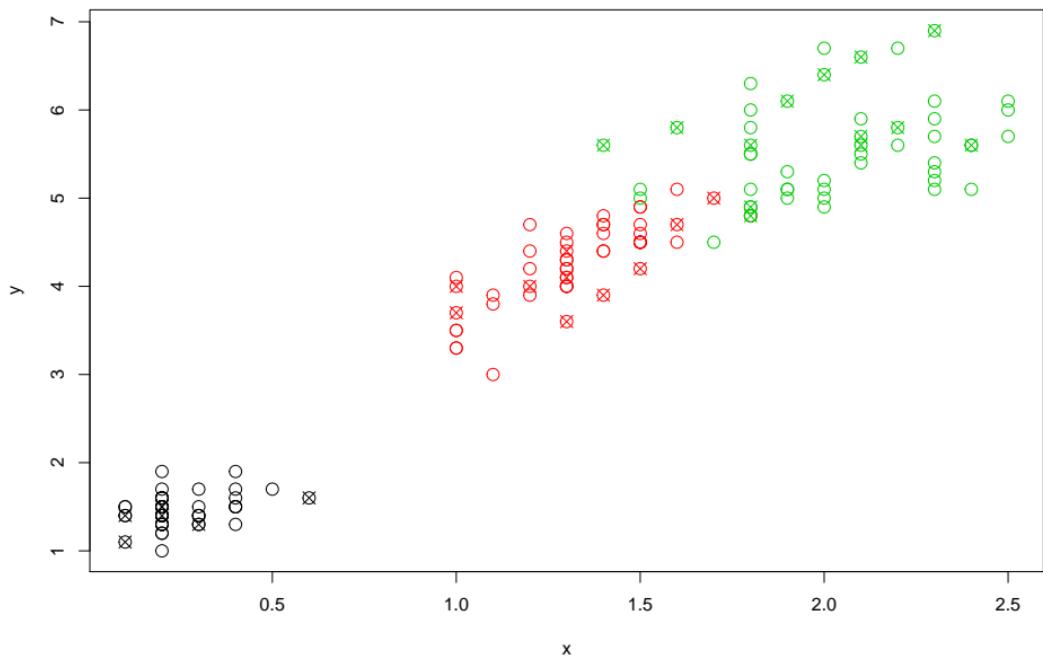
¹⁸Stichprobe

¹⁹Grundgesamtheit

- ▶ Bei einer (einfachen) **Zufallsstichprobe** hat jede Beobachtung die gleiche Wahrscheinlichkeit, Teil der Stichprobe zu sein.
- ▶ Bei **geschichtete Stichproben** setzen sich die Schichten aus ähnlichen Beobachtungen zusammen (z. B. Alter, Geschlecht). Es wird eine einfache, zufällige Stichprobe aus jeder Schicht genommen.
- ▶ Zufällige Stichproben erlauben einen Schluss auf die Grundgesamtheit (Generalisierbarkeit).
- ▶ Gelegenheitsstichproben können verzerrt sein.

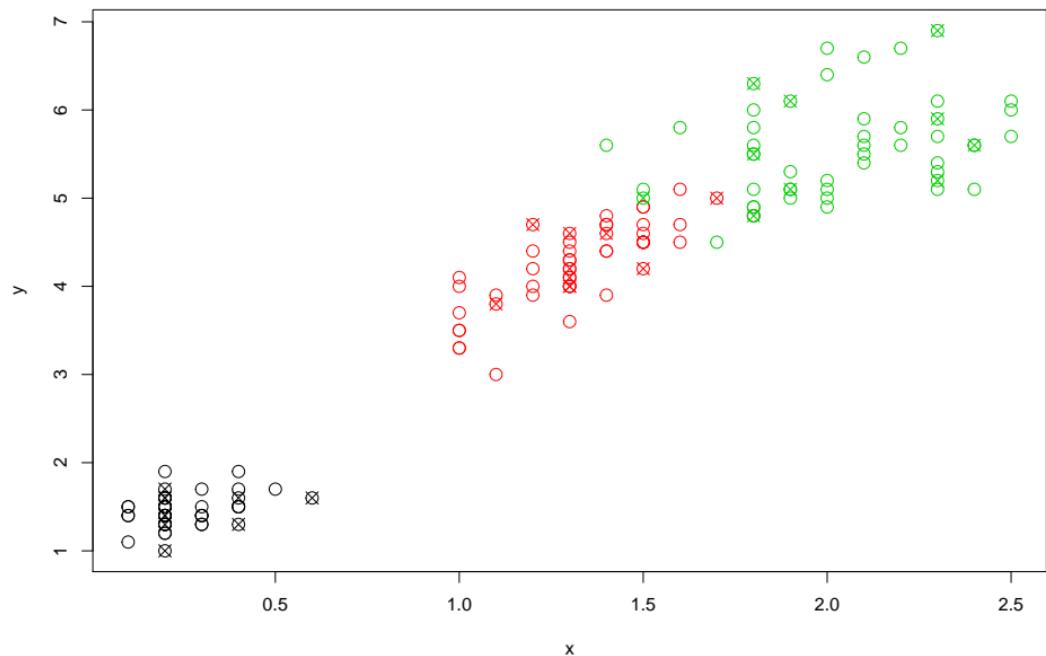
Einfache Zufallsstichprobe

30 zufällig ausgewählte Beobachtungen:



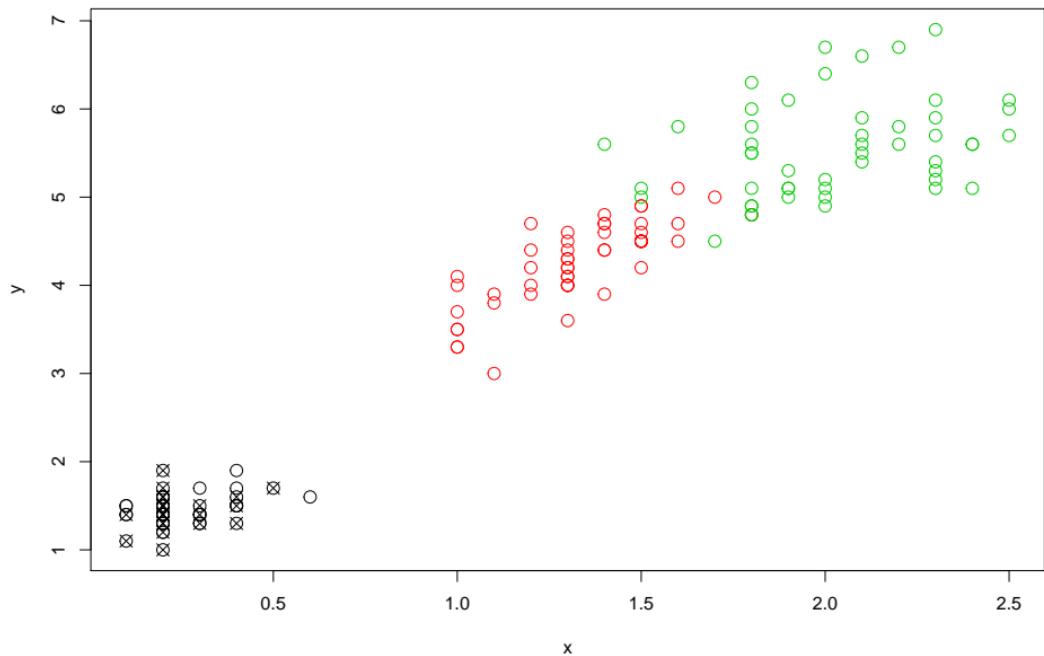
Geschichtete Zufallsstichprobe

Von jeder Farbe 10 zufällig ausgewählte Beobachtungen:



Gelegenheitsstichprobe

Von den ersten 50 Beobachtungen 30 zufällig ausgewählte:



Beobachtungsstudien und Experimente

- ▶ Bei **Beobachtungsstudien** werden Daten gesammelt, ohne die Entstehung der Daten zu beeinflussen (keine unmittelbaren Kausalaussagen möglich).
- ▶ Bei einem **Experiment** wird der Wert der unabhängigen Variable(n) manipuliert²⁰ und die Variation der abhängigen Variable gemessen.
 - ▶ Um Verzerrungen durch Kovariablen zu vermeiden, erfolgt die Zuordnung zu den Experimentalkonditionen **zufällig (randomisiert)**.²¹
 - ▶ Durch wiederholte Messung kann der Effekt der Experimentalkonditionen geschätzt werden: hohe interne Validität. Bei Quasi-Experimenten ist die Zuordnung nicht randomisiert: geringe interne Validität.

²⁰z. B. Zielgruppe erhält Werbung, Kontrollgruppe nicht

²¹Video <https://www.causeweb.org>: McLellan M © Randomize

Beispiel Experiment

Unterscheidet sich die Einschätzung (Rating) eines Unternehmens in Abhängigkeit davon, ob die Person alleiniger Entscheider oder derjenige ist, der die Entscheidungsvorlage vorbereitet?

Um dies zu untersuchen, wurde Studierenden eine Fallstudie mit der Bitte um Einschätzung gegeben, wobei **zufällig** zugeordnet wurde, ob es sich um einen alleinigen Entscheider oder um den Ersteller einer Entscheidungsvorlage handelt.²²

²²Hose, C., Lübke, K., Nolte, T., und Obermeier, T. (2012): *Ratingprozess und Ratingergebnis: Ein Experiment bei qualitativen Ratingkriterien*, Kredit & Rating Praxis (6), 12-14.

Die Dozentin stellt fest, dass die Motivation der Studierenden mit der Uhrzeit zusammenhängt, und zwar unterschiedlich für Frauen und Männer. Welche Aussage stimmt?

- A. Es handelt sich um eine Beobachtungsstudie.
- B. Es handelt sich um ein randomisiertes Experiment.

Warum ist die interne Validität bei einem randomisierten Experiment höher als z. B. bei Beobachtungsstudien?

- ▶ Bei **Laborexperimenten** erfolgt die Untersuchung innerhalb einer speziellen Versuchsanordnung (geringe externe Validität).
- ▶ Bei einem **Feldexperimenten** erfolgt die Untersuchung im natürlichen Umfeld (hohe externe Validität).

Schlussmöglichkeiten

	zufällige Zuordnung	keine zufällige Zuordnung
zufällige Stichprobe	Kausalschluss, generalisierbar für die Population	kein Kausalschluss, Aussage generalisierbar für die Population
keine zufällige Stichprobe	Kausalschluss, nur für die Stichprobe	kein Kausalschluss, Aussage nur für die Stichprobe

Big Data

*Data are not just numbers, they are numbers with a context.*²³

Big Data wird häufig durch 4 Vs charakterisiert:

- ▶ **V**olume: schiere Masse an Daten
- ▶ **V**elocity: Geschwindigkeit – immer schneller, immer neuere Daten (Sensoren, Tweeds, ...)
- ▶ **V**ariety: Vielfalt – immer mehr verschiedene Datentypen (Posts, Bilder, Videos, ...)
- ▶ **V**eracity: Stimmigkeit / Vertrauenswürdigkeit der Daten – und damit der Schlussfolgerungen

²³ Cobb, G. W. und Moore, D. S. (1997): Mathematics, Statistics, and Teaching, The American Mathematical Monthly, 104(9) 801-823.

Personenbeziehbare Daten und unternehmensinterne Daten sind sensibel!

Rechtliche Rahmen u. a.:

- ▶ Bundesdatenschutzgesetz
- ▶ EU Datenschutz-Grundverordnung

Nicht alles was möglich ist, ist auch legal!

Tipps: Datenerhebung

- ▶ Liste mit Datenquellen: <https://www.fom.de/forschung/institute/ifes/studium-und-lehre/datenquellen.html>
- ▶ Hinweise zu Fragebögen: https://www.fom.de/fileadmin/fom/forschung/ifes/141112_Leitfaden_operative_Umsetzung_quantitativer_Befragungen.pdf

Griechische Buchstaben in den Folien

- ▶ α : *alpha*, i. d. R. Symbol für das Signifikanzniveau eines Tests, auch Zeichen für Fehler 1. Art.
- ▶ β : *beta*, i. d. R. Symbol für Regressionskoeffizienten, auch Zeichen für Fehler 2. Art.
- ▶ δ : *delta*, i. d. R. Symbol für allgemeine zusammenfassende Statistik (Kennzahl).
- ▶ ϵ : *epsilon*, i. d. R. Symbol für Residuum.
- ▶ μ : *my*, i. d. R. Symbol für den Populationsmittelwert.
- ▶ σ : *sigma*, i. d. R. Symbol für die Populationsstandardabweichung.
- ▶ π : *pi*, i. d. R. Symbol für den Populationsanteil.

1. Forschungsfrage: Was soll untersucht werden?
2. Studiendesign: Operationalisierung / Variablenauswahl. Wahl des Stichprobenverfahren und / oder Versuchsplanung. Alternativ: Nutzung vorhandener Daten.
3. Datenerhebung
4. Datenanalyse: Datenvorverarbeitung (Ausreißer, fehlende Werte), Explorative Datenanalyse (Grafiken und Kennzahlen).
5. Inferenz: Schätzen und Testen; Modellierung.
6. Schlussfolgerungen: (vorläufige) Antwort auf Forschungsfrage.

Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.
– Donald Knuth

4 Einführung R

[...] she was also following a wider trend: for many academics [...] R is the data-analysis tool of choice.²⁴

Verbreitung z. B.: <http://r4stats.com/articles/popularity/>

R ist eine weit verbreitete Eintrittskarte in das globale Datenzeitalter!

²⁴Tippmann, S.. Programming tools (2015): Adventures with R. A guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis. Nature, 517, S. 109–110. <https://doi.org/10.1038%2F517109a>

- ▶ Methoden- und Anwendungsvielfalt (Finance, Marketing, HR, Psychologie, ...) ²⁵
- ▶ Neue Methoden der Datenanalyse werden häufig in R entwickelt (auch Big Data, KI, etc.).
- ▶ frei und offen; kostenlos
- ▶ Schnittstellen zu sehr vielen Datenquellen/-banken (auch SocialMedia etc.)
- ▶ Erweiterungen u. a. für Microsoft, Oracle, SAP Produkte, aber auch SPSS, SAS
- ▶ unzählige Nutzer*innen weltweit in Unternehmen und Wissenschaft
- ▶ Möglichkeiten für Reporting, Apps, etc.
- ▶ numerische Stabilität / Genauigkeit
- ▶ große Entwickler*innen-Gemeinde mit langer Geschichte (seit 1993); R Konsortium, u. a. IBM, Microsoft, TIPCO, Google, ...

²⁵ Siehe z. B. <https://cran.r-project.org/web/views/>

- ▶ **R** <https://www.r-project.org/>: ist das Basisprogramm
- ▶ **RStudio Desktop** <https://www.rstudio.com/>: ist eine komfortable Entwicklungsumgebung für R und bietet zusätzliche Tools, wie z. B. Dokumentenerstellung etc.
- ▶ **mosaic** <https://cran.r-project.org/web/packages/mosaic/>: ist ein Zusatzpaket, welches u. a. eine vereinheitlichte R Syntax bietet

1. R (<https://www.r-project.org/>)
2. RStudio Desktop (<https://www.rstudio.com/>)
3. Installation von Zusatzpaketen in RStudio:

```
install.packages("mosaic")
```

Ausführliche Installationsanleitung [hier](#).

Ich glaube, dass die Fähigkeit zum Programmieren eine der Basisfähigkeiten von jungen Menschen wird, neben Lesen, Schreiben, Rechnen. Die werden nicht wegfallen. Aber Programmieren wird nochmal dazu kommen.²⁶

²⁶ Rede von Bundeskanzlerin Merkel zur Deutsch-Französischen Digitalkonferenz am 13. Dezember 2016.

Der Prozess, eine komplexe Aufgabe auf eine Reihe einfacher Anweisungen zu reduzieren - genau darum geht es beim Programmieren -, ist eine Fähigkeit, die in vielen Aspekten des modernen Lebens nützlich ist, nicht nur für professionelle Informatiker und Programmierer.²⁷

I think everyone should learn how to program a computer, because it teaches you how to think.²⁸

²⁷Facebooks Forschungschef Yann LeCun

²⁸Steve Jobs

Don't fence off students from the computation pool, throw them in! Computing skills are essential to working with data in the 21st century. Given this fact, we feel that to shield students from computing is to ultimately do them a disservice.²⁹

²⁹ Ismay, C., Kim, A. (2017): ModernDive

- ▶ Dokumentation des Vorgehens
- ▶ (Einfache) Nachvollziehbarkeit, Wiederholung
- ▶ Möglichkeit zur Automatisierung und Übertragung
- ▶ "Direkte" Kommunikation mit dem Programm / Computer
- ▶ Speziell R: unzählige Literatur und Hilfe / Tutorials im Internet

```
analysiere( y ~ # ggfs. abhängige Variable
             x # unabhängige Variable(n)
             | z, # ggfs. bedingende (gruppierende) Variable(n)
               Optionen, # ggfs. weitere Optionen
               data = daten ) # Datensatz
```

analysiere(): Was soll R tun?³⁰

Hinweis: unter macOS: ~: alt+n, |: alt+7

³⁰Befehlsübersicht [hier](#)

1. Was soll der Computer für mich tun?
2. Was muss der Computer dafür wissen?

```
meineanalyse( meiny ~ meinx, data = meinedaten)
```

- ▶ R unterscheidet zwischen Groß- und Kleinbuchstaben.
- ▶ R verwendet den Punkt . als Dezimaltrennzeichen.
- ▶ Fehlende Werte werden in R durch NA kodiert.
- ▶ Kommentare werden mit dem Rautezeichen # eingeleitet.
- ▶ Eine Ergebniszuzuweisung erfolgt über <- .
- ▶ %>% (Paket dplyr) übergibt Ergebnisse.
- ▶ Hilfe zur Funktion foo: ?foo

5 Explorative Datenanalyse

- ▶ **Balkendiagramm:** Häufigkeit von Merkmalsausprägungen (nominal, ordinal, metrisch diskret)
- ▶ **Histogramm:** Häufigkeit von gruppierten Merkmalsausprägungen (metrisch)
- ▶ **Boxplot:** Visualisierung von Median, oberem und unterem Quartil, Minimum und Maximum, Ausreißern
- ▶ **Streudiagramm / Scatterplot:** Darstellung der Merkmalsausprägungen von zwei i. d. R. metrischen Merkmalen³¹ als Punkte
- ▶ **Mosaikplot:** Darstellung der Merkmalsausprägungen zweier nominaler Merkmale
- ▶ **Liniendiagramm:** Verlauf der Merkmalsausprägung eines Merkmals
- ▶ **Kreisdiagramm**³²

³¹bei kategorialen oder metrisch diskreten Merkmalen ggf. *verwackeln* (engl.: jitter)

³²siehe z. B. Regel 20 von <https://robjhyndman.com/hyndsiht/graphics/>

- ▶ Vermittle viele Zahlen, sonst brauchst du keine Grafik.
- ▶ Vermeide Ablenkung von der Hauptbotschaft.
- ▶ Fördere visuellen Vergleich.
- ▶ Unterschiedliche Farben nur, wenn es den Vergleich unterstützt.
- ▶ Vermeide 3D.
- ▶ Achte auf die Achsenkalierung.

Länge und Breite des Kelch- und Blütenblattes von drei verschiedenen Schwertlilienarten.³³

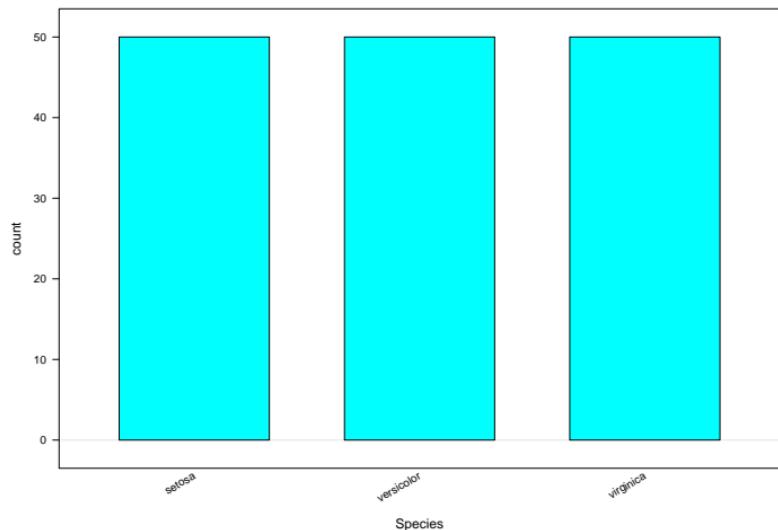


Foto: Armin Hauke

³³Fisher, R. A. (1936): *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, 7, Part II, 179–188.

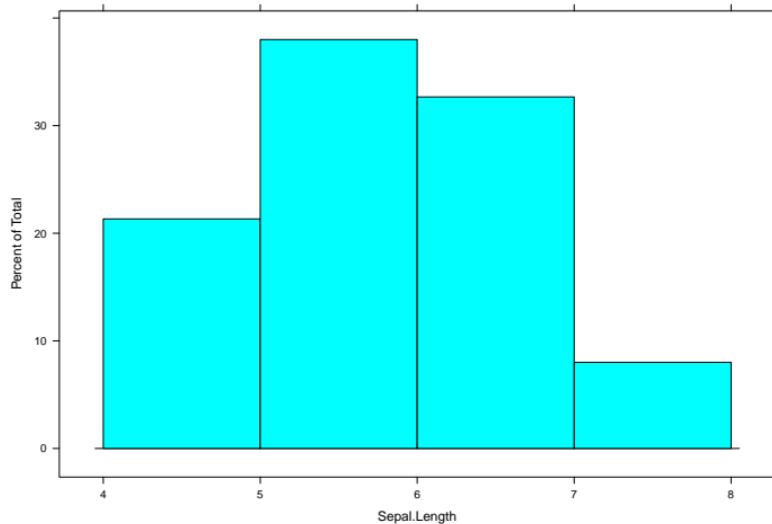
Balkendiagramm

Visualisiert die absoluten oder relativen Häufigkeiten von kategorialen oder metrisch diskreten Variablen.



Histogramm

Visualisiert die (gruppierte) Verteilung einer numerischen Variable.

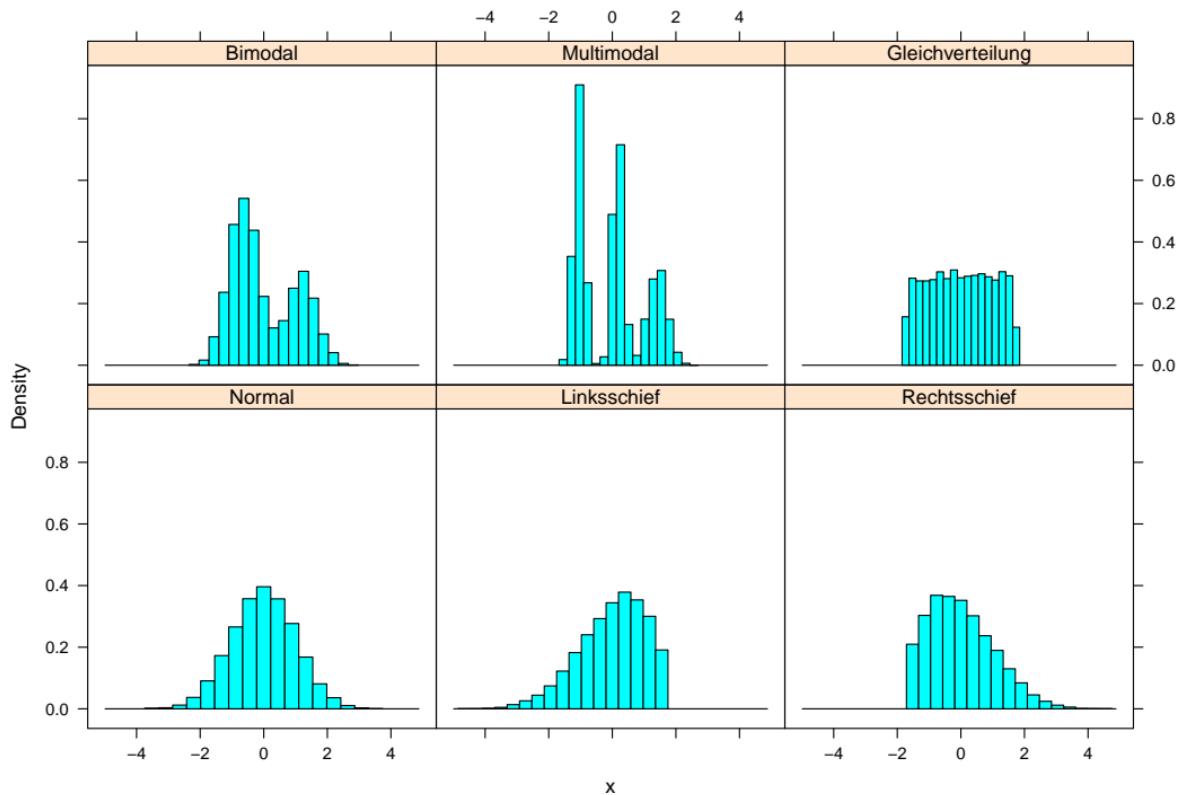


Welche Aussage stimmt?

- A. Die meisten Werte sind ≤ 5 .
- B. Die meisten Werte sind > 5 und ≤ 6 .
- C. Die meisten Werte sind > 6 und ≤ 7 .
- D. Die meisten Werte sind > 7 .

Die Verteilung gibt an, wie häufig bzw. wahrscheinlich bestimmte Werte oder Wertebereiche sind.

- ▶ Schiefe: Bei **rechtsschiefen** (linkssteilen) Verteilungen sind mehr Werte im unteren Wertebereich, bei **linksschiefen** (rechtssteilen) im oberen.
- ▶ Bei **symmetrische** Verteilungen verteilen sich die Daten symmetrisch um eine zentrale Lage.
- ▶ Bei **mehrgipfligen** Verteilungen gibt es mehr als nur ein Zentrum, um das die Werte streuen.



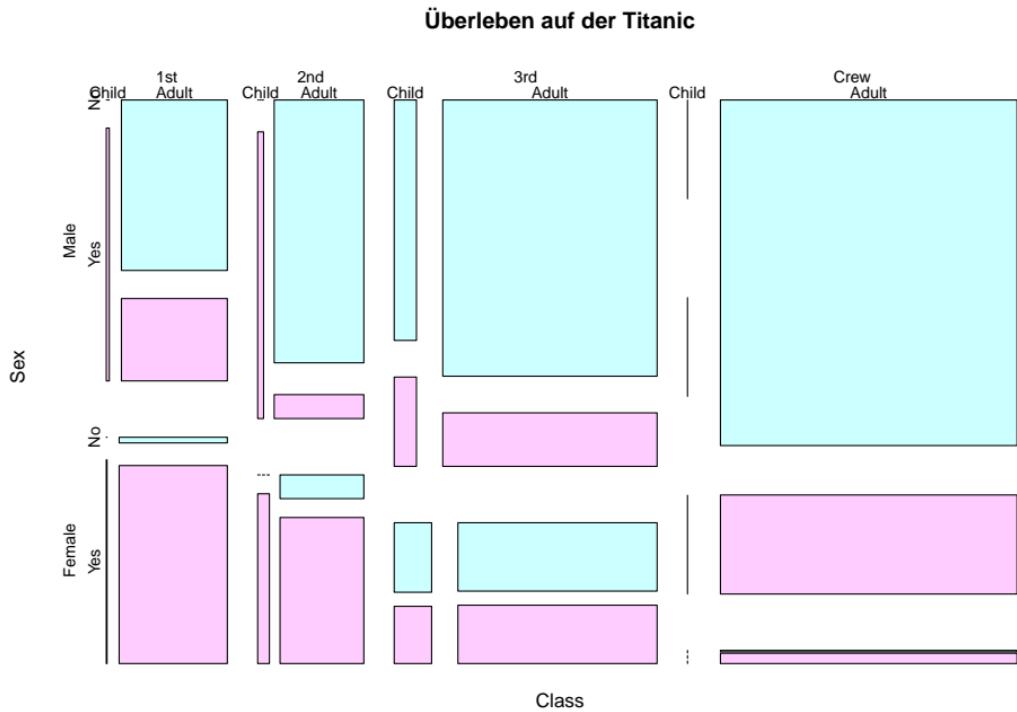
Übung 18: Verteilungsform

Welche Aussage stimmt vermutlich für die Verteilung des Einkommens?

- A. Das Einkommen ist gleichverteilt.
- B. Das Einkommen ist multimodal.
- C. Das Einkommen ist normalverteilt.
- D. Das Einkommen ist linksschief.
- E. Das Einkommen ist rechtsschief.

Mosaikplot

Visualisiert die gemeinsame Verteilung von zwei kategorialen Variablen.

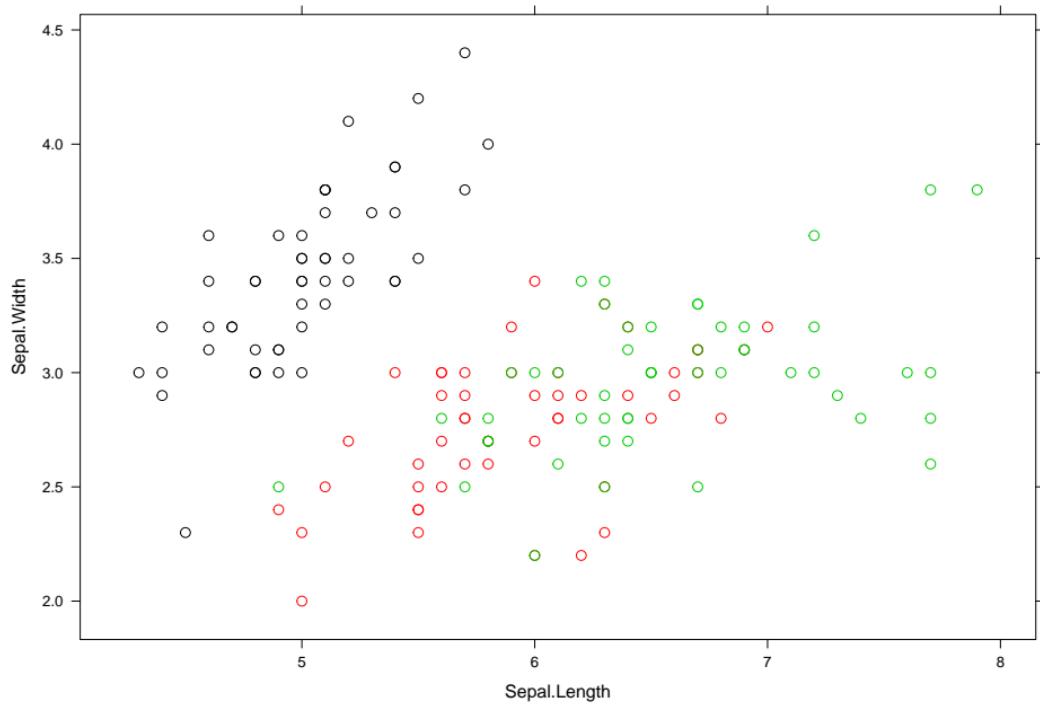


Stimmt die Aussage: Der Anteil der Überlebenden ist in der 1. Klasse größer als in den unteren Klassen?

- ▶ Ja.
- ▶ Nein.

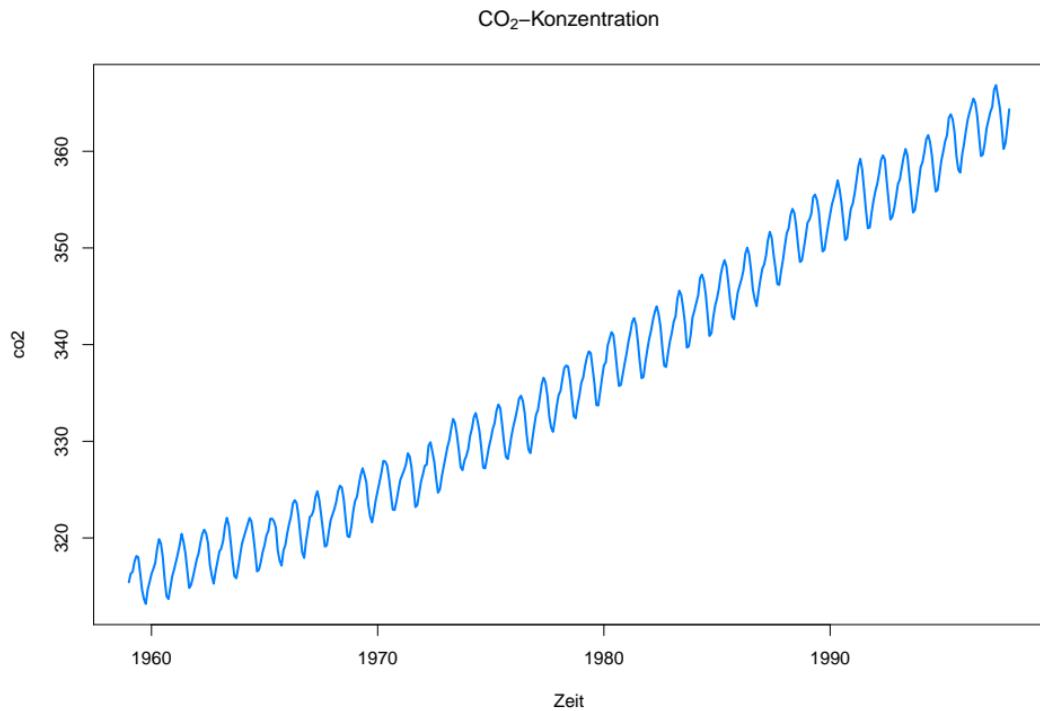
Streudiagramm

Visualisiert die gemeinsame Verteilung von zwei numerischen Variablen.



Liniendiagramm

Visualisiert den (zeitlichen) Verlauf mindestens einer numerischen Variable



Übung 20: Liniendiagramm

Stimmt die Aussage: Die CO₂-Konzentration ist im Laufe der Zeit gestiegen?

- ▶ Ja.
- ▶ Nein.

- ▶ `bargraph()`: Balkendiagramm
- ▶ `histogram()`: Histogramm
- ▶ `bwplot()`: Boxplot
- ▶ `xyplot()`: Streudiagramm
- ▶ `mosaicplot()`: Mosaikplot

Übung 21: Wahl der Visualisierung

Mit welchem Verfahren kann die Verteilung des Merkmals Stundenlohn sinnvoll visualisiert werden?

- A. Balkendiagramm
- B. Histogramm
- C. Streudiagramm

Lagemaße sollen die zentrale Tendenz der Daten beschreiben:

- ▶ **Minimum** bzw. **Maximum**: kleinste bzw. größte Merkmalsausprägung
- ▶ **Modus**/Modalwert: häufigste Merkmalsausprägung
- ▶ **Median**/Zentralwert: Merkmalsausprägung, die bei (aufsteigend) sortierten Beobachtungen in der Mitte liegt
- ▶ **Arithmetischer Mittelwert** (engl. mean)³⁴: Summe aller Werte geteilt durch die Anzahl: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- ▶ **Quantil**: Das p-Quantil ist der Wert, für den gilt, dass er von p Prozent der Werte nicht überschritten wird.

³⁴Darüberhinaus gibt es noch den Geometrischen und den Harmonischen Mittelwert.

Stimmt die Aussage: Die Berechnung des arithmetischen Mittelwertes ist bei nominalen Merkmalen *nicht* sinnvoll?

- ▶ Ja.
- ▶ Nein.

Daten: 20; 18; 24; 40; 24; 22; 21; 23; 20; 28 ($n = 10$)

- ▶ Minimum, Maximum, Modus: $x_{min} = 18$, $x_{max} = 40$, $x_{mod} = \{20; 24\}$
- ▶ Median: 18; 20; 20; 21; $\underbrace{22; 23}_{x_{0,5} = \frac{22+23}{2} = 22,5}$; 24; 24; 28; 40
- ▶ Arithmetischer Mittelwert: $\bar{x} = \frac{1}{10}(20 + 18 + 24 + \dots + 28) = \frac{240}{10} = 24$
- ▶ 25%-Quantil:³⁵ $x_{0,25} = 20$

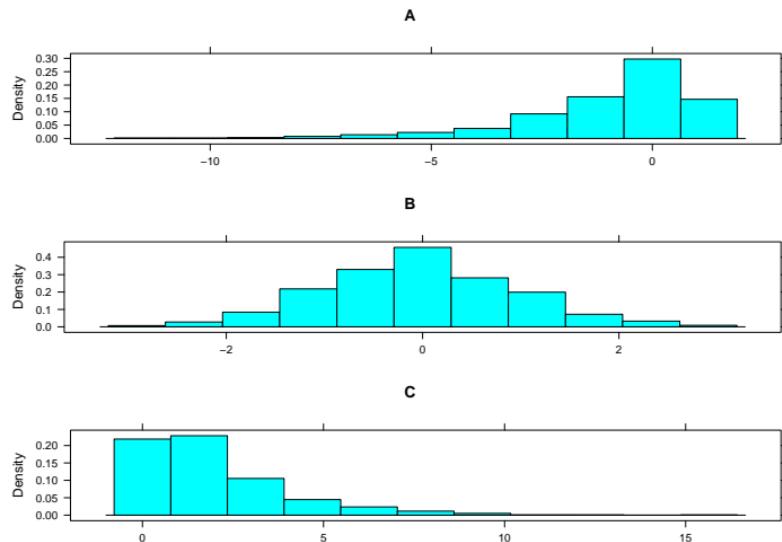
³⁵Hier sind verschiedene Berechnungen möglich. R gibt z. B. 20.25 aus.

Stimmt die Aussage: Der Median ist das 50 % Quantil einer Verteilung?

- ▶ Ja.
- ▶ Nein.

- ▶ Der arithmetische Mittelwert ist der Durchschnitt in dem Sinne, dass alle Merkmalsträger den gleichen Anteil an der Merkmalssumme haben.
- ▶ Der Median ist die Merkmalsausprägung eines (im Sinne des Merkmals) typischen, d. h. mittleren Merkmalsträgers.
- ▶ Der Median ist robust gegen Ausreißer, der arithmetische Mittelwert nicht.

Übung 24: Vergleich Median und Mittelwert



Für welche Abbildung gilt wohl Median < arithmetischer Mittelwert?

- A. Abbildung A.
- B. Abbildung B.
- C. Abbildung C.

Streuungsmaße sollen die Streuung / Variation der Daten beschreiben:

- ▶ **Varianz**: Maß für die durchschnittliche quadratische Abweichung zum Mittelwert:
 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Diese hat aber eine andere Einheit als die Daten, z. B. Daten in €, Varianz €².
- ▶ **Standardabweichung** (engl. standard deviation): Quadratwurzel der Varianz:
 $sd = s = \sqrt{s^2}$
- ▶ **Interquartilsabstand** (engl. interquartile range, IQR): oberes Quartil (75%-Quantil) – unteres Quartil (25%-Quantil)
- ▶ **Spannweite** (engl. range): Maximum – Minimum

Beispielrechnung Streuungsmaße

Daten: 20; 18; 24; 40; 24; 22; 21; 23; 20; 28, $n = 10$, $\bar{x} = 24$

- ▶ Varianz: $s^2 = \frac{1}{10-1} ((20-24)^2 + (18-24)^2 + \dots + (28-24)^2) = \frac{354}{9} \approx 39,33$
- ▶ Standardabweichung: $s = \sqrt{39,33} = 6,27$
- ▶ Interquartilsabstand:³⁶ $IQR = 24 - 20 = 4$
- ▶ Spannweite: $40 - 18 = 22$.

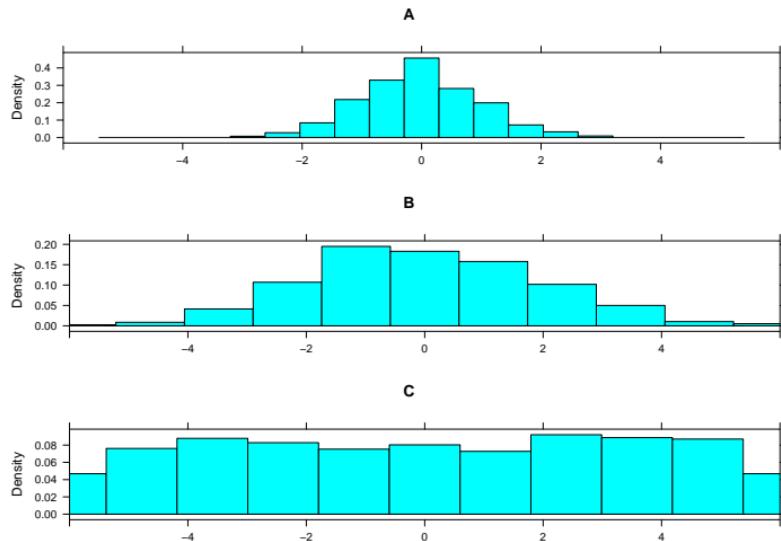
³⁶Hier sind aufgrund verschiedener Berechnungsmöglichkeiten der Quantile unterschiedliche Werte möglich. R gibt z. B. 3.75 aus.

Übung 25: Streuungsmaße

Welche Aussage stimmt?

- A. Die Standardabweichung ist robuster (gegen Ausreißer) als der Interquartilsabstand.
- B. Der Interquartilsabstand ist robuster (gegen Ausreißer) als die Standardabweichung.
- C. Interquartilsabstand und Standardabweichung sind gleich robust gegen Ausreißer.

Übung 26: Vergleich Streuung



Bei welcher Abbildung ist die Standardabweichung sd wohl am Größten?

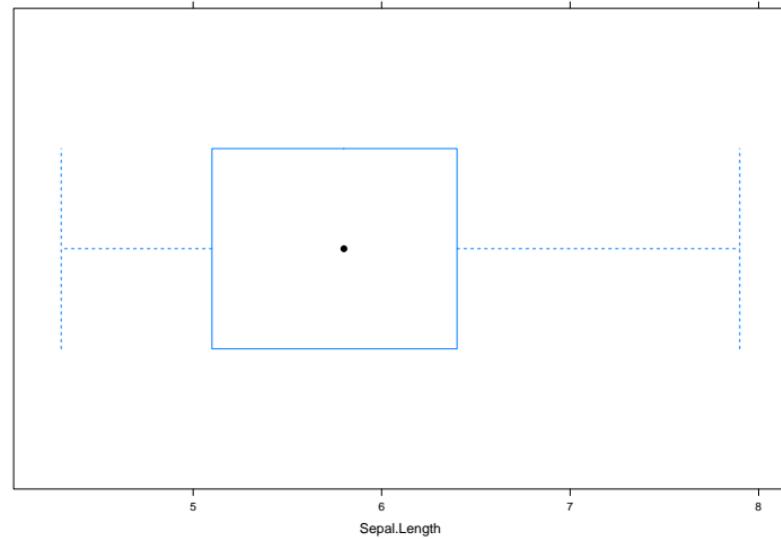
- A. Abbildung A.
- B. Abbildung B.
- C. Abbildung C.

Bilden Sie Gruppen von 4-8 Personen und analysieren Sie die Anzahl Stunden, die Sie heute Nacht geschlafen haben. Berechnen Sie arithmetischen Mittelwert, Median und Standardabweichung.³⁷

³⁷Aus Datenschutzgründen dürfen Sie lügen!

Boxplot

Visualisiert die Verteilung von deskriptiven Kennzahlen und mögliche Ausreißer einer numerischen Variable.

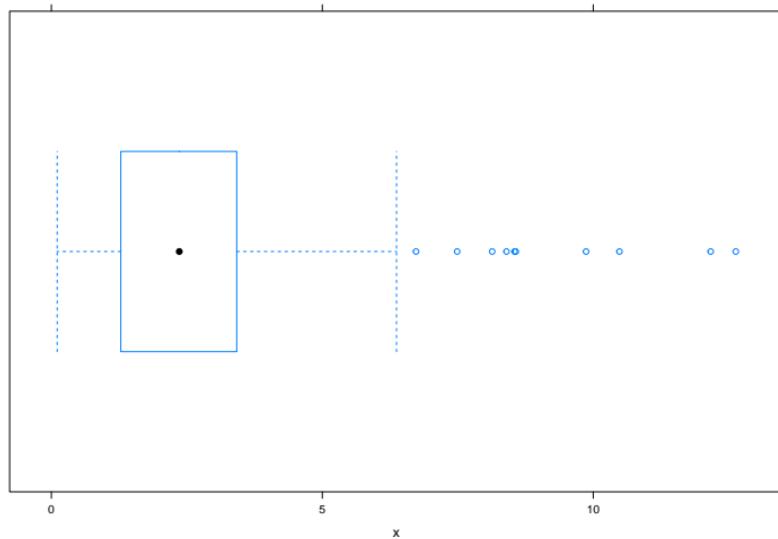


- ▶ Die untere Linie der Box ist das untere Quartil (Q1).
- ▶ Die obere Linie der Box ist das obere Quartil (Q3).
- ▶ Der Punkt in der Box (häufig auch eine Linie) ist der Median.
- ▶ Sollten Punkte außerhalb der Antennen sein, sind dies mögliche Ausreißer. Maximale Reichweite der Antennen: $1,5 \cdot IQR$ vom oberen bzw. unteren Quartil. Sollte das Maximum bzw. das Minimum der Daten kleiner bzw. größer sein, wird dies genommen.³⁸

³⁸Definition nicht immer einheitlich.

Offene Übung 28: Boxplot

Verbinde Abbildung und Kennzahlen. Ab wann ist eine Beobachtung ein potentieller Ausreißer nach oben?



```
##   min   Q1 median   Q3   max mean    sd    n missing
## 0.11 1.3  2.36 3.4 12.63 2.96 2.65 100      0
```

- ▶ **Kovarianz** beschreibt den linearen Zusammenhang zweier metrischer Merkmale:
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$
- ▶ Der **Korrelationskoeffizient** $r = \frac{s_{xy}}{sd_x \cdot sd_y}$ normiert die Kovarianz auf den Wertebereich -1 bis $+1$ durch Division der Kovarianz durch das Produkt der Standardabweichungen.
- ▶ Korrelationskoeffizienten $r > 0$ zeigen einen positiven linearen Zusammenhang an, $r < 0$ einen negativen. Je größer $|r|$, desto größer ist der lineare Zusammenhang.
- ▶ **Achtung:** Korrelation heißt nicht zwangsläufig Kausalität, keine Korrelation heißt nicht zwangsläufig kein Zusammenhang oder keine Kausalität.³⁹

³⁹ Scheinkorrelation, siehe z. B. <http://www.tylervigen.com/spurious-correlations>

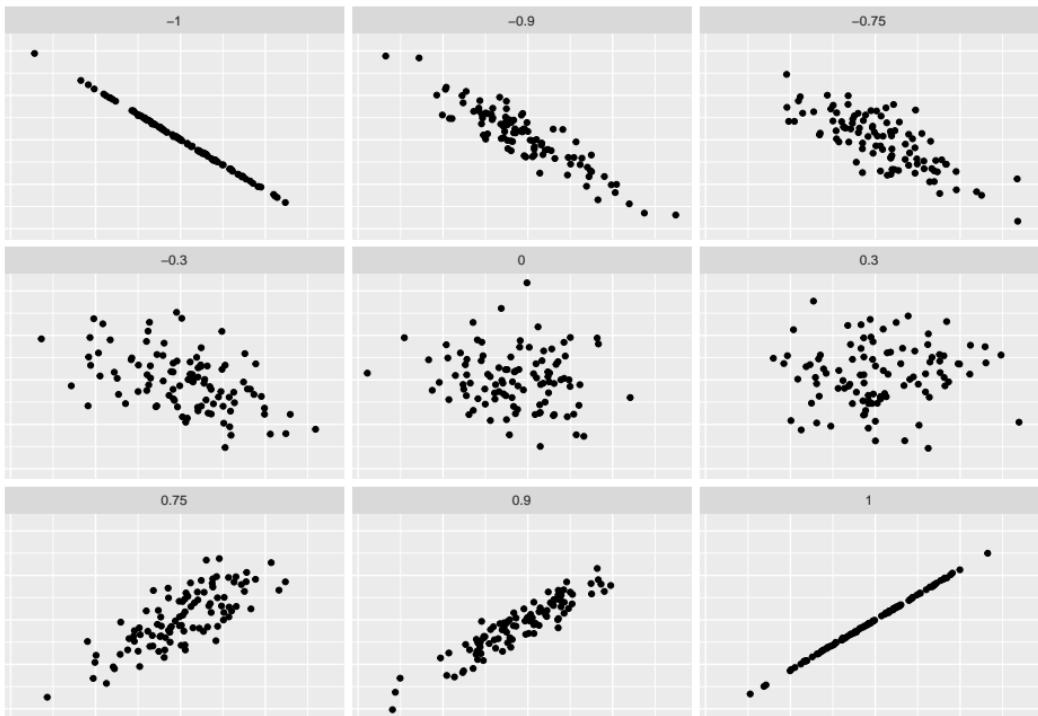
Beispielrechnung Kovarianz und Korrelation

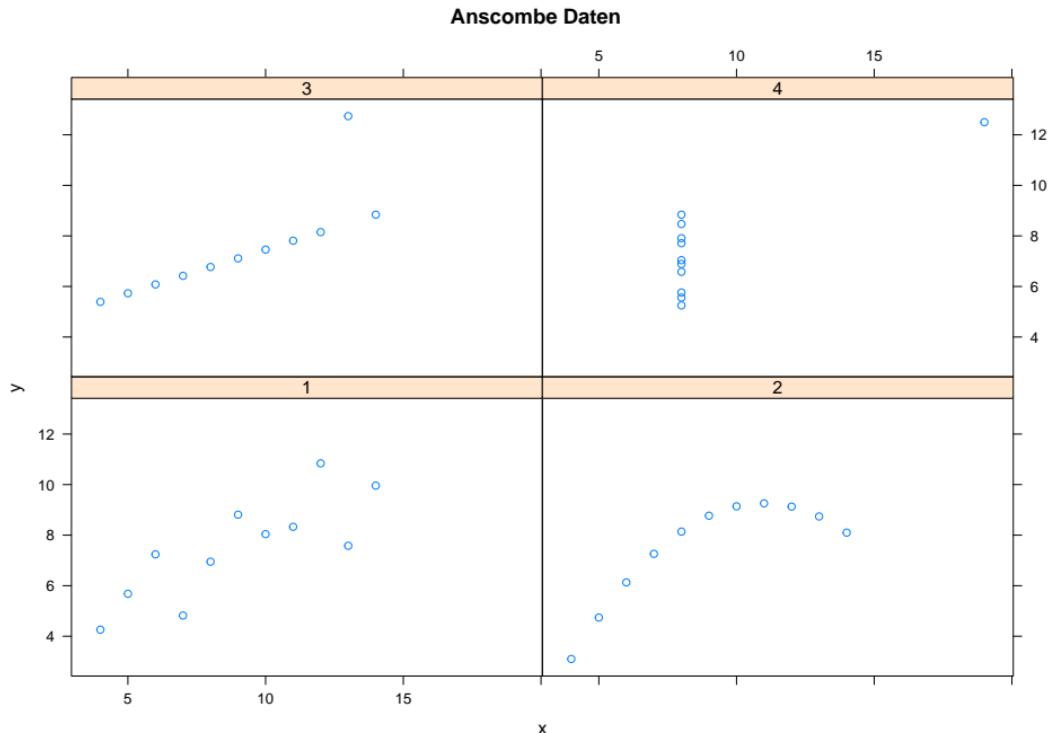
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y}_i)^2$	$(x_i - \bar{x})(y_i - \bar{y}_i)$
1	20	6	-4	-1	16	1	4
2	24	7	0	0	0	0	0
3	30	10	6	3	36	9	18
4	25	7	1	0	1	0	0
5	21	5	-3	-2	9	4	6
\sum	120	35	0	0	62	14	28

- ▶ Lagemaße: $\bar{x} = \frac{120}{5} = 24$; $\bar{y} = \frac{35}{5} = 7$
- ▶ Streuungsmaße:
 $s_x^2 = \frac{62}{4} = 15,5$; $s_y^2 = \frac{14}{4} = 3,5$; $s_x = \sqrt{15,5} = 3,94$; $s_y = \sqrt{3,5} = 1,87$
- ▶ Kovarianz: $s_{xy} = \frac{28}{4} = 7$
- ▶ Korrelation: $r = \frac{7}{3,94 \cdot 1,87} = 0,95$

5. Explorative Datenanalyse

Korrelationskoeffizienten





Die Verteilung von x und y unterscheidet sich sichtbar. Aber die deskriptiven Kennzahlen

- ▶ $\bar{x} = 9; \bar{y} = 7,5$
- ▶ $sd_x \approx 3,31; sd_y \approx 2,03$
- ▶ $r \approx 0,82$

sind nahezu identisch – in allen vier Fällen.⁴⁰

⁴⁰Weiteres Beispiel z. B. unter <https://www.autodeskresearch.com/publications/samestats>

Stimmt die Aussage: Der Korrelationskoeffizient ist robust gegen Ausreißer?

- ▶ Ja.
- ▶ Nein.

- ▶ `favstats()`: Kennzahlen numerischer Variablen
- ▶ `prop()`: Anteile
- ▶ `tally()`: (Kreuz-)tabellierung
- ▶ `cor()`: Korrelationskoeffizient

Einlesen der *Tipping*⁴¹ Daten:

```
# Herunterladen  
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")  
# Einlesen in R  
tips <- read.csv2("tips.csv")  
  
# Alternativ - heruntergeladene Datei einlesen:  
# tips <- read.csv2(file.choose())
```

⁴¹Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

Ein Kellner sammelte über mehrere Monate Daten über sein Trinkgeld:

- ▶ total_bill: Rechnungshöhe in Dollar
- ▶ tip: Trinkgeld in Dollar
- ▶ sex: Geschlecht des Rechnungszahlenden
- ▶ smoker: Gab es Raucher*innen am Tisch?
- ▶ day: Wochentag
- ▶ time: Tageszeit / Mahlzeit
- ▶ size: Anzahl Personen am Tisch

```
# Ggfs. einmalig vorab installieren  
# install.packages("mosaic")  
  
# Paket mosaic laden  
library(mosaic)
```

```
inspect(tips)
```

```
##  
## categorical variables:  
##      name   class levels   n missing          distribution  
## 1   sex   factor     2 244       0 Male (64.3%), Female (35.7%)  
## 2 smoker factor     2 244       0 No (61.9%), Yes (38.1%)  
## 3   day   factor     4 244       0 Sat (35.7%), Sun (31.1%), Thur (25.4%) ...  
## 4   time  factor     2 244       0 Dinner (72.1%), Lunch (27.9%)  
##  
## quantitative variables:  
##      name   class min    Q1 median     Q3   max      mean        sd   n missing  
## 1 total_bill numeric 3.07 13.3475 17.795 24.1275 50.81 19.785943 8.9024120 244      0  
## 2      tip numeric 1.00  2.0000  2.900  3.5625 10.00  2.998279 1.3836382 244      0  
## 3      size integer 1.00  2.0000  2.000  3.0000  6.00  2.569672 0.9510998 244      0
```

Übung 30: Metrische Variablen

Wie viele metrische Variablen liegen vor?

- A. 2
- B. 3
- C. 4
- D. 7
- E. 244

Was vermuten Sie: Um welche Form der Datenerhebung handelt es sich hier?

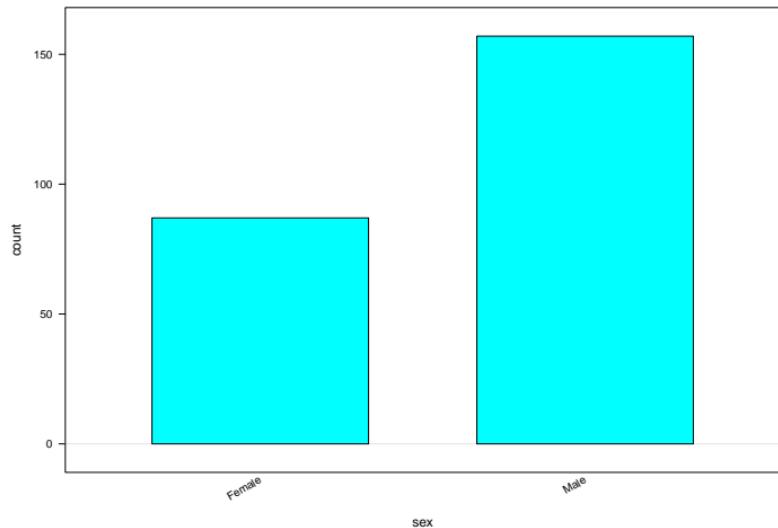
- A. Beobachtungsstudie.
- B. Experiment.

Was folgt daraus?

Analyse: Geschlecht Rechnungszahler*in

Analysiere über Balkendiagramm:

```
bargraph(~ sex, # (unabhängige) Variable, die analysiert wird  
        data = tips) # Datensatz
```



Welche Aussage stimmt?

- A. Bei einer Mehrheit der Stichprobe zahlt eine Frau.
- B. Bei einer Mehrheit der Stichprobe zahlt ein Mann.
- C. Weiß nicht.

Analysiere über Anteil:

```
prop(~ sex, # Variable, die analysiert wird  
     success = "Female", # Ausprägung  
     data = tips) # Datensatz
```

```
##      Female  
## 0.3565574
```

Analysiere über Tabellen:

```
tally( ~ sex, # Variable, die analysiert wird  
      data = tips) # Datensatz
```

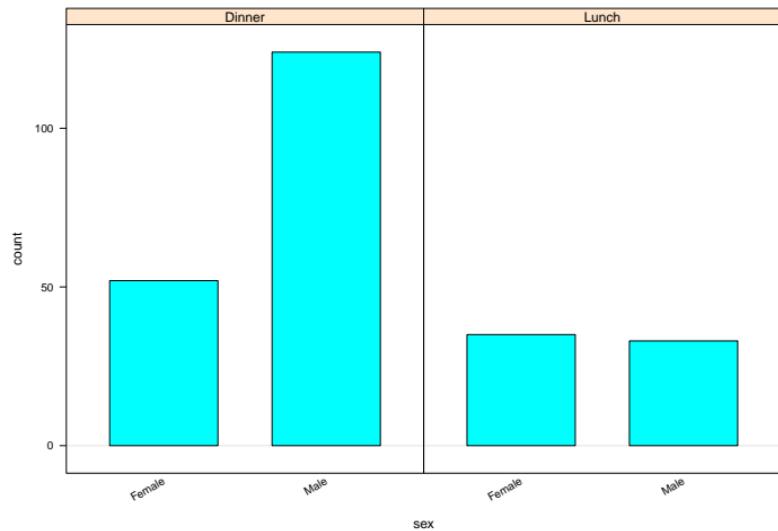
```
## sex  
## Female    Male  
##     87     157
```

```
tally( ~ sex, # Variable, die analysiert wird  
      format = "proportion", # Option: Anteile  
      data = tips) # Datensatz
```

```
## sex  
##     Female      Male  
## 0.3565574 0.6434426
```

Gruppiertes Balkendiagramm

```
bargraph( ~ sex # Variable, die analysiert wird  
| time, # Variable, nach der bedingt wird  
data = tips) # Datensatz
```



Übung 33: Geschlecht nach Tageszeit

Welche Aussage stimmt?

- A. Beim Lunch zahlen mehr Frauen als Männer.
- B. Beim Lunch zahlen weniger Frauen als Männer.
- C. Beim Lunch zahlen gleich viele Frauen wie Männer.

Kreuztabellierung Geschlecht nach Tageszeit

```
tally( ~ sex # Variable, die analysiert wird
      | time, # Variable, nach der bedingt wird
      data = tips) # Datensatz

##           time
## sex      Dinner Lunch
##   Female     52    35
##   Male       124   33
```

Kreuztabellierung Geschlecht nach Tageszeit: Anteile

```
tally( ~ sex # Variable, die analysiert wird  
      | time, # Variable, nach der bedingt wird  
      format = "proportion", # Option: Anteile  
      data = tips) # Datensatz
```

```
##           time  
## sex      Dinner     Lunch  
##   Female  0.2954545 0.5147059  
##   Male    0.7045455 0.4852941
```

Übung 34: Raucher je Wochentag

Welcher Befehl führt eine Kreuztabellierung der Anteile der Raucher je Wochentag durch?

- A. `tally(~ smoker | day, format = 'proportion', data = tips)`
- B. `tally(~ day | smoker, format = 'proportion', data = tips)`

Kreuztabellierung Raucher und Wochentag

```
tally( ~ smoker | day,
      format = "proportion", data = tips)

##          day
## smoker      Fri       Sat       Sun      Thur
##   No  0.2105263 0.5172414 0.7500000 0.7258065
##   Yes 0.7894737 0.4827586 0.2500000 0.2741935

tally( ~ day | smoker,
      format = "proportion", data = tips)

##          smoker
## day           No        Yes
##   Fri  0.02649007 0.16129032
##   Sat  0.29801325 0.45161290
##   Sun  0.37748344 0.20430108
##   Thur 0.29801325 0.18279570
```

Was ist an diesem Befehl falsch?

```
tally( ~ x data = daten)
```

- A. Es fehlt eine Option.
- B. Es fehlt eine bedingende Variable.
- C. Es fehlt ein Komma.
- D. Gar nichts.

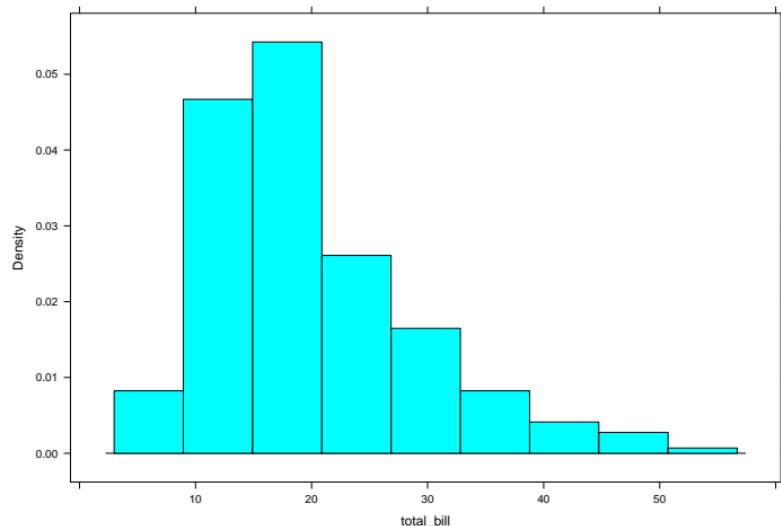
Was ist an diesem Befehl falsch?

```
Tally( ~ x, data = daten)
```

Analyse Rechnungshöhe

Analysiere über Histogramm:

```
histogram( ~ total_bill, # Variable, die analysiert wird  
          data = tips) # Datensatz
```



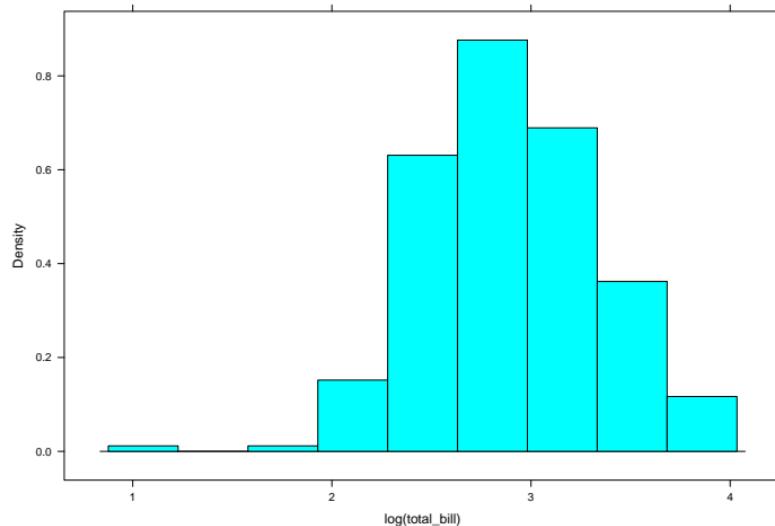
Übung 37: Rechnungshöhe

Welche der folgenden Aussagen stimmt?

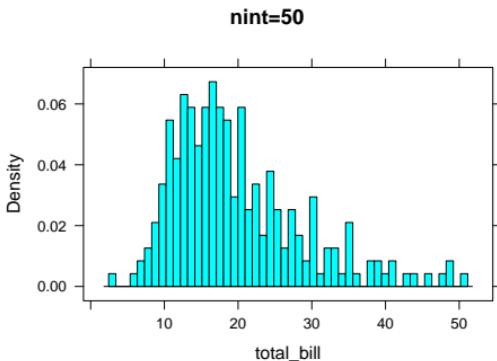
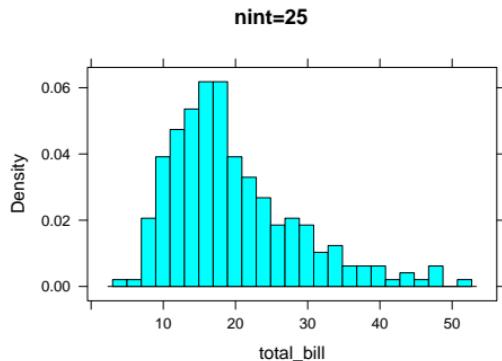
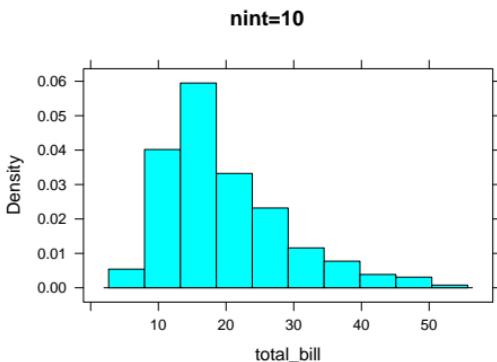
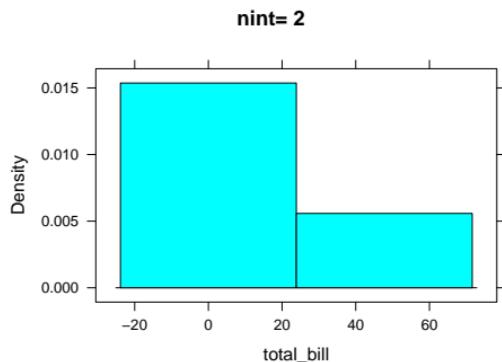
- A. Die Rechnungshöhe ist gleichverteilt.
- B. Die Rechnungshöhe ist multimodal.
- C. Die Rechnungshöhe ist normalverteilt.
- D. Die Rechnungshöhe ist linksschief.
- E. Die Rechnungshöhe ist rechtsschief.

Ggf. können Variablen durch Transformationen (z. B. $\sqrt{()}$, $\ln()$, ...) in Richtung einer symmetrischen Normalverteilung transformiert werden:

```
histogram( ~ log(total_bill), # logarithmierte Variable  
           data = tips) # Datensatz
```



Histogram: Anzahl der Rechtecke festlegen mit Option nint=



Analysiere über Kennzahlen:

```
favstats( ~ total_bill, # Variable, die analysiert wird  
          data = tips) # Datensatz  
  
##      min      Q1 median      Q3      max      mean       sd      n missing  
##  3.07 13.3475 17.795 24.1275 50.81 19.78594 8.902412 244          0
```

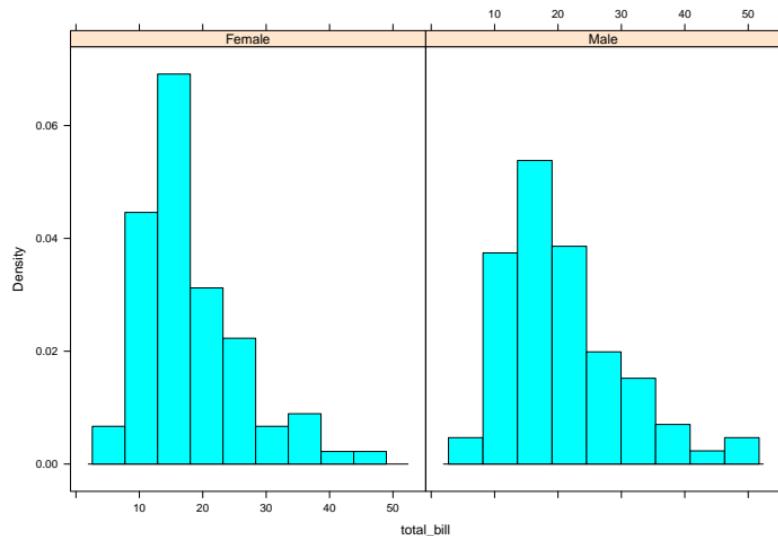
Übung 38: Kennzahlen

Welche Aussage stimmt?

- A. Die durchschnittliche Rechnungshöhe ist kleiner als die Rechnungshöhe einer im Bezug auf die Rechnungshöhe typischen Rechnung.
- B. Die durchschnittliche Rechnungshöhe ist größer als die Rechnungshöhe einer im Bezug auf die Rechnungshöhe typischen Rechnung.
- C. Die durchschnittliche Rechnungshöhe ist gleich der Rechnungshöhe einer im Bezug auf die Rechnungshöhe typischen Rechnung.

Rechnungshöhe je Geschlecht

```
histogram( ~ total_bill # Variable, die analysiert wird  
| sex, # Variable, nach der bedingt wird  
data = tips) # Datensatz
```



Übung 39: Rechnungshöhe nach Geschlecht

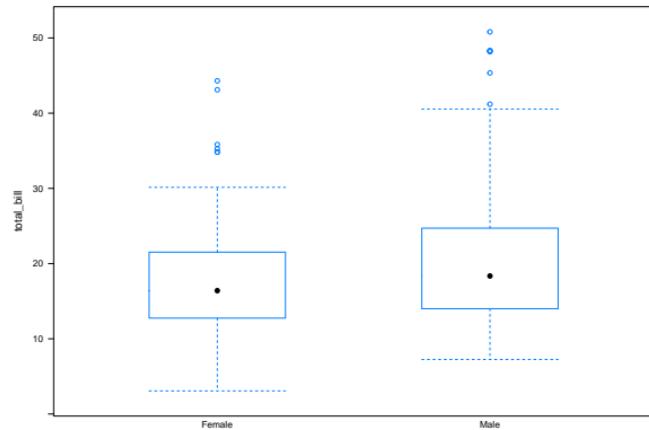
Welche Aussage stimmt nach der Abbildung?

- A. Männer haben einen höheren Anteil an höheren Rechnungen.
- B. Frauen haben einen höheren Anteil an höheren Rechnungen.
- C. Die Verteilung ist bei den Frauen linksschief.
- D. Die Verteilung ist bei den Männern linksschief.

Boxplot Rechnungshöhe abhängig vom Geschlecht

Analysiere über Boxplot⁴²

```
bwplot(total_bill ~ # abhängige Variable  
       sex, # unabhängige Variable  
       data = tips) # Datensatz
```



⁴²Beachte ~ "als Funktion von", | "bedingt, gruppiert nach".

Übung 40: Übung Boxplot

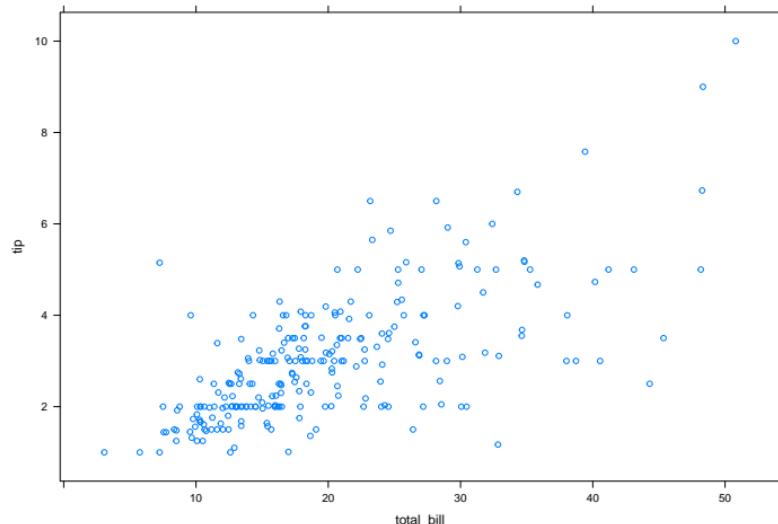
Welche Aussage stimmt nach der Abbildung?

- A. Der Mittelwert der Rechnungshöhe ist bei den Männern unter 20 \$.
- B. Der Mittelwert der Rechnungshöhe ist bei den Männern über 20 \$.
- C. Der Median der Rechnungshöhe ist bei den Männern unter 20 \$.
- D. Der Median der Rechnungshöhe ist bei den Männern über 20 \$.

```
favstats( ~ total_bill # Variable, die analysiert wird  
          | sex, # Variable, nach der bedingt wird  
          data = tips) # Datensatz  
  
##      sex   min   Q1 median   Q3   max     mean       sd      n missing  
## 1 Female 3.07 12.75 16.40 21.52 44.30 18.05690 8.009209  87      0  
## 2   Male 7.25 14.00 18.35 24.71 50.81 20.74408 9.246469 157      0
```

Analysiere über Streudiagramm:

```
xyplot( tip # abhängige Variable  
       ~ total_bill, # unabhängige Variable  
       data = tips) # Datensatz
```



Übung 41: Zusammenhang Rechnungshöhe und Trinkgeld

Welche Aussage stimmt?

- A. Es scheint keinen Zusammenhang zwischen Rechnungshöhe und Trinkgeld zu geben.
- B. Es scheint einen negativen Zusammenhang zwischen Rechnungshöhe und Trinkgeld zu geben.
- C. Es scheint einen positiven Zusammenhang zwischen Rechnungshöhe und Trinkgeld zu geben.

Analysiere über Korrelationskoeffizienten:

```
cor( tip # abhängige Variable  
      ~ total_bill, # unabhängige Variable  
      data = tips) # Datensatz
```

```
## [1] 0.6757341
```

Übung 42: Rechnungs- und relative Trinkgeldhöhe (I / II)

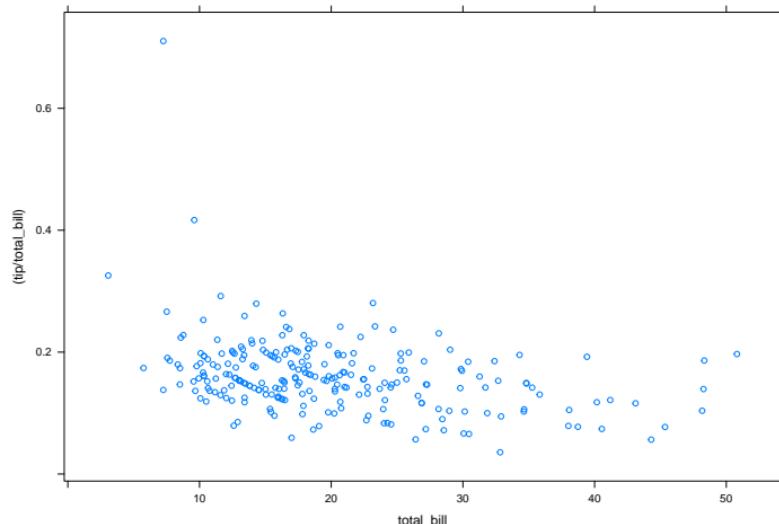
Welcher Befehl visualisiert den Zusammenhang zwischen Rechnungshöhe und der relativen Trinkgeldhöhe $\frac{\text{tip}}{\text{total_bill}}$?

- A. `xyplot(~ (tip/total_bill) | total_bill, data=tips)`
- B. `xyplot((tip/total_bill) ~ total_bill, data=tips)`

Zusammenhang Rechnungs- und relative Trinkgeldhöhe

```
#           AV ~ UV
```

```
xyplot( (tip/total_bill) ~ total_bill, # unabhängige Variable  
       data = tips)                 # Datensatz
```



Übung 43: Rechnungs- und relative Trinkgeldhöhe (II / II)

Welche Aussage stimmt ?

- A. Es gibt Ausreißer nach oben bei der relativen Trinkgeldhöhe.
- B. Es gibt Ausreißer nach unten bei der relativen Trinkgeldhöhe.
- C. Es gibt Ausreißer nach oben bei der Rechnungshöhe.
- D. Es gibt Ausreißer nach unten bei der Rechnungshöhe.

Offene Übung44: Rechnungshöhe für Raucher bzw. Nichtraucher

Was können Sie über die Verteilung der Rechnungshöhe für Raucher bzw. Nichtraucher aussagen?⁴³

⁴³Video <https://www.causeweb.org>: McLellan M © Describe the Distribution

6 Einführung Inferenz

Dreieckstest

- ▶ Drei gleichaussehende Proben, zwei sind gleich, eine **zufällige** ist anders.
- ▶ Der / die Kandidat*in muss herausfinden, welche Probe anders ist.⁴⁴



⁴⁴vgl. ISO 4120 <https://www.iso.org/standard/33495.html>

Übung 45: Dreieckstest

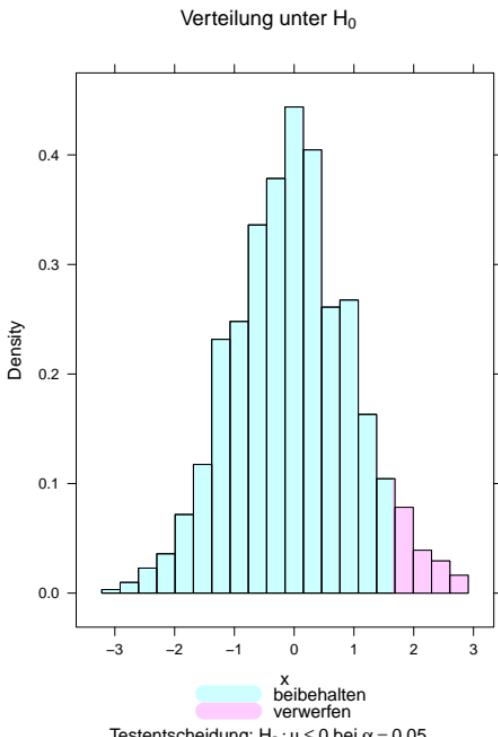
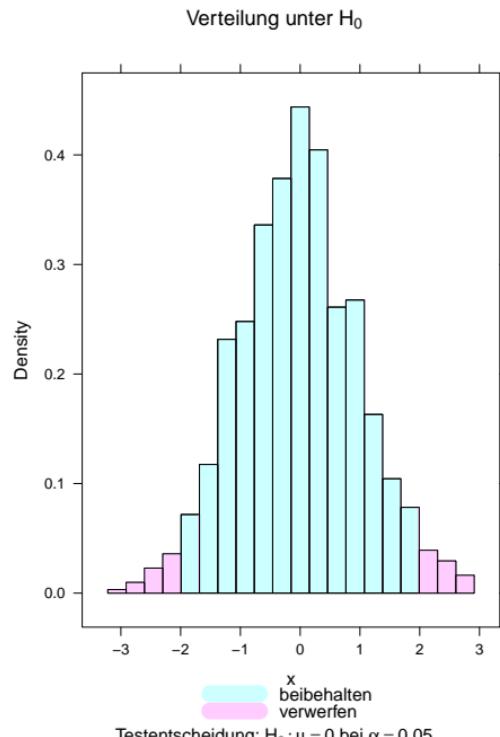
Wie groß ist die Wahrscheinlichkeit π , zufällig, d. h., ohne einen Unterschied zu schmecken, auf die richtige (sprich abweichende) Probe zu tippen?

- A. $\pi = 0$
- B. $\pi = \frac{1}{3}$
- C. $\pi = \frac{1}{2}$
- D. $\pi = \frac{2}{3}$
- E. $\pi = 1$

- ▶ Inhaltliche Hypothesen und Modelle werden mathematisch/statistisch operationalisiert. Dabei ist die **Nullhypothese** H_0 i. d. R. die, dass es keinen Zusammenhang, keinen Unterschied gibt, die **Alternativhypothese** (H_A Forschungshypothese) ist das logische Gegenteil der Nullhypothese. Die Rollen von H_0 und H_A können nicht vertauscht werden.
- ▶ Hypothesen beziehen sich auf die Population / Grundgesamtheit, d. h. π, μ, β .
- ▶ Nullhypothesen können **einseitig, gerichtet** (\leq, \geq) oder **zweiseitig, ungerichtet** ($=$) sein.
- ▶ Anhand einer geeigneten **Teststatistik** δ werden die Stichprobendaten zusammengefasst. Ist die Wahrscheinlichkeit der Teststatistik unter H_0 klein, fällt der Wert der Teststatistik in den **Ablehnungsbereich** und H_0 wird verworfen, andernfalls nicht.
- ▶ Das vorab festgelegte **Signifikanzniveau** α^{45} eines Tests gibt die maximal zugebilligte Irrtumswahrscheinlichkeit dafür an, H_0 zu verwerfen, obwohl H_0 gilt.

⁴⁵üblich: $\alpha = 1\%, 5\%, 10\%$

Gerichtete und ungerichtete Hypothesen



Übung 46: Gerichtete und ungerichtete Hypothesen

Standard ist i. d. R. eine ergebnisoffene, ungerichtete Hypothese.

Was ist der Vorteil einer gerichteten Hypothese gegenüber einer ungerichteten Hypothese, wenn das Interesse nur auf einer Seite liegt?⁴⁶

- A. Wenn die Forschungsthese H_A stimmt, so wird H_0 eher verworfen.
- B. Wenn die Forschungsthese H_A stimmt, so wird H_0 eher beibehalten.
- C. Wenn die Nullhypothese H_0 stimmt, so wird H_0 eher verworfen.
- D. Wenn die Nullhypothese H_0 stimmt, so wird H_0 eher beibehalten.

⁴⁶Eine gerichtete Hypothese muss aber inhaltlich (z. B. Literatur) begründet sein!

- ▶ Die Beweislast liegt bei der Forschungsthese: Wir gehen von H_0 aus: der Angeklagte ist unschuldig, da ist nichts.
- ▶ Wenn die Beweise (Daten) gegen den Angeklagten (H_0) sprechen⁴⁷, haben wir berechtigten Zweifel an der Unschuld. Es gibt Belege für die Forschungsthese H_A .
- ▶ Wenn die Daten nicht ausreichen, um zu zeigen, dass der Angeklagte schuldig ist, so sagen wir nicht: er ist unschuldig. **Daher nie:** wir bestätigen die Nullhypothese, sondern nur, wir können die Nullhypothese nicht verwerfen.

⁴⁷d.h., unter der Unschuldsvermutung (sehr) unwahrscheinlich sind

Übung 47: Hypothese

Wie lautet die richtige Hypothese, um zu überprüfen, ob ein Geschmacksunterschied beim Dreieckstest vorliegt, d. h., die Erkennenswahrscheinlichkeit liegt über reinem "Raten"?

- A. $H_0 : \pi = \frac{1}{2}$ vs. $H_A : \pi \neq \frac{1}{2}$
- B. $H_0 : \pi = \frac{1}{3}$ vs. $H_A : \pi \neq \frac{1}{3}$
- C. $H_0 : \pi \leq \frac{1}{3}$ vs. $H_A : \pi > \frac{1}{3}$
- D. $H_0 : \pi \geq \frac{1}{3}$ vs. $H_A : \pi < \frac{1}{3}$

Schema Hypothesentest

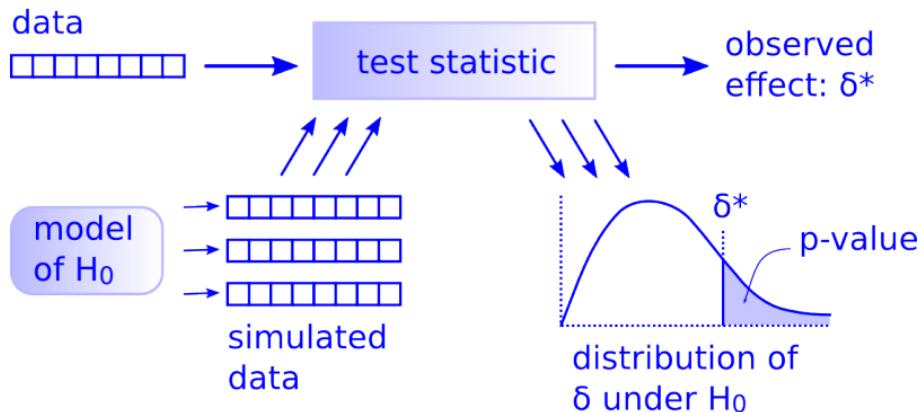


Abbildung: Quelle: Blogbeitrag Allen Downey⁴⁸

Alternative: Verwende theoretische Verteilungsannahmen unter H_0 , häufig approximativ oder asymptotisch.⁴⁹

⁴⁸ <http://allendowney.blogspot.de/2016/06/there-is-still-only-one-test.html>

⁴⁹ Bspw. t -, χ^2 -, F -Verteilungen.

- ▶ Der **p-Wert** berechnet sich aus der (Rand-)Wahrscheinlichkeit der Teststatistik unter H_0 .
- ▶ Gilt $p\text{-Wert} \leq \alpha$ (Signifikanzniveau), so wird H_0 verworfen.
- ▶ **Achtung:** Der p-Wert sagt nicht aus, wie wahrscheinlich H_0 bei den vorliegenden Daten (Teststatistik) ist ($p(H_0|\delta^*)$), sondern wie wahrscheinlich die vorliegende Teststatistik⁵⁰ unter H_0 ist ($p(\delta^*|H_0)$). Der p-Wert sagt nicht, wie relevant ein Ergebnis ist.
- ▶ **Keine** inhaltliche Entscheidung sollte rein auf Basis des p-Wertes getroffen werden.
- ▶ Vor der Testentscheidung **immer** eine explorative Datenanalyse durchführen.

⁵⁰oder noch extremere Werte

Im Rahmen einer Sonderveranstaltung der FOM Dortmund (6.10.2016)⁵¹ und Münster (9.11.2017) tippten von $n = 34$ Teilnehmer*innen $x = 12$ im Rahmen eines Dreieckstest auf die richtige Probe. d. h. das *andere* Bier: Krombacher bzw. Perlenbacher.



Abbildung: Quelle: Anzeige Westfälische Rundschau, 19.9.2016

⁵¹<https://www.fom.de/2016/oktober/kneipe-statt-hoersaal-chef-mit-humor-lohnt-sich.html>

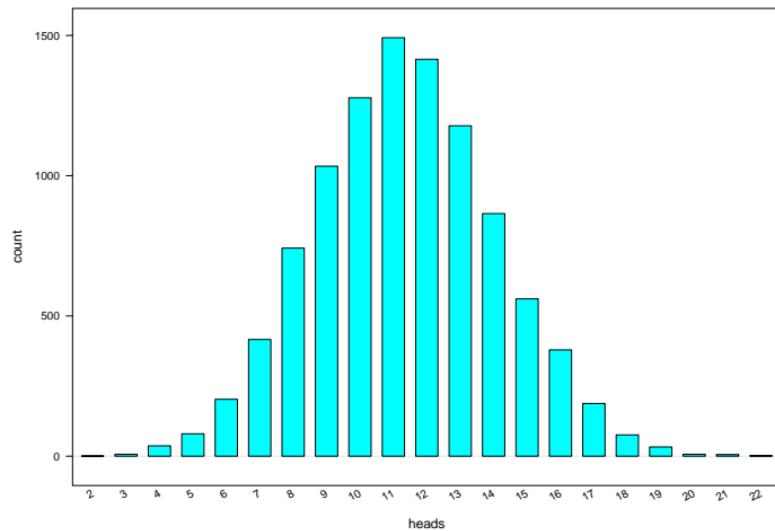
```
# Paket laden, ggf. vorher einmalig installieren
library(mosaic)

# 34facher Münzwurf mit Erfolgsw.keit 1/3
rflip(n = 34, prob = 1/3)

##
## Flipping 34 coins [ Prob(Heads) = 0.333333333333333 ] ...
##
## H T T T H H T T H T T T H T H T T H T T H H H T T H T T T H
## T H
##
## Number of Heads: 14 [Proportion Heads: 0.411764705882353]
```

Simulation Verteilung unter der Nullhypothese

```
set.seed(1896) # Reproduzierbarkeit  
Nullvtlg <- do(10000) * rflip(n = 34, prob = 1/3)  
bargraph(~heads, data = Nullvtlg)
```



Übung 48: Simulation I

Welche der Aussagen stimmt?

- A. Unter der Nullhypothese ist $x = 12$ ein unüblicher, d. h. unwahrscheinlicher, Wert.
- B. Unter der Nullhypothese ist $x = 12$ ein üblicher, d. h. wahrscheinlicher, Wert.
- C. Unter der Nullhypothese ist $x = 10$ ein unüblicher, d. h. unwahrscheinlicher, Wert.
- D. Unter der Nullhypothese ist $x = 20$ ein üblicher, d. h. wahrscheinlicher, Wert.

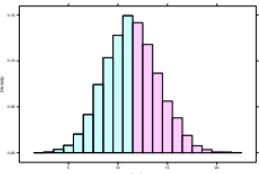
Bei welchem Wert für x hätten Sie wohl Belege für die Forschungsthese $\pi > \frac{1}{3}$?

- A. Bei $x = 5$.
- B. Bei $x = 10$.
- C. Bei $x = 15$.
- D. Bei $x = 20$.

Wie lautet der richtige Befehl, um die Randwahrscheinlichkeit des Stichprobenergebnis $x = 12$ unter der Nullhypothese $H_0 : \pi \leq \frac{1}{3}$ zu berechnen?

- A. `prop(~ (heads>=12), data=NULLvtlg)`
- B. `prop(~ (heads<=12), data=NULLvtlg)`
- C. `prop(~ (heads==12), data=NULLvtlg)`

Übung 51: Testentscheidung



```
# Anteil mindestens so viele Erfolge unter H_0 wie in der Stichprobe
prop( ~ (heads>=12), data=NULLvtlg)
```

```
## TRUE
## 0.471
```

Welche Aussage stimmt?

- A. Die Daten liefern Belege für die Forschungsthese, dass die Teilnehmer*innen einen Geschmacksunterschied schmecken, d. h., mehr Personen tippen richtig als zufällig zu erwarten wäre, d. h., H_0 wird verworfen.
- B. Die Daten liefern nicht ausreichend Belege für die Forschungsthese, dass die Teilnehmer*innen einen Geschmacksunterschied schmecken, d. h., nicht mehr Personen tippen richtig als zufällig zu erwarten wäre, d. h., H_0 wird nicht verworfen.
- C. Weiß nicht.

		Testentscheidung H_0	Testentscheidung H_A
Realität H_0	Ok	Fehler 1. Art ⁵²	
Realität H_A	Fehler 2. Art ⁵³	Ok	

⁵²Auch α -Fehler genannt. Die Wahrscheinlichkeit dieses Fehlers wird durch das Signifikanzniveau nach oben beschränkt.

⁵³Auch β -Fehler genannt. Die Wahrscheinlichkeit dieses Fehlers ist schwieriger zu bestimmen, aber siehe z. B. Paket [pwr](#). Bei guten Tests sinkt sie mit größerem Stichprobenumfang n .

Angenommen in Wirklichkeit liegt ein Geschmacksunterschied zwischen Krombacher und Perlenbacher vor. Welcher Fehler wurde begangen?

- A. Fehler 1. Art.
- B. Fehler 2. Art.
- C. Kein Fehler.

Einlesen der *Tipping*⁵⁴ Daten.

```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
tips <- read.csv2("tips.csv")

# Alternativ - heruntergeladene Datei einlesen:
# tips <- read.csv2(file.choose())
```

⁵⁴Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

- **Kategoriale** Variable: z. B. **Anteil p**

```
prop(smoker ~ time,
     success = "Yes",
     data=tips)

## Yes.Dinner  Yes.Lunch
## 0.3977273 0.3382353

diffprop(smoker ~ time,
         success = "Yes",
         data=tips)

## diffprop
## -0.05949198
```

- **Numerische** Variable, z. B.
Mittelwert \bar{x}

```
mean(total_bill ~ time,
      data=tips)
```

```
## Dinner      Lunch
## 20.79716 17.16868
```

```
diffmean(total_bill ~ time,
          data=tips)
```

```
## diffmean
## -3.628483
```

Original:

Name	Geschlecht	Größe
Ahmet	m	180
Gabi	w	170
Max	m	186
Susi	w	172

Permutation Geschlecht:

Name	Geschlecht	Größe
Ahmet	m	180
Gabi	m	170
Max	w	186
Susi	w	172

$$\bar{x}_m - \bar{x}_f = \frac{180 + 186}{2} - \frac{170 + 172}{2} = 12$$

$$\bar{x}_m - \bar{x}_f = \frac{180 + 170}{2} - \frac{186 + 172}{2} = -4$$

Wenn H_0 gilt, sollte es für den Wert der abhängigen Variable egal sein, welchen Wert die unabhängige hat: Permutiere Werte und berechne Teststatistik. Wiederhole dies oft und gucke ob der Wert der beobachtete Wert der Stichprobe unter den permutierten unwahrscheinlich ist.

Voraussetzung: Zufällige Stichprobe (Permutation) oder zufällige Zuordnung (Randomisation).

Beispiel: Zwei-Stichproben-Fall:

- ▶ Wiederhole z. B. 10000 ×
 - ▶ Mische die $n_1 + n_2$ Beobachtungen.
 - ▶ Ordne zufällig n_1 Beobachtungen der ersten Stichprobe zu, die restlichen der zweiten.
 - ▶ Berechne die Differenz der Mittelwerte $\bar{x}_1 - \bar{x}_2$. Analog für andere Teststatistiken, z. B. Anteil.
- ▶ Zeichne Histogramm der Verteilung der Teststatistik des Modells unter $H_0 : \mu_1 - \mu_2 = 0$. Vergleiche mit dem beobachteten Wert der Teststatistik (der Stichprobe).
- ▶ Der p-Wert ist der Anteil der zufälligen Teststatistiken, die mindestens so groß sind wie der beobachtete Wert.⁵⁵

⁵⁵Bei ungerichteten, zweiseitigen Tests im Absolutbetrag.

6. Einführung Inferenz **shuffle()**

```
# Stichprobe: Beobachtungen 1 bis 10
obs <- 1:10
obs

## [1] 1 2 3 4 5 6 7 8 9 10

# Zufallszahlengenerator setzen: Reproduzierbarkeit
set.seed(1896)

# Permutation:
shuffle(obs)

## [1] 10 7 5 8 9 4 3 1 2 6
```

Permutation der unabhängigen Variable

Unter H_0 : Kein Unterschied / Zusammenhang in der **Population**:

- Kategorial: Unterschied Raucheranteil *eigentlich* $\pi_{Lunch} - \pi_{Dinner} = 0$, beobachteter Unterschied $p_{Lunch} - p_{Dinner} = -0.06$ *zufällig* $\neq 0$.

```
diffprop(smoker ~ shuffle(time), success = "Yes", data=tips)
```

```
## diffprop
## 0.1036096
```

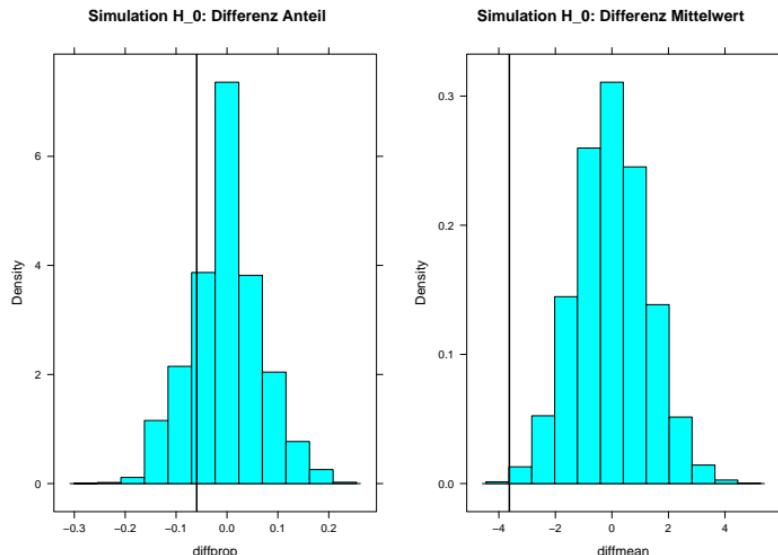
- Numerisch: Unterschied Mittelwerte Rechnungshöhe *eigentlich* $\mu_{Lunch} - \mu_{Dinner} = 0$, beobachteter Unterschied $\bar{x}_{Lunch} - \bar{x}_{Dinner} = -3.63$ *zufällig* $\neq 0$.

```
diffmean(total_bill ~ shuffle(time), data=tips)
```

```
## diffmean
## 0.6655749
```

6. Einführung Inferenz Permutationstest

```
set.seed(1896) # Reproduzierbarkeit
Nullvtlg_k <- do(10000) * diffprop(smoker ~ shuffle(time),
                                         success = "Yes", data=tips)
Nullvtlg_n <- do(10000) * diffmean(total_bill ~ shuffle(time),
                                         data=tips)
```



Übung 53: Testentscheidung

Unter welcher Nullhypothese ist der in der Stichprobe beobachtete Unterschied unwahrscheinlicher?

- A. Anteil Raucher: $H_0 : \pi_{Lunch} - \pi_{Dinner} = 0$
- B. Mittelwert Rechnungshöhe $H_0 : \mu_{Lunch} - \mu_{Dinner} = 0$
- C. Bei beiden ungefähr gleich.

- ▶ Bei einer **Punktschätzung** wird z. B. ein unbekannter Parameter / Wert der *Population* anhand eines Wertes der *Stichprobe* geschätzt, z. B. $\hat{\mu} = \bar{x}$, $\hat{\sigma} = sd$, $\hat{\pi} = p$. Das "Dach" ^ ist das Symbol für die Schätzung (engl.: (point) estimate).
- ▶ Der **Standardfehler** (engl.: standard error, *se*) beschreibt die Streuung (Standardabweichung) eines Schätzwertes, z. B. für den arithmetischen Mittelwert \bar{x} : $se = \frac{sd}{\sqrt{n}}$, d. h. *se* sinkt mit steigendem *n* (c. p.).
- ▶ Die **Anzahl Freiheitsgrade** (engl.: degrees of freedom, *df*) gibt an, wie viele Beobachtungen dabei *frei* sind: Ist der Mittelwert von *n* Beobachtungen bekannt, so ist $df = n - 1$.

Übung 54: Punktschätzung Dreieckstest

Im Dreieckstest Krombacher vs. Perlenbacher: Wie lautet der Punktschätzer für die Erfolgswahrscheinlichkeit, die richtige Probe zu finden?

- A. $\hat{\pi} = \frac{1}{3}$
- B. $\hat{\pi} = \frac{1}{2}$
- C. $\hat{\pi} = \frac{12}{34}$
- D. $\hat{\pi} = \frac{34}{34}$

Übung 55: Ergebnis Punktschätzung

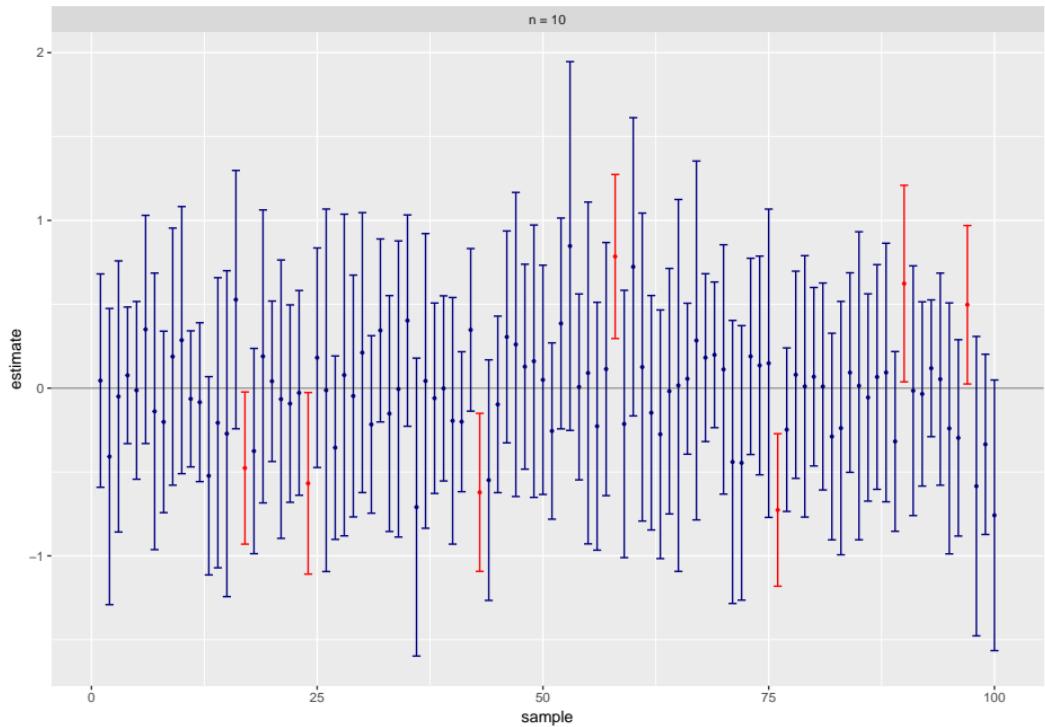
Wird mit Sicherheit in der Population gelten $\pi = \hat{\pi} = \frac{12}{34}$?

- ▶ Ja.
- ▶ Nein.

- ▶ Ein **Konfidenzintervall** gibt einem Bereich an, der den wahren, unbekannten Wert der Population mit einer gegebenen Sicherheit (z. B. $95\% = 1 - \alpha = 100\% - 5\%$) überdeckt, d. h., den Anteil der so konstruierten Konfidenzintervalle, die den Wert enthalten.
- ▶ Je größer die Sicherheit (z. B. 99 % statt 95 %), desto breiter ist das Intervall.⁵⁶
- ▶ Je größer der Stichprobenumfang, desto kleiner das Konfidenzintervall (unter sonst gleichen Umständen): der Standardfehler se fällt mit n .

⁵⁶Häufig bei $n > 30$: $95\%-KI \approx \delta^* \pm (2 \cdot se)$

Überdeckung durch Konfidenzintervall



Original:

Name	Geschlecht	Größe
Ahmet	m	180
Gabi	w	170
Max	m	186
Susi	w	172

Bootstrap-Stichprobe:

Name	Geschlecht	Größe
Ahmet	m	180
Ahmet	m	180
Max	m	186
Susi	w	172

$$\bar{x}_m - \bar{x}_f = \frac{180 + 186}{2} - \frac{170 + 172}{2} = 12$$

$$\bar{x}_m - \bar{x}_f = \frac{180 + 180 + 186}{3} - \frac{172}{1} = 10$$

Ziehe aus der Originalstichprobe mit Zurücklegen eine neue vom gleichen Umfang und berechne Stichprobenstatistik. Wiederhole dies oft und schätze damit die Verteilung der Stichprobenstatistik.

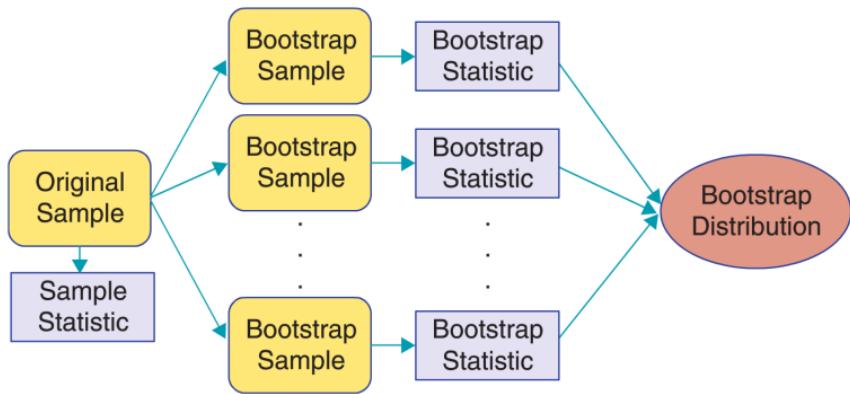


Abbildung: Quelle: Lock, Robin, Patti Frazer Lock, Kari Lock Morgan, Eric F. Lock, and Dennis F. Lock (2012): Statistics: UnLOCKing the Power of Data. Wiley.

Simuliere zufällige Stichprobenziehung: **Ziehe mit Zurücklegen** aus der gegebenen Stichprobe eine neue vom gleichen Umfang:

```
# Stichprobe: Beobachtungen 1 bis 10
obs <- 1:10
obs

## [1] 1 2 3 4 5 6 7 8 9 10

# Zufallszahlengenerator setzen: Reproduzierbarkeit
set.seed(1896)

# Resampling:
resample(obs)

## [1] 10 7 6 6 8 8 6 1 6 2
```

Voraussetzung: Zufällige Stichprobe oder zufällige Zuordnung. Nicht zu kleine Stichprobe.⁵⁷

Beispiel: Bootstrap-Perzentil-Intervall⁵⁸ für eine Stichprobe:

- ▶ Wiederhole z. B. 10000 ×
 - ▶ Ziehe mit Zurücklegen eine Stichprobe vom Umfang n aus der Originalstichprobe.
 - ▶ Berechne Statistik, z. B. Mittelwert \bar{x} der Bootstrap-Stichprobe. Analog für andere Statistiken, z. B. Anteil.
- ▶ Zeichne Histogramm der Bootstrap-Verteilung der Statistik.
- ▶ Das 95 %-Bootstrap-Perzentil-Intervall sind die mittleren 95 % der Bootstrap-Verteilung.

⁵⁷ $n \geq 35$

⁵⁸ Es gibt weitere, teilweise exaktere Bootstrap-Methoden.

```
# Datenvektor
testerg <- factor(c(rep("r",12), # 3 richtig
                     rep("f",22))) # 11 falsch

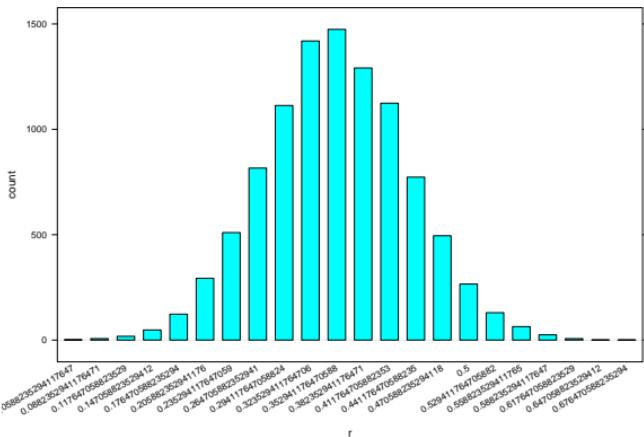
prop( ~ testerg, success = "r")

##          r
## 0.3529412

set.seed(1896) # Reproduzierbarkeit
Bootvtlg <- do(10000) * prop( ~ resample(testerg),
                           success = "r")
```

Verteilung Bootstrap-Statistik

```
bargraph( ~ r, data = Bootvtlg)
```



```
quantile(~ r, data = Bootvtlg, probs=c(0.025, 0.975))
```

```
##      2.5%      97.5%
## 0.2058824 0.5000000
```

Übung 56: Konfidenzintervall

Auf Basis der Daten und der Bootstrap-Simulation. Welche Aussage stimmt?

- A. In 95% der simulierten Stichproben liegt der Anteil der richtig erkannten Proben zwischen 0.21 und 0.5.
- B. Ein Anteil von 0.33 ist bei den Daten ein ungewöhnlicher Wert.
- C. Mehr Bootstrap-Stichproben (d. h., z. B. `do(1000000)`) verkleinert das Konfidenzintervall.

Hinweis: Aufgrund der kleinen Stichprobe ist hier ein Vorgehen über die zugrundeliegende, theoretische Binomialverteilung exakter. Bei kleinen Stichproben ist das Bootstrap-Konfidenzintervall i. d. R. zu klein und die Verteilung der Stichprobe entspricht evtl. nicht der Verteilung der Population.

Überprüfung ob das Bootstrap-Intervall ggf. verzerrt ist: $\overline{\hat{\delta}^*} - \hat{\delta}$

```
mean( ~ r, data = Bootvtlg) - 12/34
```

```
## [1] -0.0006441176
```

Bootstrap-Schätzung des Standardfehlers:

```
sd( ~ r, data = Bootvtlg)
```

```
## [1] 0.0805043
```

Hinweis: Theoretisch beträgt $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\frac{12}{34}(1-\frac{12}{34})}{34}} = 0.0820.$

- ▶ **Permutationstest**, hier: simuliere zufällige Zuordnung⁵⁹. Simulierte Verteilung einer Statistik unter der Annahme, dass kein Zusammenhang vorliegt (Modell H_0), u. a. zur Bestimmung von p-Werten.

```
statistik(y ~ shuffle(x), data = Daten)
```

- ▶ **Bootstrap**, hier: simuliere zufälliges Ziehen einer Stichprobe⁶⁰. Schätze Verteilung einer Statistik der Stichprobe, u. a. zur Bestimmung von Konfidenzintervallen oder Standardfehlern.

```
statistik(y ~ x, data = resample(Daten))
```

⁵⁹d. h. ohne Zurücklegen

⁶⁰d. h. mit Zurücklegen

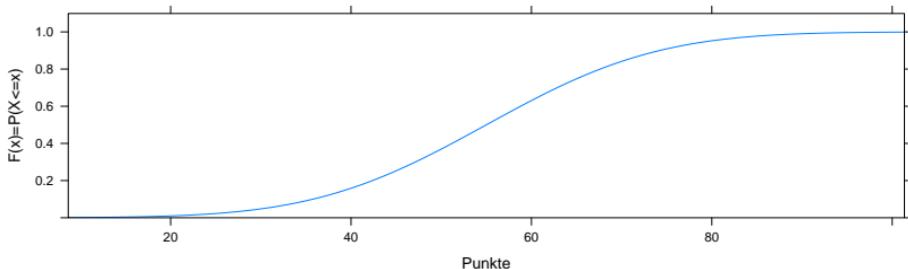
7 Normalverteilung

- ▶ Eine **Zufallsvariable** X ist eine Variable, deren Wert x vom **Zufall** abhängt.
- ▶ Beobachtungen x_i können aufgefasst werden als Realisationen von Zufallsvariablen X .
- ▶ Die **Verteilungsfunktion** $F(x)$ sagt, wie wahrscheinlich es ist, einen Wert $\leq x$ zu beobachten: $F(x) = P(X \leq x)$, und damit $0 \leq F(x) \leq 1$.
- ▶ Durch eine *zufällige* Stichprobe oder eine *zufällige* Zuordnung im Rahmen eines Experimentes soll sichergestellt werden, dass die Beobachtungen x unabhängig und identisch verteilt sind.

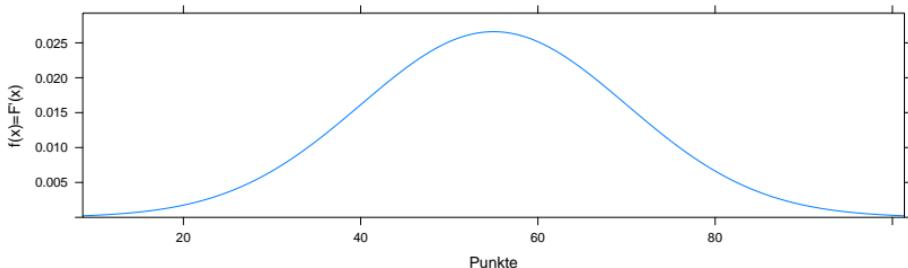
Normalverteilung

Die Punkte einer Klausur seien **normalverteilt** mit $\mu = 55$ (Mittelwert der Population / Erwartungswert) und $\sigma = 15$ (Standardabweichung Population).

Verteilungsfunktion



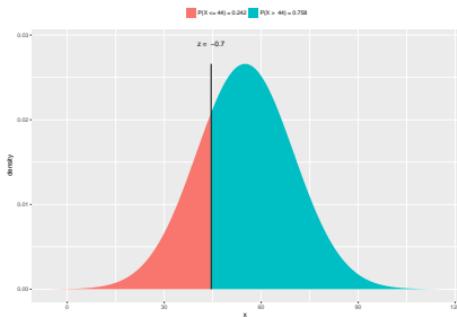
Dichtefunktion



Angenommen, unter 45 Punkten, d. h., mit 44.5 Punkten oder weniger, fällt mensch leider durch, dann liegt die Durchfallwahrscheinlichkeit bei 0.24:

$$p = F(x) = P(X \leq x)$$

```
xpnorm(44.5, mean = 55, sd = 15)
```



```
## [1] 0.2419637
```

Stimmt die Aussage: Die Verteilungsfunktion ist monoton nicht fallend, d. h., mit größerem x wird $F(x)$ zumindest nicht kleiner?

- ▶ Ja.
- ▶ Nein.

Übung 58: Eigenschaften Verteilungsfunktion II

Was gilt für $x \rightarrow \infty$?

- A. $F(x)$ geht gegen 0.
- B. $F(x)$ geht gegen 1.
- C. $F(x)$ geht gegen Unendlich.
- D. Kann nicht beantwortet werden.

z-Transformation, Standardisierung: Überführung einer beliebigen Verteilung in eine mit $\mu = 0$ und $\sigma = 1$:

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

```
zscore(c(0,1,2))
```

```
## [1] -1 0 1
```

```
zscore(c(100,200,300))
```

```
## [1] -1 0 1
```

Übung 58: z-Wert

Welche der Interpretationen von $z = -2$ ist *falsch*?

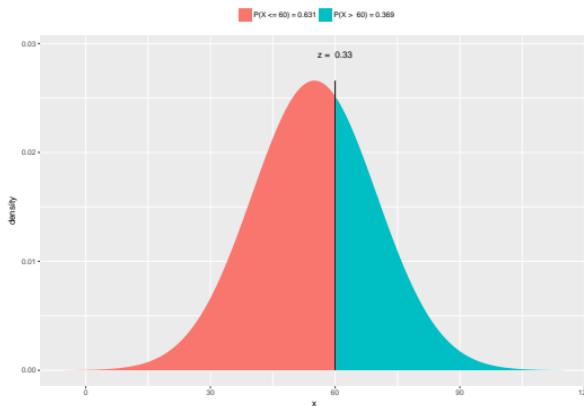
- A. Der Wert ist unterdurchschnittlich.
- B. Der Wert liegt 2 Standardabweichungen vom Mittelwert entfernt.
- C. x^{61} ist kleiner als 0.

⁶¹ $x = z \cdot \sigma + \mu$

Übung 59: Verteilungsfunktion (I/II)

Ein Studierender hat 60 Punkte erreicht:

```
xpnorm(60, mean = 55, sd = 15)
```



```
## [1] 0.6305587
```

Übung 59: Verteilungsfunktion (II/II)

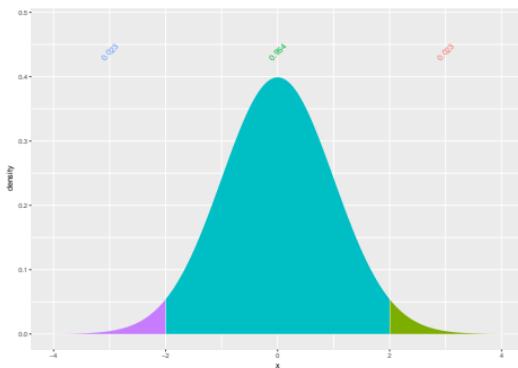
Welche Aussage stimmt?

- A. ca. 37% der Population schneiden schlechter ab als er.
- B. ca. 63% der Population schneiden besser ab als er.
- C. Er ist $\frac{1}{3}$ Standardabweichungen über dem Durchschnitt.
- D. Keine der Antworten A–C stimmt.

Bei einer Normalverteilung liegen ca.

- ▶ 68% der Werte im Bereich $\bar{x} \pm 1 \cdot sd$
- ▶ 95% der Werte im Bereich $\bar{x} \pm 2 \cdot sd$
- ▶ 99,7% der Werte im Bereich $\bar{x} \pm 3 \cdot sd$.

```
xpnorm(c(-2,2)) # Standardnormalverteilung
```



```
## [1] 0.02275013 0.97724987
```

Übung 60: 68-95-99,7% Regel

Die Daten seien normalverteilt mit $\mu = 100$ und $\sigma = 15$. Ist dann $x = 150$ ein üblicher Wert?

- ▶ Ja.
- ▶ Nein.

Offene Übung 61: Standardisierung

Welchen z -Wert⁶² hat ein Kandidat mit $x = 130$, wenn $\mu = 100$ und $\sigma = 15$ ist?

⁶²standardisiert, z-transformiert

Übung 62: Verteilungsfunktion in R

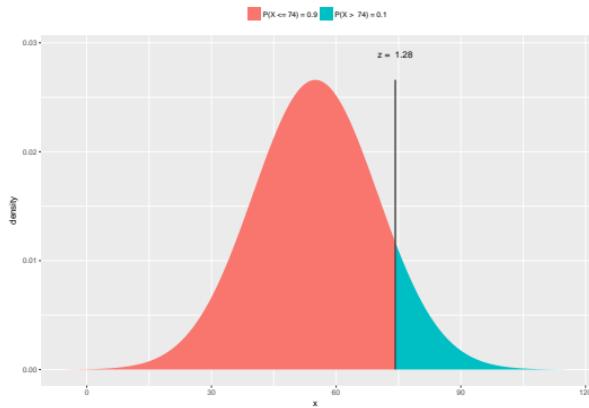
Welcher Befehl liefert den Wert der Verteilungsfunktion an der Stelle 115, wenn $\mu = 100$ und $\sigma = 15$ ist?

- A. `xpnorm(115, 100, 15)`
- B. `xpnorm(100, 15, 115)`
- C. `xpnorm(15, 115, 100)`

Man gehört zu den 10% besten ab 74 Punkten:

$$q = F^{-1}(p)$$

```
xqnorn(0.9, mean = 55, sd = 15)
```



```
## [1] 74.22327
```

Übung 63: Quantilsfunktion

Welche Aussage stimmt? (p : Wahrscheinlichkeit, dass q nicht überschritten wird)

- A. Je kleiner p , desto kleiner q .
- B. Je kleiner p , desto größer q .
- C. p und q stehen in keinem Zusammenhang.

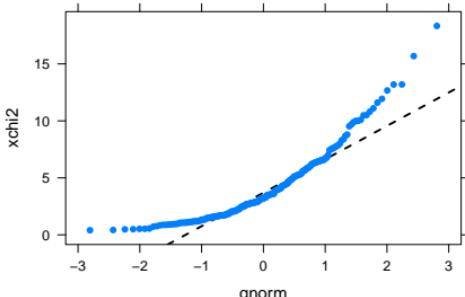
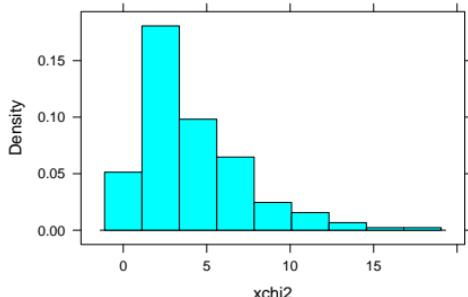
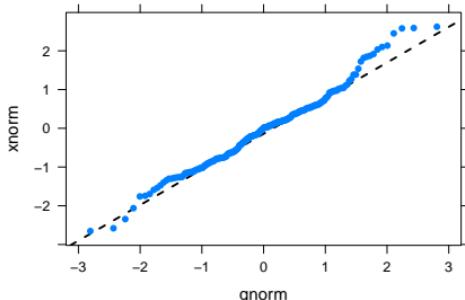
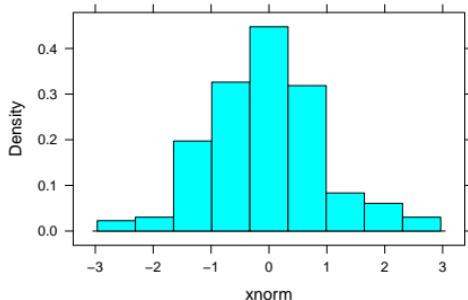
Übung 64: Quantil berechnen

Mit bis zu wie vielen Punkten zählt man zu dem oberen Drittel der Klausur?

- A. `xqnorm(1/3, mean = 55, sd = 15)`: 49
- B. `xqnorm(2/3, mean = 55, sd = 15)`: 61

Überprüfung Normalverteilungsannahme

Ein Q-Q Plot (`xqqmath()`) vergleicht die Quantile einer Verteilung z. B. mit den theoretischen einer Normalverteilung. Bei einer guten Übereinstimmung liegen die Punkte auf der Diagonalen.



Ein Fertigungsprozess funktioniere zum sog. 3σ Niveau, d. h., die erlaubten Abweichungen (Toleranz) sind innerhalb von 3 Standardabweichungen um den Mittelwert:

```
x3 <- pnorm(-3)
```

Dann wird bei $1.000.000 = 10^6$ Produkten ein Ausschuss von

```
(2*x3) * 1e06
```

```
## [1] 2699.796
```

erwartet.⁶³

⁶³Innerhalb von Six Sigma wird zur Berechnung des erwarteten Fehleranteils ("DPMO") zusätzlich eine langfristige Mittelwertsverschiebung um 1.5σ berücksichtigt, die hier ignoriert wird.

Übung 65: Six Sigma

Wie viele Fehler pro Million werden erwartet, wenn der akzeptierte Bereich innerhalb von 6σ liegt?⁶⁴

- A. ≈ 2700
- B. ≈ 65
- C. ≈ 1
- D. < 0.1

⁶⁴Ohne Berücksichtigung der Mittelwertsverschiebung.

Gabi und Klaus haben zwei verschiedene Tests geschrieben. Dabei hat Gabi bei Test A ($\mu = 60, \sigma = 10$) 75 Punkte erreicht, Klaus bei Test B ($\mu = 250, \sigma = 25$) 275 Punkte. Beide Tests sollen das Gleiche testen und die Testergebnisse seien normalverteilt. Wer von beiden hat besser abgeschnitten?

- A. Gabi
- B. Klaus
- C. Beide gleich gut.

8 Inferenz kategorialer Daten

Idee: Schluss von einer (zufälligen / randomisierten) Stichprobe auf eine Population:

- ▶ Punktschätzung
- ▶ Konfidenzintervall
- ▶ Hypothesentest

Ziel: Aussagen treffen, die über die Stichprobe hinausgehen – und dabei berücksichtigen, dass Variation allgegenwärtig ist und Schlussfolgerungen unsicher.⁶⁵

⁶⁵Vgl. Moore, D. (2007) *The Basic Practice of Statistics*, 4th edn. New York: Freeman, S. xxviii.

Übung 67: Statistik Essenszeit

Durch welche Statistik kann die Verteilung der Variable Essenzeit (Lunch / Dinner) sinnvoll beschrieben werden?

- A. Anteil.
- B. Arithmetischer Mittelwert.

Übung 68: Visualisierung Essenszeit

Durch welche Grafik kann die Verteilung der Variable Essenzeit (Lunch / Dinner) sinnvoll dargestellt werden?

- A. Balkendiagramm.
- B. Histogramm.
- C. Boxplot.

Übung 69: Gültigkeit Inferenz

Wann ist aufgrund einer quantitativen Datenanalyse eine Aussage über die Population gerechtfertigt?

- A. Nie.
- B. Bei einer zufälligen Stichprobe.
- C. Bei einer randomisierten Zuordnung innerhalb eines Experimentes.
- D. Bei einem hohen Stichprobenumfang n .
- E. Immer.

- ▶ **Test eines Anteilswertes:** Anteil eines (binären⁶⁶) Merkmals in der Population.

`prop.test()`. Mögliche Hypothesen:

- ▶ ungerichtet / zweiseitig: $H_0 : \pi = \pi_0$ vs. $H_A : \pi \neq \pi_0$. Option:
`alternative="two.sided"` (Standardeinstellung)
- ▶ gerichtet / einseitig:
 - ▶ $H_0 : \pi \leq \pi_0$ vs. $H_A : \pi > \pi_0$. Option: `alternative="greater"`
 - ▶ $H_0 : \pi \geq \pi_0$ vs. $H_A : \pi < \pi_0$. Option: `alternative="less"`

- ▶ **Test zweier Anteilswerte:** Vergleicht zwei Anteilswerte. `prop.test()`. Mögliche Hypothesen:

- ▶ ungerichtet / zweiseitig: $H_0 : \pi_1 = \pi_2$ vs. $H_A : \pi_1 \neq \pi_2$. Option:
`alternative="two.sided"` (Standardeinstellung)
- ▶ gerichtet / einseitig:
 - ▶ $H_0 : \pi_1 \leq \pi_2$ vs. $H_A : \pi_1 > \pi_2$. Option: `alternative="greater"`
 - ▶ $H_0 : \pi_1 \geq \pi_2$ vs. $H_A : \pi_1 < \pi_2$. Option: `alternative="less"`

- ▶ **Chi-Quadrat Unabhängigkeitstest:** Zusammenhang / Unabhängigkeit zweier nominaler Merkmale. Hypothese H_0 : Die Merkmale sind unabhängig, es gibt keinen Zusammenhang. H_A : Die Merkmale sind nicht unabhängig, es gibt einen Zusammenhang. `xchisq.test()`

⁶⁶kategorial mit zwei Ausprägungen

Beispiele

- ▶ Analyse des Anteils der Studierenden, die die Vorlesung nachbereiten – ggf. je nach Geschlecht oder Studiengang.
- ▶ Untersuchung des Anteils der Mitarbeiter*innen, die während der Arbeit SocialMedia nutzen – ggf. je nach Geschlecht.
- ▶ Analyse des Anteils der betrügerischen Versicherungsvorgänge – ggf. je nach Vertragsart.
- ▶ Vergleich des Anteils der Dividendenzahlenden Unternehmen je Index.
- ▶ Anteil von “Blockbuster-Movies” pro Film-Genre (s. Datensatz [ggplot2movies](#)).

Wo können Sie die Verfahren einsetzen?

Frauenanteil der Rechnungszahler*innen: Einlesen der Daten

Einlesen der *Tipping*⁶⁷ Daten:

```
# Herunterladen  
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")  
  
# Einlesen in R  
tips <- read.csv2("tips.csv")  
  
# Alternativ - heruntergeladene Datei einlesen:  
# tips <- read.csv2(file.choose())  
  
library(mosaic) # Paket mosaic laden
```

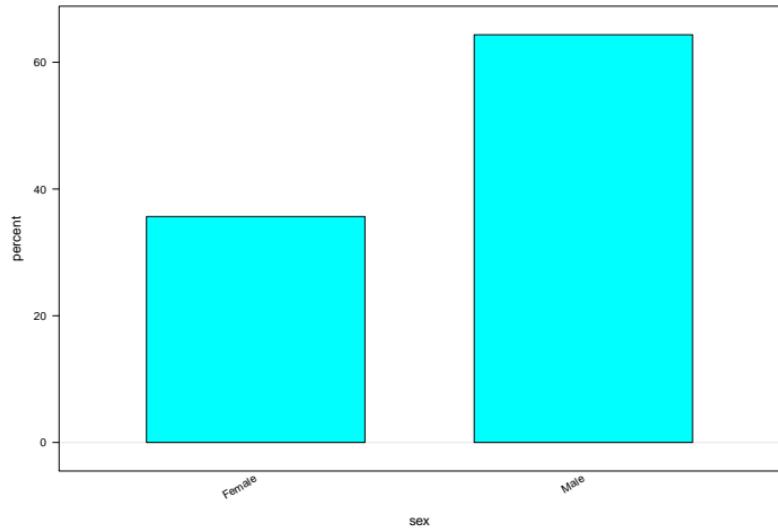
⁶⁷Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

```
tally(~ sex, # (unabhängige) Variable  
      data=tips) # Datensatz
```

```
## sex  
## Female    Male  
##     87     157
```

Frauenanteil der Rechnungszahler*innen: Deskriptive Analyse (II / II)

```
bargraph( ~ sex, data = tips, type = "percent")
```



Übung 70: Frauenanteil Stichprobe

Stimmt die Aussage: In der Stichprobe liegt der Frauenanteil nicht bei 50 %?

- ▶ Ja.
- ▶ Nein.

Übung 71: Testverfahren Frauenanteil der Rechnungszahler*innen

Welches ist das richtige Testverfahren, um die Forschungsthese zu untersuchen, dass der Anteil der Rechnungszahlerinnen , d. h. `sex=="Female"`, nicht bei 50 % liegt – in der Population?

- A. Test eines Anteilswertes.
- B. Chi-Quadrat Unabhängigkeitstest.

Übung 72: Hypothesen Frauenanteil der Rechnungszahler*innen

Wie lautet das korrekte Hypothesenpaar?

- A. $H_0 : \pi = 0.5$ vs. $H_A : \pi \neq 0.5$
- B. $H_0 : \pi \leq 0.5$ vs. $H_A : \pi > 0.5$
- C. $H_0 : \pi \geq 0.5$ vs. $H_A : \pi < 0.5$

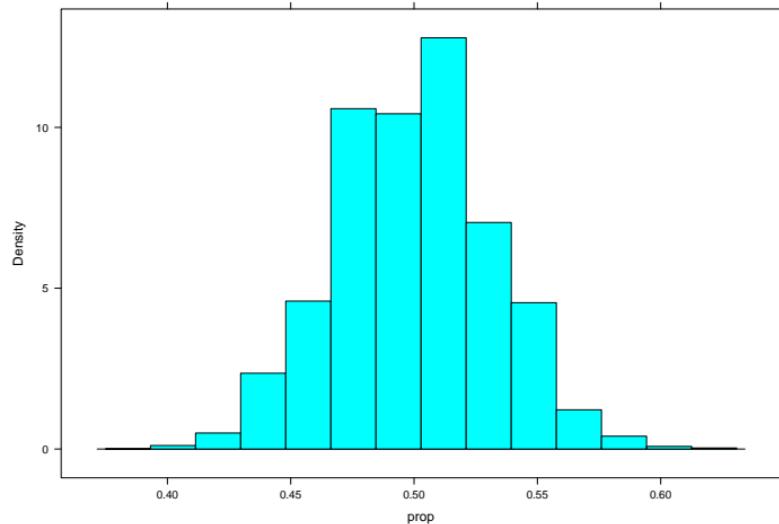
Simulation unter H_0

```
# Zufallszahlengenerator setzen
set.seed(1896)

Nullvtlg <- do(10000) * # 10000 Wiederholungen
rflip(n=nrow(tips)) # n-facher Münzwurf
```

Verteilung des Anteils unter H_0

```
histogram(~ prop, data = Nullvtlg)
```



Vergleich beobachteter Anteil in Verteilung unter H_0

Punktschätzer

```
propdach <- prop(~ sex, data = tips,  
                  success = "Female")
```

```
propdach
```

```
##      Female
```

```
## 0.3565574
```

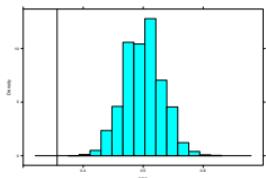
2.5% und 97.5% Quantile der Verteilung unter H_0

```
quantile(~ prop, data = Nullvtlg,  
         probs=c(0.025, 0.975))
```

```
##      2.5%      97.5%
```

```
## 0.4385246 0.5614754
```

Übung 73: Interpretation Simulationsergebnis



Welche der folgenden Aussagen stimmt?

- A. Ein Frauenanteil von 0.36 in der Stichprobe ist unter der Annahme, der Anteil in der Population liegt bei 0.5, ein üblicher Wert.
- B. Ein Frauenanteil von 0.36 in der Stichprobe ist unter der Annahme, der Anteil in der Population liegt bei 0.5, kein üblicher Wert.

```
# Abweichung zu p_0=0.5
abw0 <- abs(propdach - 0.5)

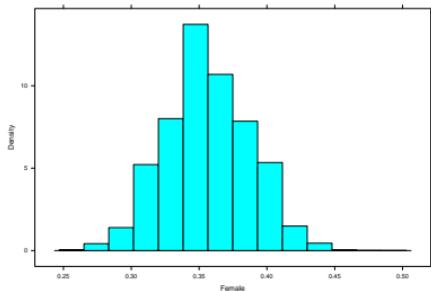
pvalue <- prop( ~ abs(prop-0.5) >= abw0,
                 data = Nullvtlg)

pvalue
## TRUE
##    0
```

Bootstrap Konfidenzintervall

```
set.seed(1896) # Reproduzierbarkeit
Bootvtlg <- do(10000) *
  prop(~ sex, data = resample(tips), success = "Female")

histogram(~ Female, data = Bootvtlg)
```



```
quantile(~ Female, data = Bootvtlg, probs = c(0.025, 0.975))
```

```
##      2.5%    97.5%
## 0.2991803 0.4180328
```

Test des Anteilswertes

```
prop.test(~ sex, # Variable, die gestestet wird  
          p = 0.5, # hypothetischer Wert p_0  
          alternative = "two.sided", # Alternativhypothese  
          data = tips) # Datensatz
```

Ergebnis Test des Anteilswertes

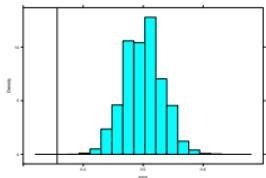
```
##  
## 1-sample proportions test with continuity correction  
##  
## data: tips$sex [with success = Female]  
## X-squared = 19.512, df = 1, p-value = 9.995e-06  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.2971654 0.4205463  
## sample estimates:  
## p  
## 0.3565574
```

Übung 74: Testergebnis: Stichprobenanteil

Wie hoch ist der Anteil der Frauen unter den Rechnungszahler*innen in der Stichprobe?

- A. 19.51%
- B. 29.72%
- C. 42.05%
- D. 35.66%
- E. 50.00%

Übung 75: Testergebnis: Testentscheidung



Liefern die Daten der Stichprobe Belege dafür, dass der Frauenanteil in der Population der Rechnungszahler*innen nicht bei 50 % liegt?

- ▶ Ja.
- ▶ Nein.

Vergleich Geschlecht je Tageszeit

```
prop(sex ~ time, success = "Female", data = tips)

## Female.Dinner  Female.Lunch
##      0.2954545   0.5147059

diffdach <- diffprop(sex ~ time, success = "Female", data = tips)
diffdach

##  diffprop
## 0.2192513
```

In der Stichprobe:

$$\hat{\pi}_{\text{Lunch}} - \hat{\pi}_{\text{Dinner}} = 0.51 - 0.3 = 0.22$$

Permutationstest Geschlecht je Tageszeit

```
set.seed(1896) # Reproduzierbarkeit

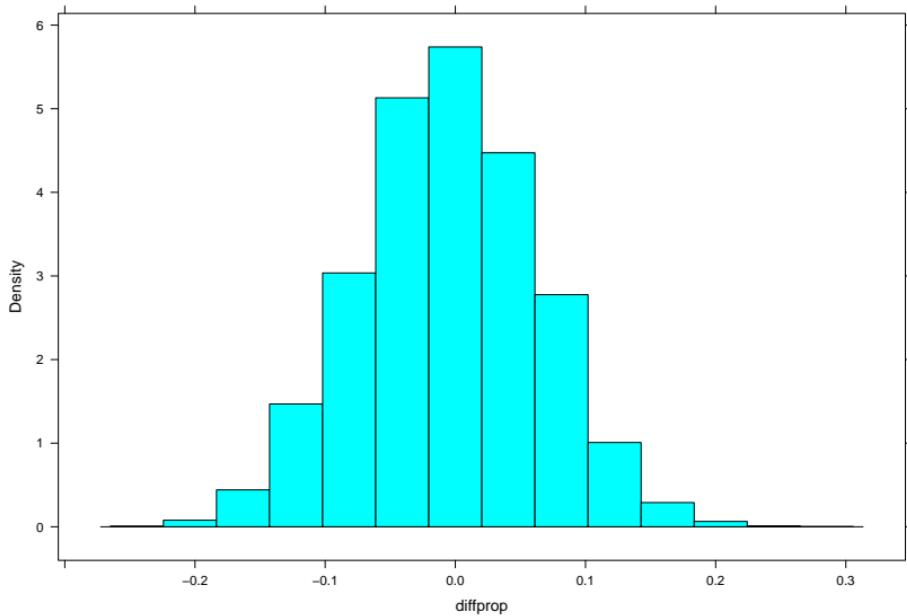
Nullvtlg <- do(10000) * diffprop(sex ~ shuffle(time),
                                    success = "Female", data = tips)

quantile(~ diffprop, data = Nullvtlg, probs = c(0.025, 0.975))

##          2.5%      97.5%
## -0.1273396  0.1377005
```

Verteilung unter H_0

```
histogram(~ diffprop, data = Nullvtlg)
```



76: Bestimmung p-Wert

Für welches Hypothesenpaar erhalten Sie den p-Wert über:

```
prop( ~ abs(diffprop) >= abs(diffdach), data = Nullvtlg)
```

```
##    TRUE
```

```
## 0.0014
```

- A. Für $H_0: \hat{\pi}_{\text{Lunch}} - \hat{\pi}_{\text{Dinner}} = 0$.
- B. Für $H_0: \pi_{\text{Lunch}} - \pi_{\text{Dinner}} = 0$.

Offene Übung 77: Geschlecht je Tageszeit

Fassen Sie die vorangegangene Analyse zusammen. Wie lautete die Forschungsfrage, Hypothesen und die Antwort auf die Forschungsfrage.

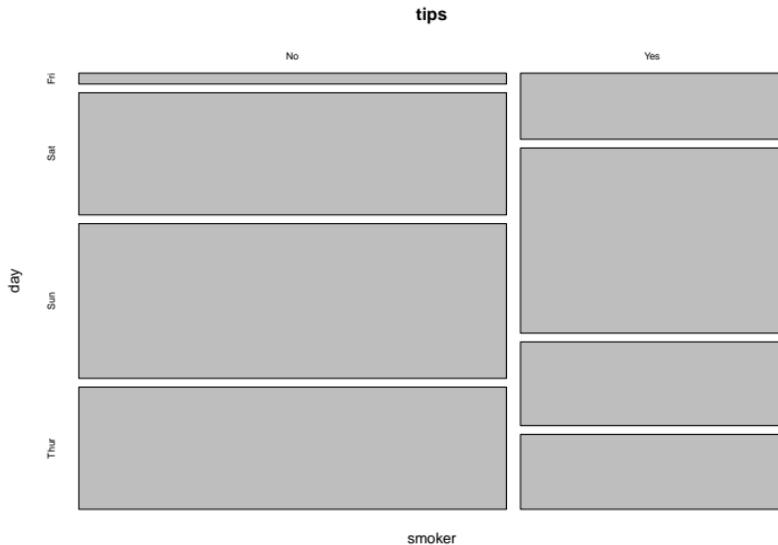
Anzahl der Raucher*innen je Wochentag

```
tally(smoker ~ # abhängige Variable  
      day, # unabhängige Variable  
      data = tips) # Datensatz
```

```
##          day  
## smoker Fri Sat Sun Thur  
##   No     4   45   57   45  
##   Yes    15  42   19   17
```

Verteilung Raucher*innen je Wochentag

```
mosaicplot(smoker ~ day, data = tips)
```



- ▶ Test des Zusammenhangs zweier kategorialer (nominaler) Variablen. Dabei werden die **beobachteten** Häufigkeiten O (*observed*) der Merkmalsausprägungskombinationen mit den **unter Unabhängigkeit erwarteten** Werten E (*expected*) verglichen:

$$\chi^2 = \sum_i^{\text{Zeilen}} \sum_j^{\text{Spalten}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- ▶ Nullhypothese: Die beiden nominalen Variablen sind unabhängig voneinander, d. h., die Verteilung der einen Variable hängt nicht vom Wert der anderen Variable ab. Große Werte von χ^2 sind unter H_0 unwahrscheinlich.

Übung 78: χ^2 -Teststatistik

Eine Forscherin stellt innerhalb einer Untersuchung eine Abweichung zwischen beobachtet O und erwartet E von 42 fest.

Welche Aussage stimmt?

- A. Die Abweichung ist groß.
- B. Die Abweichung ist klein.
- C. Weiß nicht.

Welches ist das richtige Testverfahren, um die Forschungsthese zu untersuchen, dass es einen Zusammenhang zwischen Rauchen und Wochentag gibt?

- A. Test eines Anteilswertes.
- B. Chi-Quadrat Unabhängigkeitstest.

Chi-Quadrat Test

```
xchisq.test(smoker ~ day, data = tips)

##
## Pearson's Chi-squared test
##
## data: tally(x, data = data)
## X-squared = 25.787, df = 3, p-value = 1.057e-05
##
##      4      45      57      45
## (11.76) (53.84) (47.03) (38.37)
## [5.12]   [1.45]   [2.11]   [1.15]
## <-2.26> <-1.20> < 1.45> < 1.07>
##
##      15      42      19      17
## ( 7.24) (33.16) (28.97) (23.63)
## [8.31]   [2.36]   [3.43]   [1.86]
## < 2.88> < 1.54> <-1.85> <-1.36>
##
## key:
## observed
## (expected)
## [contribution to X-squared]
## <Pearson residual>
```

Übung 80: Testergebnis: Testentscheidung

Liefern die Daten genug Belege für einen Zusammenhang zwischen Rauchen und Wochentag – in der Population?

- ▶ Ja.
- ▶ Nein.

Offene Übung 81: Mahlzeit und Rauchen

Untersuchen Sie den Zusammenhang zwischen der Mahlzeit (Tageszeit) und Rauchen am Tisch.

9 Inferenz numerischer Daten

Inferenz

Idee: Schluss von einer (zufälligen / randomisierten) Stichprobe auf eine Population:

- ▶ Punktschätzung
- ▶ Konfidenzintervall
- ▶ Hypothesentest

Ziel: Aussagen treffen, die über die Stichprobe hinausgehen – und dabei berücksichtigen, dass Variation allgegenwärtig ist und Schlussfolgerungen unsicher.⁶⁸

⁶⁸Vgl. Moore, D. (2007) *The Basic Practice of Statistics*, 4th edn. New York: Freeman, S. xxviii.

Durch welche Statistik kann die zentrale Tendenz der Variable Rechnungshöhe sinnvoll beschrieben werden?

- A. Anteil.
- B. Arithmetischer Mittelwert.

Übung 83: Visualisierung Rechnungshöhe

Durch welche Grafik kann die Verteilung der Variable Rechnungshöhe *nicht* sinnvoll dargestellt werden?

- A. Balkendiagramm.
- B. Histogramm.
- C. Boxplot.

Übung 84: Gültigkeit Inferenz

Wann ist aufgrund einer quantitativen Datenanalyse eine Kausalaussage gerechtfertigt?

- A. Nie.
- B. Bei einer zufälligen Stichprobe.
- C. Bei einer randomisierten Zuordnung innerhalb eines Experimentes.
- D. Bei einem hohen Stichprobenumfang n .
- E. Immer.

- ▶ **Einstichproben t-Test:** Testet den Mittelwert eines Merkmals einer Stichprobe mit einem hypothetisch richtigen Mittelwert der Population. `t.test()`
 - ▶ ungerichtet, zweiseitig: $H_0 : \mu = \mu_0$, vs. $H_A : \mu \neq \mu_0$ `alternative='two.sided'`
 - ▶ gerichtet, einseitig:
 - ▶ $H_0 : \mu \leq \mu_0$, vs. $H_A : \mu > \mu_0$ `alternative='greater'`
 - ▶ $H_0 : \mu \geq \mu_0$, vs. $H_A : \mu < \mu_0$ `alternative='less'`
- ▶ **Gepaarter Test / t-Test für abhängige Stichproben:** Testet die Differenz der Mittelwerte zweier Merkmale (A, B) einer Stichprobe mit einer hypothetisch richtigen Differenz in der Population⁶⁹. `t.test()`
 - ▶ ungerichtet, zweiseitig: $H_0 : \mu_{A-B} = \delta_0$, vs. $H_A : \mu_{A-B} \neq \delta_0$ `alternative='two.sided'`
 - ▶ gerichtet, einseitig:
 - ▶ $H_0 : \mu_{A-B} \leq \delta_0$, vs. $H_A : \mu_{A-B} > \delta_0$ `alternative='greater'`
 - ▶ $H_0 : \mu_{A-B} \geq \delta_0$, vs. $H_A : \mu_{A-B} < \delta_0$ `alternative='less'`

⁶⁹häufig: $\delta_0 = 0$

- ▶ **Zweistichproben Test / t-Test für unabhängige Stichproben:** Testet die Mittelwerte eines Merkmals zweier Stichproben 1, 2 in der Population⁷⁰. `t.test()`
 - ▶ ungerichtet, zweiseitig: $H_0 : \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$, vs.
 $H_A : \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0$ `alternative='two.sided'`
 - ▶ gerichtet, einseitig:
 - ▶ $H_0 : \mu_1 \leq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \leq 0$, vs. $H_A : \mu_1 > \mu_2 \Leftrightarrow \mu_1 - \mu_2 > 0$
`alternative='greater'`
 - ▶ $H_0 : \mu_1 \geq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \geq 0$, vs. $H_A : \mu_1 < \mu_2 \Leftrightarrow \mu_1 - \mu_2 < 0$ `alternative='less'`
- ▶ **Varianzanalyse / Anova:** Testet die Gleicheit der Mittelwerte zweier oder mehr Stichproben (Merkmale) in der Population: $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ vs. H_A : mindestens ein Mittelwert unterscheidet sich ($\mu_i \neq \mu_j$). `aov()`

⁷⁰ auch $\delta_0 \neq 0$ möglich

Beispiele

- ▶ Analyse des Workloads der Studierenden – ggf. je nach Geschlecht oder Studiengang
- ▶ Untersuchung des Humors⁷¹ der Mitarbeiter*innen, ggf. je Geschlecht oder Abteilung
- ▶ Vergleich der Kaufkraft der Kund*innen mit oder ohne Kundenkarte
- ▶ Analyse der Rendite von Investitionsalternativen
- ▶ Vergleich der Mitarbeiter-Zufriedenheit zwischen Abteilungen

Wo können Sie die Verfahren einsetzen?

⁷¹latente Variable, daher Operationalisierung erforderlich

- ▶ Einstichproben t-Test: eine Stichprobe, ein Merkmal: $t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{sd^2}{n}}} = \frac{\bar{x} - \mu_0}{se}$.
- ▶ t-Test für abhängige Stichproben, gepaarter t-Test: eine Stichprobe, zwei Merkmale, es wird die Differenz je Beobachtung analysiert.
- ▶ t-Test für unabhängige Stichproben: zwei Stichproben, ein Merkmal.
- ▶ Idee⁷²: Setze Differenz der Mittelwerte ins Verhältnis zur Streuung der Schätzung (Standardfehler, se):

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$$

Große Werte von $|t|^{73}$ sind unter der Nullhypothese unwahrscheinlich.

- ▶ Voraussetzung: Daten innerhalb der Stichprobe(n) unabhängig, identisch, normalverteilt

⁷²hier im Fall für zwei unabhängige Stichproben, analog für die anderen Fälle

⁷³im zweiseitigen Fall

Übung 85: t-Test

Bei einem gerichteten Einstichproben t-Test für

$$H_0 : \mu \leq 42 \quad vs. \quad H_A : \mu > 42$$

kommt als Schätzwert der Stichprobe $\hat{\mu} = \bar{x} = 40$ raus.

Wird der t-Test die Nullhypothese verwerfen?

- A. Ja.
- B. Nein.
- C. Vielleicht. Hängt von $se = \frac{sd}{\sqrt{n}}$ ab.

Einlesen der *Tipping*⁷⁴ Daten sowie Laden des Pakets `mosaic`. Zufallszahlengenerator setzen.

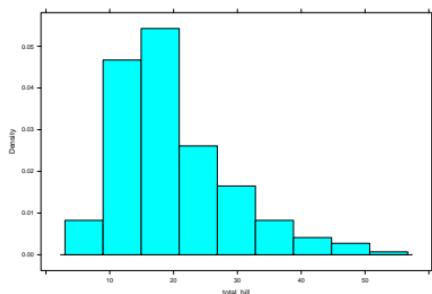
```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
tips <- read.csv2("tips.csv")
# Alternativ - heruntergeladene Datei einlesen:
# tips <- read.csv2(file.choose())

library(mosaic) # Paket laden
```

⁷⁴Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

Deskriptive Analyse Rechnungshöhe

```
histogram( ~ total_bill, data = tips)
```



```
favstats( ~ total_bill, data = tips)
```

```
##   min      Q1 median      Q3    max      mean       sd     n missing
##  3.07 13.3475 17.795 24.1275 50.81 19.78594 8.902412 244         0
```

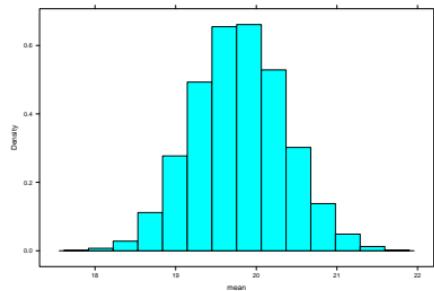
Übung 86: Verteilung Rechnungshöhe

Welche der folgenden Aussagen stimmt?

- A. Die Rechnungshöhe ist gleichverteilt.
- B. Die Rechnungshöhe ist multimodal.
- C. Die Rechnungshöhe ist normalverteilt.
- D. Die Rechnungshöhe ist linksschief.
- E. Die Rechnungshöhe ist rechtsschief.

Bootstrap Verteilung mittlere Rechnungshöhe

```
set.seed(1896) # Reproduzierbarkeit  
  
# 10000 Bootstrap Stichproben, Mittelwert berechnen  
Bootvtlg <- do(10000) *  
  mean(~ total_bill, data = resample(tips))  
  
histogram(~ mean, data = Bootvtlg)
```

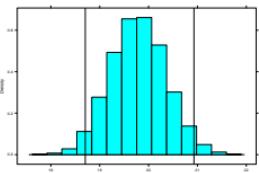


Übung 87: Verteilung mittlere Rechnungshöhe

Welche der folgenden Aussagen stimmt?

- A. Der Mittelwert der Rechnungshöhe ist gleichverteilt.
- B. Der Mittelwert der Rechnungshöhe ist multimodal.
- C. Der Mittelwert der Rechnungshöhe ist normalverteilt.
- D. Der Mittelwert der Rechnungshöhe ist linksschief.
- E. Der Mittelwert der Rechnungshöhe ist rechtsschief.

Übung 88: Konfidenzintervall



```
quantile( ~ mean, probs = c(0.025, 0.975), data = Bootvtlg)
```

```
##      2.5%    97.5%
```

```
## 18.70401 20.92869
```

Stimmt die Aussage: Mit 95 % Sicherheit überdeckt der Bereich 18.7\$ bis 20.93\$ eine zufällig ausgewählte Beobachtung?

- ▶ Ja.
- ▶ Nein.

Übung 89: Rechnungshöhe

Welches ist das richtige Testverfahren, um zu testen, ob die mittlere Rechnungshöhe in der Population über 15 \$⁷⁵ liegt (Forschungsthese)?

- A. t-Test für eine Stichprobe, ungerichtet.
- B. t-Test für eine Stichprobe, gerichtet.
- C. gepaarter t-Test, ungerichtet.
- D. gepaarter t-Test, gerichtet.
- E. Varianzanalyse.

⁷⁵hier μ_0 (willkürlich) gewählt

Übung 90: Rechnungshöhe R Befehl

Welches ist der richtige Befehl um diesen Test durchzuführen?

A.

```
t.test(~ total_bill, mu=15, alternative="greater",
       data = tips)
```

B.

```
t.test(~ total_bill, mu=15, alternative="less",
       data = tips)
```

```
t.test(~ total_bill, # Variable, die analysiert wird  
       mu=15, alternative="greater", # Optionen  
       data = tips) # Datensatz  
  
##  
##  One Sample t-test  
##  
## data: total_bill  
## t = 8.3976, df = 243, p-value = 1.909e-15  
## alternative hypothesis: true mean is greater than 15  
## 95 percent confidence interval:  
## 18.84492      Inf  
## sample estimates:  
## mean of x  
## 19.78594
```

Übung 91: Testergebnis Rechnungshöhe

Wird die Nullhypothese $H_0 : \mu \leq 15$ gegen $H_A : \mu > 15$ zum Signifikanzniveau $\alpha = 5\%$ verworfen?

- ▶ Ja.
- ▶ Nein.

Übung 92: p-Wert

Was würde passieren, wenn die vorher festgelegte Hypothese⁷⁶ nicht $H_0 : \mu \leq 15$ gegen $H_A : \mu > 15$ sondern $H_0 : \mu \leq 19.5$ gegen $H_A : \mu > 19.5$ lauten würde?

- A. Der p-Wert wird kleiner.
- B. Der p-Wert wird größer.
- C. Der p-Wert ändert sich nicht.

⁷⁶Hypothesen dürfen **nicht** nach der Analyse angepasst werden!

Gepaarter t-Test

Zeigen die Daten, dass die mittlere relative Trinkgeldhöhe signifikant über 10 % liegt?

Betrachte dazu je Beobachtung die Differenz $x_d = x_{\text{tip}} - 0.1 \cdot x_{\text{total_bill}}$:

```
t.test(~ I(tip - 0.1*total_bill), data=tips, alternative="greater")
```

```
##  
## One Sample t-test  
##  
## data: I(tip - 0.1 * total_bill)  
## t = 15.602, df = 243, p-value < 2.2e-16  
## alternative hypothesis: true mean is greater than 0  
## 95 percent confidence interval:  
## 0.9117688      Inf  
## sample estimates:  
## mean of x  
## 1.019684
```

Übung 93: Gepaarter t-Test

Was sagt der p-value $< 2.2\text{e-}16$ aus?

- A. Die Wahrscheinlichkeit, dass die Nullhypothese stimmt, ist kleiner als $2.2 \cdot 10^{-16}$.
- B. Die Wahrscheinlichkeit, dass die Alternativhypothese stimmt, ist kleiner als $2.2 \cdot 10^{-16}$.
- C. Weder A noch B.

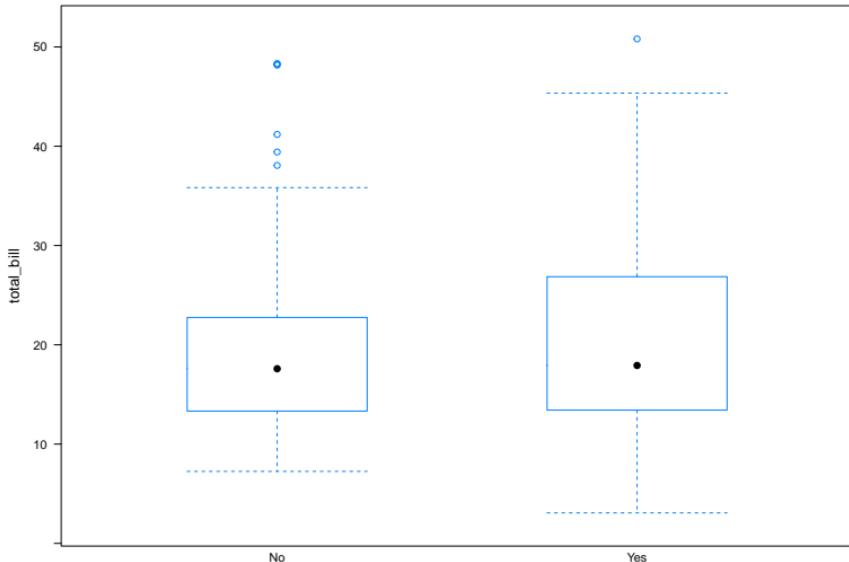
Übung 94: Fehlerart t-Test

Angenommen, in Wirklichkeit gilt $\mu_d = 0$. Welcher Fehler wurde begangen?

- A. Fehler 1. Art, α -Fehler.
- B. Fehler 2. Art, β -Fehler.

Boxplot Rechnungshöhe Raucher / Nichtraucher

```
bwplot(total_bill ~ smoker, data = tips)
```



Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

```
# Mittelwert Stichprobe  
mean(total_bill ~ smoker, data = tips)  
  
##          No        Yes  
## 19.18828 20.75634  
  
# Differenz Mittelwert Stichprobe  
diffmean(total_bill ~ smoker, data = tips)  
  
## diffmean  
## 1.568066
```

Übung 95: Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

Welche Aussage stimmt – für die Stichprobe?

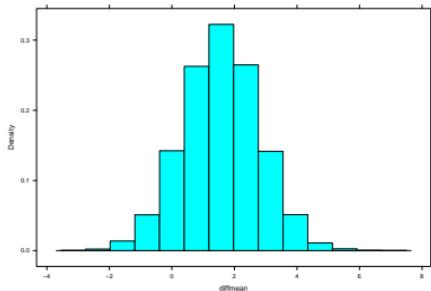
- A. $\bar{x}_{\text{Smoker Yes}} - \bar{x}_{\text{Smoker No}} = 0$
- B. $\bar{x}_{\text{Smoker Yes}} - \bar{x}_{\text{Smoker No}} \neq 0$

9. Inferenz numerischer Daten

Bootstrap Konfidenzintervall Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

```
set.seed(1896) # Reproduzierbarkeit
Bootvtlg <- do(10000) *
  diffmean(total_bill ~ smoker, data = resample(tips))

histogram(~ diffmean, data = Bootvtlg)
```



```
quantile(~ diffmean, data = Bootvtlg, probs = c(0.025, 0.975))
```

```
##          2.5%      97.5%
## -0.7991132  3.9476430
```

Übung 96: Testverfahren Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

Welches ist das Testverfahren, um zu testen, ob die mittlere Rechnungshöhe in der Population bei Rauchern und Nichtrauchern gleich ist, d. h., die Forschungsthese lautet:
Es gibt einen Unterschied im Mittelwert der Population?

- A. gepaarter t-Test, ungerichtet.
- B. gepaarter t-Test, gerichtet.
- C. t-Test für unabhängige Stichproben, ungerichtet.
- D. t-Test für unabhängige Stichproben, gerichtet.

Übung 97: Hypothese Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

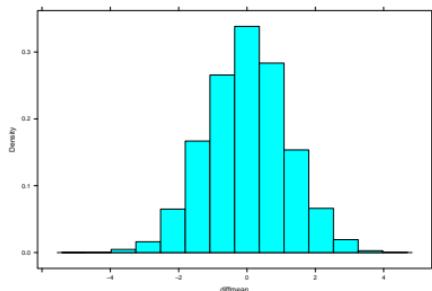
Wie lautet das richtige Hypothesenpaar?

- A. $H_0 : \mu_{\text{Smoker Yes}} \neq \mu_{\text{Smoker No}}$ vs. $H_A : \mu_{\text{Smoker Yes}} = \mu_{\text{Smoker No}}$
- B. $H_0 : \mu_{\text{Smoker Yes}} = \mu_{\text{Smoker No}}$ vs. $H_A : \mu_{\text{Smoker Yes}} \neq \mu_{\text{Smoker No}}$
- C. $H_0 : \bar{x}_{\text{Smoker Yes}} \neq \bar{x}_{\text{Smoker No}}$ vs. $H_A : \bar{x}_{\text{Smoker Yes}} = \bar{x}_{\text{Smoker No}}$
- D. $H_0 : \bar{x}_{\text{Smoker Yes}} = \bar{x}_{\text{Smoker No}}$ vs. $H_A : \bar{x}_{\text{Smoker Yes}} \neq \bar{x}_{\text{Smoker No}}$
- E. $H_0 : \pi_{\text{Smoker Yes}} \neq \pi_{\text{Smoker No}}$ vs. $H_A : \pi_{\text{Smoker Yes}} = \pi_{\text{Smoker No}}$

Permutationstest Differenz mittlere Rechnungshöhe Raucher / Nichtraucher

```
set.seed(1896) # Reproduzierbarkeit
Nullvtlg <- do(10000) *
  diffmean(total_bill ~ shuffle(smoker), data = tips)

histogram(~ diffmean, data = Nullvtlg)
```



```
# Absolute Abweichung Stichprobe
dm <- abs(diffmean(total_bill ~ smoker, data = tips))

# Anteil Abweichungen unter H_0 größer als in Stichprobe
prop(~ abs(diffmean) >= dm, data = Nullvtlg)

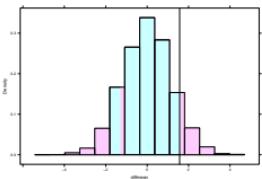
##    TRUE
## 0.1865
```

t-Test Rechnungshöhe Raucher / Nichtraucher

```
t.test(total_bill ~ # Abhängige Variable
       smoker, # Unabhängige Variable
       data = tips) # Datensatz

##
## Welch Two Sample t-test
##
## data: total_bill by smoker
## t = -1.2843, df = 169.63, p-value = 0.2008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.9783704 0.8422385
## sample estimates:
## mean in group No mean in group Yes
##           19.18828          20.75634
```

Übung 98: Testentscheidung Rechnungshöhe Raucher / Nichtraucher



Liefern die Daten genug Belege für die Forschungsthese $H_A : \mu_{\text{Smoker Yes}} \neq \mu_{\text{Smoker No}}$, d. h., kann die Nullhypothese verworfen werden?

- ▶ Ja.
- ▶ Nein.

```
set.seed(1896) # Reproduzierbarkeit
t.test(total_bill ~ smoker,
       data = sample(tips, size = 100))

##
## Welch Two Sample t-test
##
## data: total_bill by smoker
## t = 0.26883, df = 71.558, p-value = 0.7888
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.437518 4.509063
## sample estimates:
## mean in group No mean in group Yes
##           20.24919          19.71342
```

```
set.seed(1896) # Reproduzierbarkeit
t.test(total_bill ~ smoker,
       data = sample(tips, size = 200))

##
## Welch Two Sample t-test
##
## data: total_bill by smoker
## t = -0.47295, df = 127.15, p-value = 0.6371
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.368647 2.068999
## sample estimates:
## mean in group No mean in group Yes
##           19.33061          19.98043
```

Übung 99: Stichprobengröße

Welche Auswirkungen hat, unter sonst gleichen Umständen, ein größerer Stichprobenumfang n ?

- A. Das Konfidenzintervall wird schmäler.
- B. Das Konfidenzintervall wird breiter.

Der p-Wert gibt (nur) die Randwahrscheinlichkeit der Teststatistik unter der Nullhypothese an. Er sagt nicht, wie groß / relevant ein Unterschied ist.

Cohens d⁷⁷ ist ein Maß für die Überlappung:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{sd_{\text{pool}}}$$

mit

$$sd_{\text{pool}} = \sqrt{\frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2 \right)}$$

Einmalige Installation

```
install.packages("lsr")
```

Paket laden

```
library(lsr)
```

⁷⁷Anwendbar für den Vergleich zweier Mittelwerte. Es gibt auch weitere Effektgrößen. Siehe z. B. Paket [compute.es](#).

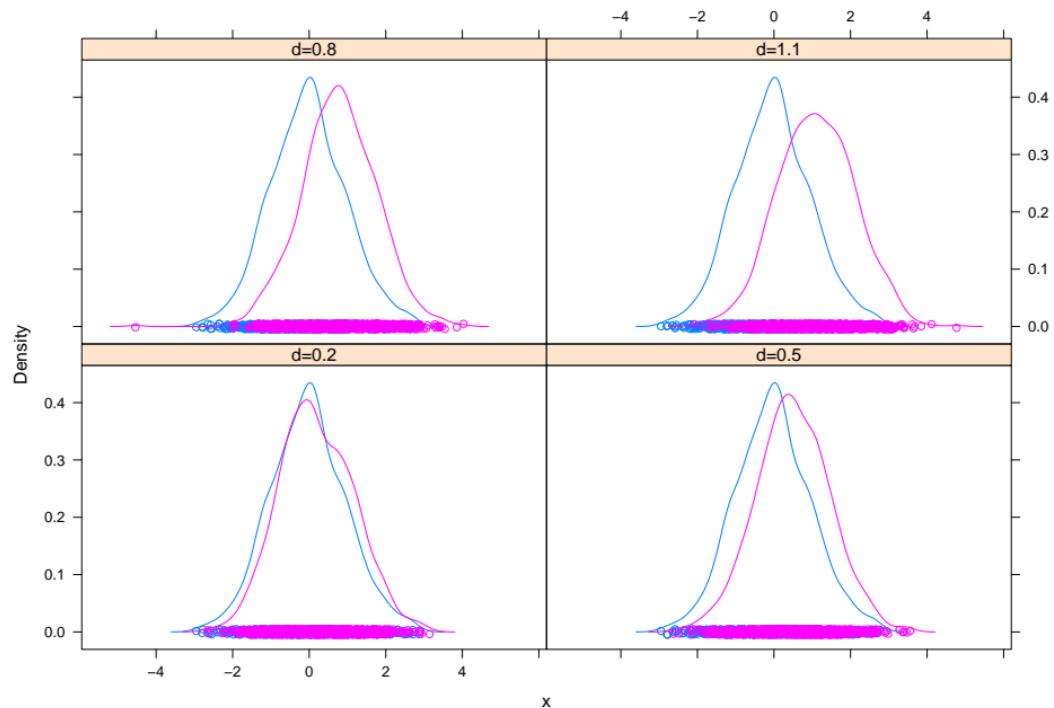
Daumenregel:

- ▶ $|d| \geq 0.2$ kleiner Effekt.
- ▶ $|d| \geq 0.5$ mittlerer Effekt.
- ▶ $|d| \geq 0.8$ großer Effekt.

```
cohensD(total_bill ~ smoker, data=tips)
```

```
## [1] 0.176426
```

Beispiel Effektgrößen



Übung 100: Effektgröße Rauchen

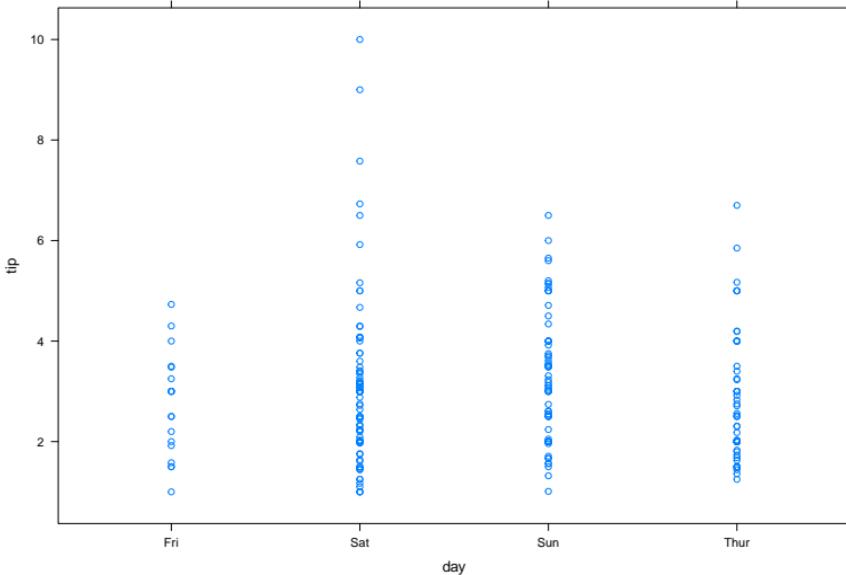
Welche Aussage stimmt? Zwischen Rauchen und Trinkgeld liegt

- A. kein,
- B. ein kleiner,
- C. ein mittlerer,
- D. ein großer

Effekt vor.

Zusammenhang Trinkgeld und Wochentag

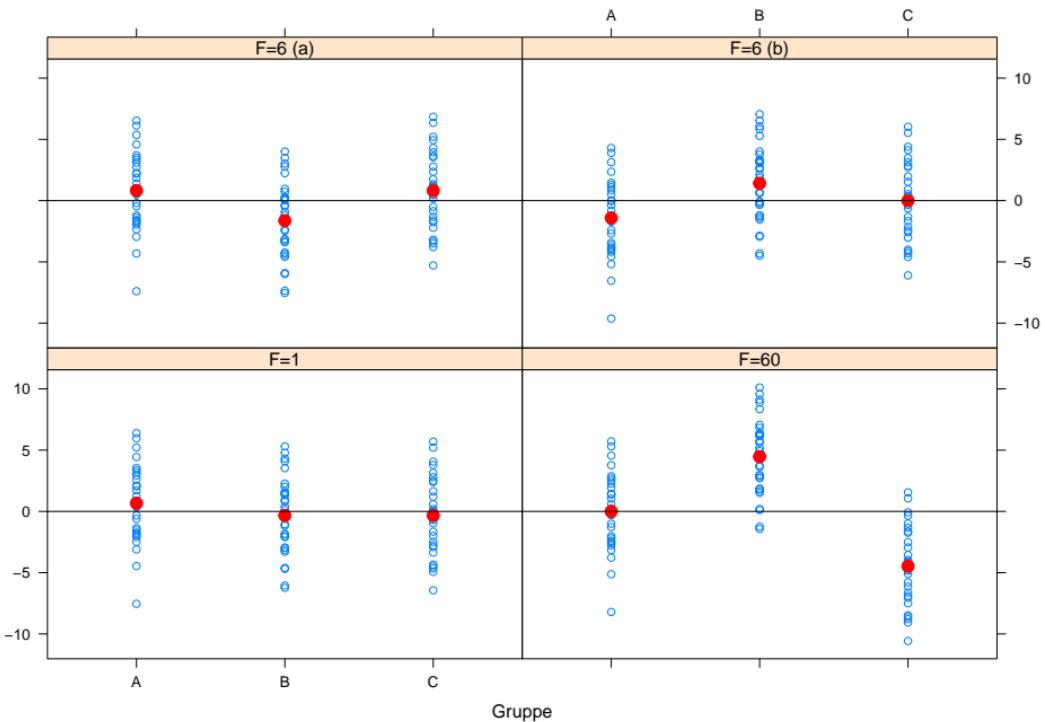
```
xyplot(tip ~ day, data = tips)
```



- ▶ Vergleich des Lagemaßes μ_i bei zwei oder mehr Stichproben. Ein- oder mehrfaktoriell möglich, bei mehr als einem Einfluss auch Wechselwirkungen.
- ▶ Nullhypothese: Lagemaß μ_i für alle Gruppen gleich.
- ▶ Die **Gesamtstreuung (SST)** wird zerlegt in die **Streuung zwischen den Stichproben/Gruppen (SSG)** und die **Streuung innerhalb der Stichproben/Gruppen (SSE)**:

$$\underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{SST} = \underbrace{\sum_{j=1}^K n_j (\bar{x}_j - \bar{x})^2}_{SSG} + \underbrace{\sum_{j=1}^K \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2}_{SSE}$$

- ▶ Ist das Verhältnis der Streuung zwischen den Gruppen im Verhältnis zur Streuung innerhalb der Gruppen groß (Teststatistik F), so ist dies unter der Nullhypothese unwahrscheinlich.
- ▶ Voraussetzung: Daten innerhalb der Stichproben / Gruppen unabhängig, identisch, normalverteilt



⁷⁸Video <https://www.causeweb.org>: Crawford S © Use ANOVA

Übung 101: Testverfahren Trinkgeld und Wochentag

Welches ist das richtige Testverfahren, um zu testen, ob die mittlere Trinkgeldhöhe in der Population über alle Wochentage gleich ist, d. h., die Forschungsthese lautet: Es gibt mindestens einen Unterschied im Mittelwert der Population zwischen den Wochentagen?

- A. gepaarter t-Test, ungerichtet.
- B. gepaarter t-Test, gerichtet.
- C. t-Test für unabhängige Stichproben, ungerichtet.
- D. t-Test für unabhängige Stichproben, gerichtet.
- E. Varianzanalyse.

Varianzanalyse in R

```
# Speichere Ergebnis der Varianzanalyse aov() in "ergaov"
ergaov <- aov(tip ~ # Abhängige Variable
                day, # Unabhängige Variable
                data = tips) # Datensatz

# Zeige Zusammenfassung von "ergaov"
summary(ergaov)
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## day                  3    9.5   3.175   1.672  0.174
## Residuals     240  455.7   1.899
```

Kann die Nullhypothese $H_0 : \mu_{\text{Fri}} = \mu_{\text{Sat}} = \mu_{\text{Sun}} = \mu_{\text{Thu}}$ verworfen werden, d. h., kann anhand der Stichprobenunterschiede der Mittelwerte

```
mean(tip ~ day, data=tips)
```

```
##      Fri      Sat      Sun      Thur
## 2.734737 2.993103 3.255132 2.771452
```

auf mindestens einen Unterschied in den Mittelwerten in der Population geschlossen werden ($\alpha = 0.05$)?

- A. Ja.
- B. Nein.
- C. Weiß nicht.

Wenn man statt einer ANOVA alle $\binom{4}{2} = \frac{4 \cdot (4-1)}{2} = 6$ Kombinationen (d. h. Donnerstag und Freitag, Donnerstag und Samstag usw.) ausprobiert hätte, hätte sich der α -Fehler kummuliert⁷⁹:

$$P(\text{Fehler 1. Art}) = 1 - (1 - 0.05)^6 = 0.265$$

Das globale Signifikanzniveau $\alpha = 0.05$ wäre nicht eingehalten!

⁷⁹hier: $\alpha = 0.05$

Offene Übung 103: Trinkgeld Mann / Frau

Analysieren Sie die Höhe des Trinkgeldes und inwieweit sich dies zwischen den Geschlechtern unterscheidet.

10 Lineare Regression

Übung 104: Skalenniveau Trinkgeldhöhe

Welches Skalenniveau hat die Variable Trinkgeldhöhe?

- A. Kategorial - nominal.
- B. Kategorial - ordinal.
- C. Numerisch - Intervallskala.
- D. Numerisch - Verhältnisskala.

Modellierung: Lineare Regression

- **Überwachtes Lernen** (engl.: supervised learning): Kann ein Teil der Variation einer abhängigen Variable y durch unabhängige Variable(n) x modelliert werden:

$$y = f(x) + \epsilon^{80}$$

- Schätze \hat{f} anhand der Daten / Stichprobe
- **Annahme:** f ist eine *lineare* Funktion, d. h., $f(x) = \beta_0 + \beta_1 \cdot x$ Hier: y **numerisch**, nur eine unabhängige Variable x . Siehe `lm()`
 - β_0 : Achsenabschnitt
 - β_1 : Steigung, d. h. durchschnittliche Änderung von y , wenn x eine Einheit größer wird
- **Methode der kleinsten Quadrate:** Bestimme Vektor $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ so, dass für $\hat{\epsilon}_i = y_i - \hat{f}(x_i) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ der Wert $\sum \hat{\epsilon}_i^2$ minimal ist.

⁸⁰ ϵ : (zufälliger) Fehler, Residuum

- ▶ **Nullhypothese des Koeffiziententests:** Variable x_j hat keinen linearen Einfluss auf y , d. h., $H_0 : \beta_j = 0$
- ▶ **Voraussetzung:**
 - ▶ linearer Zusammenhang zwischen x und y
 - ▶ keine (einflussreichen) Ausreißer
 - ▶ Residuen unabhängig (d. h. keine (Auto)korrelation), identisch (insbesondere konstante Varianz) normalverteilt
- ▶ Das **Bestimmtheitsmaß** R^2 gibt den Anteil der im Modell erklärten Variation von y an:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- ▶ Modellierung der Klausurpunktzahl eines Studierenden auf Basis z. B. der Schulnote.
- ▶ Analyse des Gehaltes einer Mitarbeiter*in auf Basis von z. B. Ausbildungsdauer.
- ▶ Vorhersage der Seitenabrufe auf Basis der Fans, Follower und Art des Inhalts⁸¹.
- ▶ Modellierung der Risikos einer Analage (Betafaktor).
- ▶ Vorhersage der Verspätung von Flügen (s. Datensatz [nycflights13](#)).
- ▶ Vorhersage der Persönlichkeit anhand von Social-Media-Daten (s. [dieses Paper](#)).

Wo können Sie dies Verfahren einsetzen?

⁸¹z. B. Gewinnspiel, Rabatt.

Einlesen der *Tipping*⁸² Daten sowie laden des Pakets mosaic. Zufallszahlengenerator setzen.

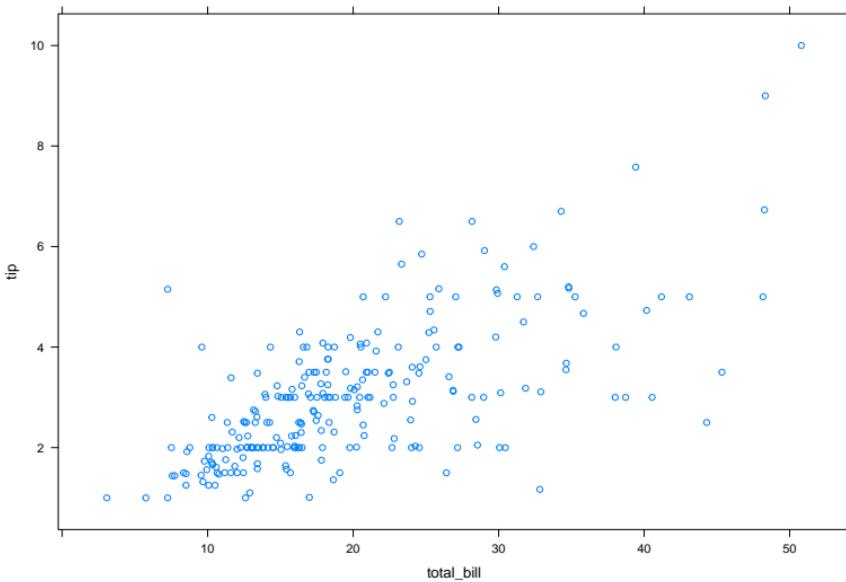
```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
tips <- read.csv2("tips.csv")
# Alternativ - heruntergeladene Datei einlesen:
# tips <- read.csv2(file.choose())

library(mosaic) # Paket laden
```

⁸²Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

Streudiagramm: Trinkgeld und Rechnungshöhe

```
xyplot(tip ~ total_bill, data = tips)
```



Übung 105: Korrelation Trinkgeld und Rechnungshöhe

Welche Aussage stimmt vermutlich für den Korrelationskoeffizient zwischen Trinkgeld und Rechnungshöhe?

- A. Der Korrelationskoeffizient liegt bei $r = -0.68$.
- B. Der Korrelationskoeffizient liegt bei $r = -0.23$.
- C. Der Korrelationskoeffizient liegt bei $r = 0.68$.
- D. Der Korrelationskoeffizient liegt bei $r = 0.23$.

Übung 106: Zusammenhang Trinkgeld und Rechnungshöhe

Welche Aussage stimmt vermutlich – aus inhaltlichen Gründen?

- A. Die Trinkgeldhöhe hängt ab von der Rechnungshöhe.
- B. Die Rechnungshöhe hängt ab von der Trinkgeldhöhe.
- C. Trinkgeld und Rechnungshöhe sind unabhängig.

Lineare Regression Trinkgeld auf Rechnungshöhe

```
# Speichere Ergebnis der Regression lm() in "erglm1"
erglm1 <- lm(tip ~ # abhängige Variable
              total_bill, # unabhängige Variable(n)
              data = tips) # Datensatz

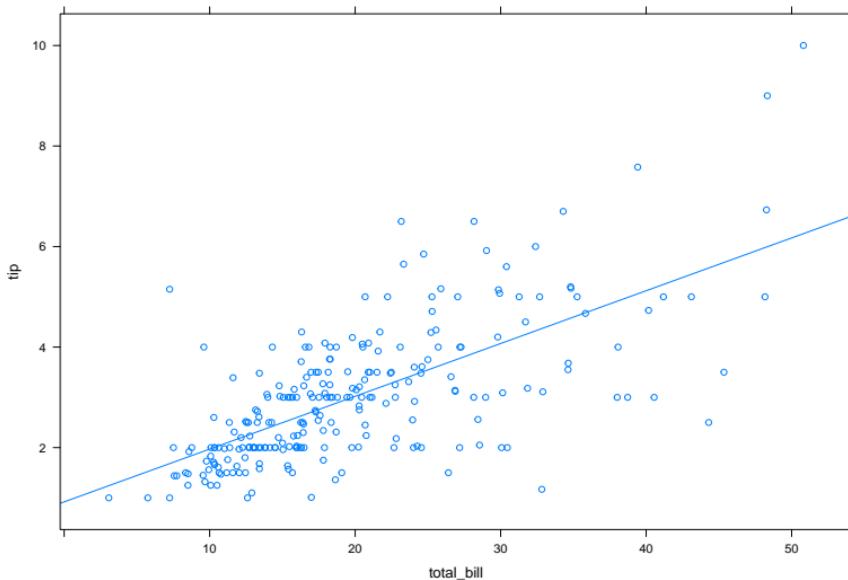
# Zeige Zusammenfassung von "erglm1"
summary(erglm1)

##
## Call:
## lm(formula = tip ~ total_bill, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.1982 -0.5652 -0.0974  0.4863  3.7434 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.920270  0.159735  5.761 2.53e-08 ***
## total_bill   0.105025  0.007365 14.260 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 242 degrees of freedom
## Multiple R-squared:  0.4566, Adjusted R-squared:  0.4544 
## F-statistic: 203.4 on 1 and 242 DF,  p-value: < 2.2e-16
```

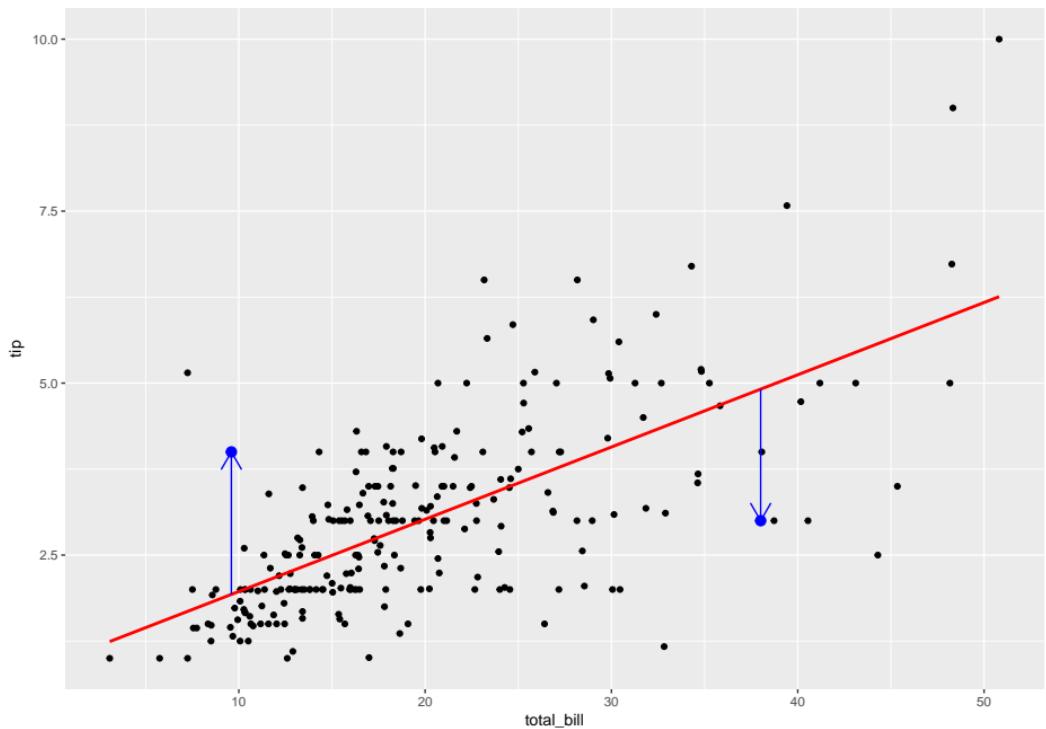
10. Lineare Regression

Regressionsgerade

```
plotModel(erglm1)
```



Residuen



Übung 107: Regression Trinkgeld auf Rechnungshöhe

Welche Aussage stimmt?

- A. Im Mittel steigt mit jedem Dollar Trinkgeld die Rechnungshöhe um 0.92.
- B. Im Mittel steigt mit jedem Dollar Trinkgeld die Rechnungshöhe um 0.11.
- C. Im Mittel steigt mit jedem Dollar Rechnungshöhe das Trinkgeld um 0.92.
- D. Im Mittel steigt mit jedem Dollar Rechnungshöhe das Trinkgeld um 0.11.

Geschätzte Regressionsgleichung

Die geschätzte Gleichung lautet:

$$\hat{y} = 0.9203 + 0.105 \cdot x$$

Übung 108: Prognose der Trinkgeldhöhe aus Rechnungshöhe

Für ein gegebenes $x_0 = 10$ lautet die Prognose $\hat{y}_0 = 0.9203 + 0.105 \cdot 10 = 1.9703$.

Stimmt die Aussage: Bei einer Rechnungshöhe von 10 \$ wird das Trinkgeld mit Sicherheit bei 1.97 \$ liegen?

- ▶ Ja.
- ▶ Nein.

```
predict(erglm1, # Modell
        # Neue Beobachtung mit x=10:
        newdata = data.frame(total_bill = 10),
        # Prognoseintervall:
        interval = "prediction")
```

```
##          fit      lwr      upr
## 1 1.970515 -0.05184074 3.99287
```

Übung 109: Bestimmtheitsmaß

Welche Aussage stimmt?⁸³

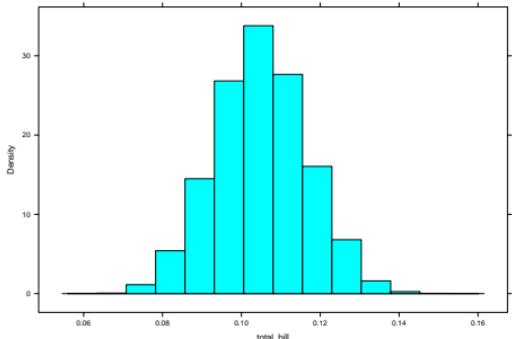
- A. Die Wahrscheinlichkeit, dass das Modell stimmt, liegt bei 46 %.
- B. 46 % der Beobachtungen werden richtig modelliert.
- C. 46 % der Variation der Rechnungshöhe werden modelliert.
- D. 46 % der Variation der Trinkgeldhöhe werden modelliert.

⁸³R Ausgabe: Multiple R-squared =0.4566.

Bootstrap Verteilung Steigungskoeffizient

```
set.seed(1896) # Reproduzierbarkeit
Bootvtlg <- do(10000) *
  lm(tip ~ total_bill, data = resample(tips))

histogram(~ total_bill, data = Bootvtlg)
```



```
quantile(~ total_bill, data = Bootvtlg,
          probs = c(0.025, 0.975))
```

```
##           2.5%      97.5%
## 0.08235625 0.12797229
```

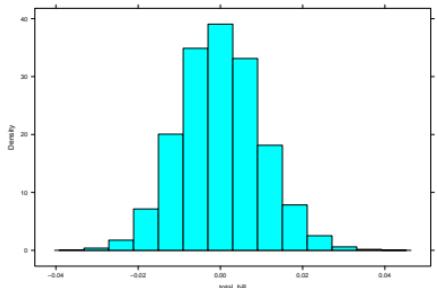
Permutationstest Verteilung Steigung (I/II)

Wenn $H_0 : \beta_1 = 0$ gilt, so sollte y in keinem (linearen) Zusammenhang zu x stehen:

```
set.seed(1896) # Reproduzierbarkeit
Nullvltlg <- do(10000) *
  lm(tip ~ shuffle(total_bill), data = tips)
```

Permutationstest Verteilung Steigung (II/II)

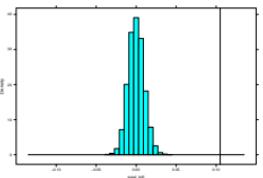
```
histogram(~total_bill, data = Nullvtlg)
```



```
quantile(~total_bill, data = Nullvtlg,  
        probs = c(0.025, 0.975))
```

```
##           2.5%      97.5%  
## -0.01876834  0.02000266
```

Übung 110: Permutationstest Steigung

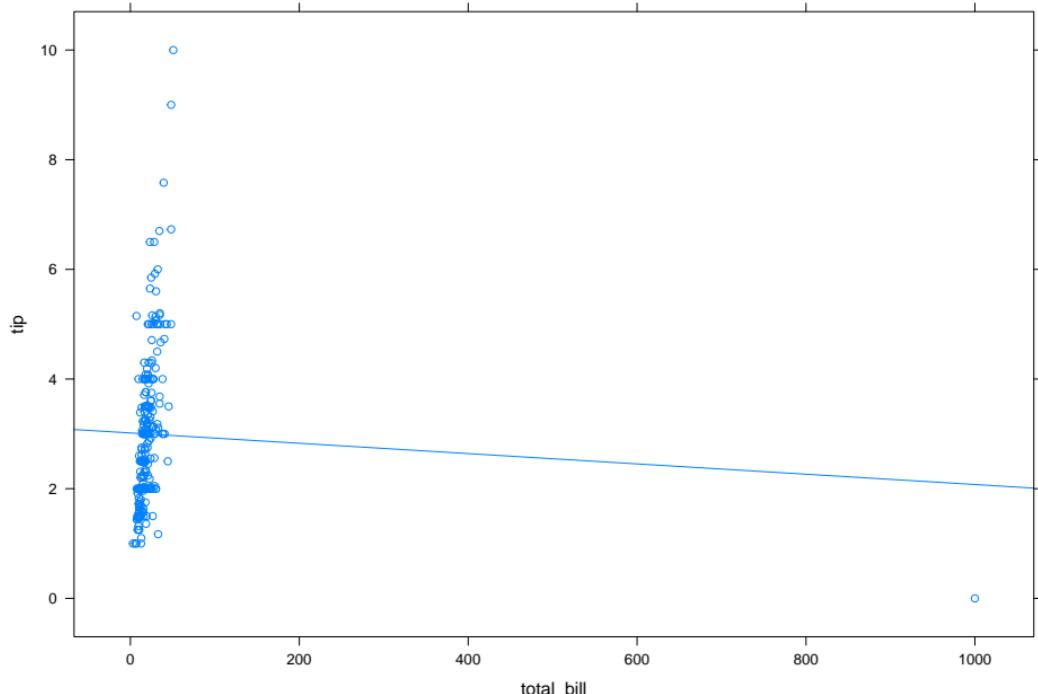


Welche Aussage stimmt?

- A. Die beobachtete Steigung der Stichprobe $\hat{\beta}_1 = 0.11$ ist unter $H_0 : \beta_1 = 0$ ein üblicher Wert.
- B. Die beobachtete Steigung der Stichprobe $\hat{\beta}_1 = 0.11$ ist unter $H_0 : \beta_1 = 0$ kein üblicher Wert.

Ausreißer

Beobachtungen, die horizontal und vertikal vom üblichen Zusammenhang abweichen, können die Regressionsgerade und die Modellgüte verändern.



Extrapolation

Vorsicht bei Vorhersagen für Werte außerhalb des bekannten, üblichen Wertebereiches.⁸⁴

```
predict(erglm1, # Modell
        # Neue Beobachtung mit x=1000:
        newdata = data.frame(total_bill = 1000),
        # Prognoseintervall:
        interval = "prediction")
```

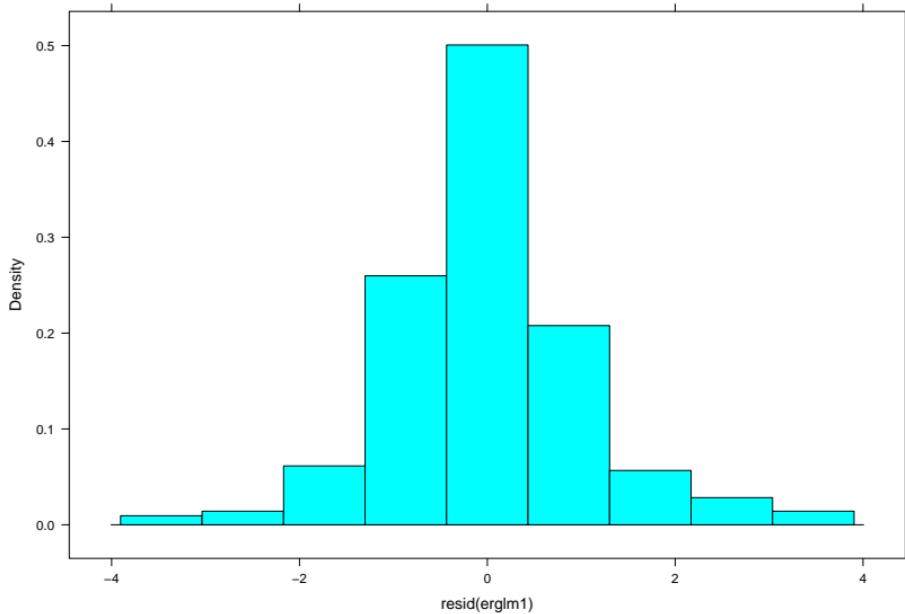
```
##          fit      lwr      upr
## 1 105.9448 91.58217 120.3074
```

⁸⁴Video <https://www.causeweb.org>: Posner M © How Far He'll Go

10. Lineare Regression

Verteilung Residuen

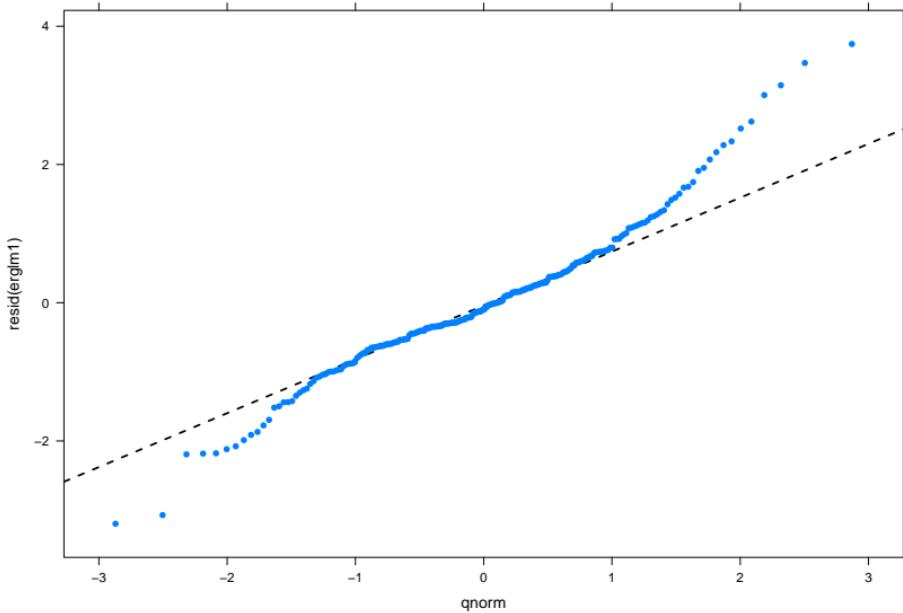
```
histogram( ~ resid(erglm1))
```



10. Lineare Regression

Q-Q Plot Residuen

```
xqqmath(~ resid(erglm1))
```



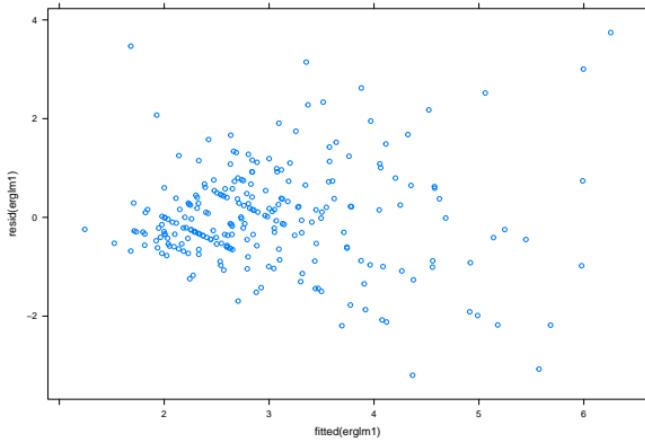
Übung 111: Verteilung Residuen

Stimmt die Aussage: Die Annahme einer Normalverteilung für die Residuen ist hier unkritisch?

- ▶ Ja.
- ▶ Nein.

Verteilung Residuen und angepasste Werte

```
xyplot(resid(erglm1) ~ fitted(erglm1))
```



Übung 112: Verteilung Residuen und angepasste Werte

Welche Aussage stimmt?

- A. Die Varianz der Residuen scheint unabhängig von der Höhe der angepassten Werte zu sein.
- B. Die Varianz der Residuen scheint mit der Höhe der angepassten Werte zu steigen.
- C. Die Varianz der Residuen scheint mit der Höhe der angepassten Werte zu fallen.

Regression nur mit Achsenabschnitt

```
mean(tip~1, data = tips)  
  
##           1  
## 2.998279  
  
lm(tip~1, data = tips)  
  
##  
## Call:  
## lm(formula = tip ~ 1, data = tips)  
##  
## Coefficients:  
## (Intercept)  
##           2.998
```

Übung 113: Regression nur mit Achsenabschnitt

Was gilt bei $\text{lm}(y \sim 1)$ für das Bestimmtheitsmaß?

- A. $R^2 = 0$
- B. $0 < R^2 < 1$
- C. $R^2 = 1$

Trinkgeld und Geschlecht

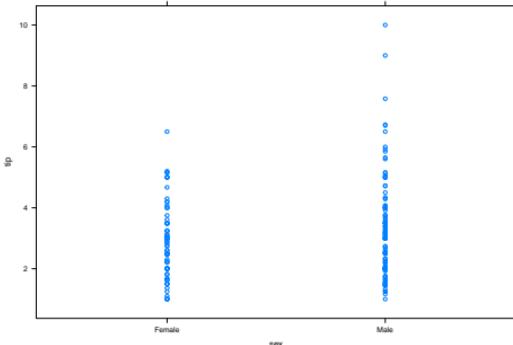
```
mean(tip ~ sex, data = tips)

##   Female      Male
## 2.833448 3.089618

diffmean(tip ~ sex, data = tips)

## diffmean
## 0.2561696

xyplot(tip ~ sex, data = tips)
```



Kategoriale Variablen werden numerisch / logisch kodiert.

Geschlecht (sex):

	Male
Female	0
Male	1

Wochentag (day):

	Sat	Sun	Thur
Fri	0	0	0
Sat	1	0	0
Sun	0	1	0
Thur	0	0	1

Regression Trinkgeld auf Geschlecht

```
erglm2 <- lm(tip ~ sex, data = tips)
summary(erglm2)
```

```
##
## Call:
## lm(formula = tip ~ sex, data = tips)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0896 -1.0896 -0.0896  0.6666  6.9104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.8334    0.1481 19.137  <2e-16 ***
## sexMale     0.2562    0.1846  1.388    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.381 on 242 degrees of freedom
## Multiple R-squared:  0.007896, Adjusted R-squared:  0.003797
## F-statistic: 1.926 on 1 and 242 DF,  p-value: 0.1665
```

Übung 114: Regression Trinkgeld und Geschlecht

Welche Aussage stimmt für die Stichprobe?

- A. Im Mittel geben Männer 0.26\$ mehr Trinkgeld als Frauen.
- B. Im Mittel geben Frauen 0.26\$ mehr Trinkgeld als Männer.
- C. Männer geben immer 0.26\$ mehr Trinkgeld als Frauen.
- D. Frauen geben immer 0.26\$ mehr Trinkgeld als Männer.

Offene Übung 115: Trinkgeld je Geschlecht

Fassen Sie die vorangegangene Analyse zusammen. Wie lautete die Forschungsfrage, Modell, Hypothesen und die Antwort auf die Forschungsfrage.

Regression eines Anteils

```
prop(smoker ~ time, success = "Yes", data = tips)

## Yes.Dinner  Yes.Lunch
## 0.3977273 0.3382353

diffprop(smoker ~ time, success = "Yes", data = tips)

##      diffprop
## -0.05949198

lm( (smoker=="Yes") ~ time, data = tips)

##
## Call:
## lm(formula = (smoker == "Yes") ~ time, data = tips)
## 
## Coefficients:
## (Intercept)    timeLunch
##          0.39773     -0.05949
```

- ▶ Eine **Lineare** Regression eines Anteils kann nicht so interpretiert werden wie die lineare Regression eines numerischen Merkmals.⁸⁵ Insbesondere ist $\hat{y} \notin \{0, 1\}$ und die **Annahmen sind verletzt**, d. h., p-Werte etc. stimmen **nicht**.
- ▶ Die richtige Herangehensweise wäre z. B. eine **Logistische** Regression.

⁸⁵ $\hat{\beta}, R^2$

Übung 116: Beurteilung lineares Modell

Woran können Sie erkennen, ob Sie ein *gutes* Modell haben – bei einer metrischen abhängigen Variable y ?

- A. An einem kleinen p-Wert.
- B. An einem großen p-Wert.
- C. An einer im Betrag kleinen geschätzten Steigung.
- D. An einer im Betrag großen geschätzten Steigung.
- E. An einem großen R^2 .

Modellgleichung:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_p \cdot x_{ip} + \epsilon_i$$

Interpretation der Koeffizienten (Schätzwerte, p-Werte): unter sonst gleichen Umständen, d. h., die anderen Variablen bleiben im Modell konstant/unverändert (*ceteris paribus*): marginaler Effekt.⁸⁶

⁸⁶Durch Versuchsplanung oder eine vorgelagerte Hauptkomponentenanalyse können unabhängige erklärende Variablen x_j erzeugt werden.

Übung 117: Multiple Regressionskoeffizienten

Können sich die geschätzten Werte und deren p-Werte ändern, wenn Variablen ins Modell hinzugenommen oder weggenommen werden?

- ▶ Ja.
- ▶ Nein.

Übung 118: Bestimmtheitsmaß

Kann sich das Bestimmtheitsmaß R^2 ändern, wenn Variablen ins Modell hinzugenommen oder weggemommen werden?

- ▶ Ja.
- ▶ Nein.

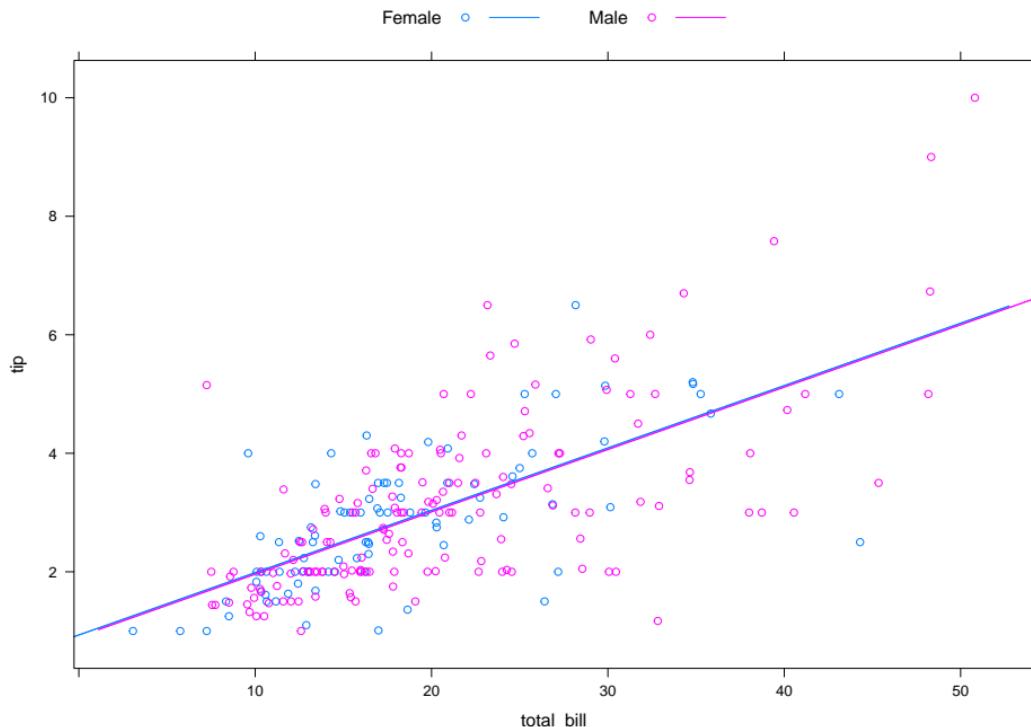
Trinkgeldhöhe als Funktion von Rechnungshöhe und Geschlecht

Modelliere Trinkgeldhöhe als lineare Funktion von Rechnungshöhe und Geschlecht:

```
erglm3 <- lm(tip ~ # abhängige Variable  
               total_bill + sex, # unabhängige Variablen  
               data = tips) # Datensatz  
  
summary(erglm3)  
  
##  
## Call:  
## lm(formula = tip ~ total_bill + sex, data = tips)  
##  
## Residuals:  
##     Min      1Q  Median      3Q     Max  
## -3.1914 -0.5596 -0.0875  0.4845  3.7465  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.933278   0.173756   5.371 1.84e-07 ***  
## total_bill    0.105232   0.007458  14.110 < 2e-16 ***  
## sexMale     -0.026609   0.138334  -0.192   0.848  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.024 on 241 degrees of freedom  
## Multiple R-squared:  0.4567, Adjusted R-squared:  0.4522  
## F-statistic: 101.3 on 2 and 241 DF,  p-value: < 2.2e-16
```

Modell Multiple Regression

```
plotModel(erglm3)
```



Übung 119: Regression Trinkgeld auf Rechnungshöhe und Geschlecht

Stimmt die Aussage: Bei gleicher Rechnungshöhe geben Männer in der Stichprobe im Mittel mehr Trinkgeld als Frauen.

- ▶ Ja.
- ▶ Nein.

Bootstrap Multiple Regression

```
set.seed(1896) # Reproduzierbarkeit
Bootvtlg <- do(10000) * lm(tip ~ total_bill + sex,
                             data = resample(tips))
confint(Bootvtlg)

##          name    lower      upper level   method   estimate
## 1 Intercept 0.52516536 1.3474283 0.95 percentile 0.93327849
## 2 total_bill 0.08255927 0.1284278 0.95 percentile 0.10523236
## 3 sexMale -0.26823214 0.2186610 0.95 percentile -0.02660871
## 4 sigma    0.89069928 1.1389718 0.95 percentile 1.02408737
## 5 r.squared 0.33253246 0.5806577 0.95 percentile 0.45670000
## 6 F        60.03312632 166.8547403 0.95 percentile 101.29274612
```

Übung 120: Inferenz Regression Trinkgeld und Geschlecht

Gegeben die Rechnungshöhe, kann die Nullhypothese $\beta_2 = \beta_{\text{sex}} = 0$ zum Signifikanzniveau $\alpha = 5\%$ verworfen werden?

- ▶ Ja.
- ▶ Nein.

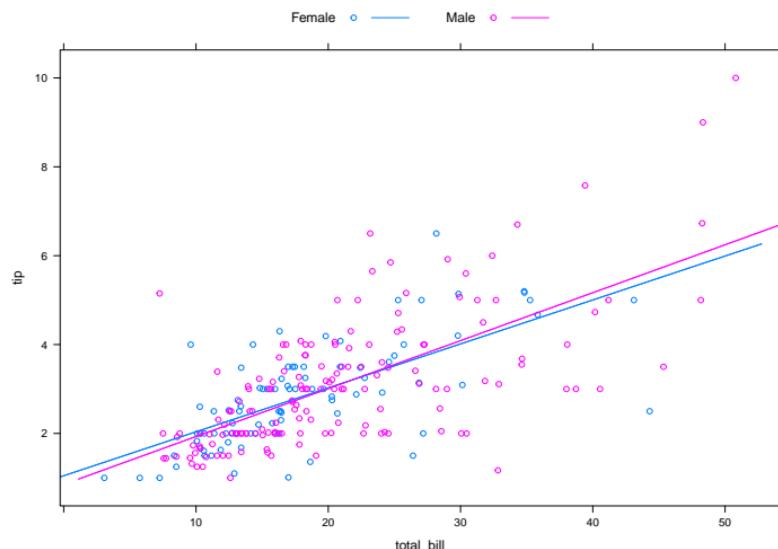
Welches ist die korrekteste Interpretation von $\hat{\beta}_1 = \hat{\beta}_{\text{total_bill}} = 0.11$?

- A. Mit jedem \$ Rechnungshöhe steigt das Trinkgeld um 0.11 \$.
- B. Mit jedem \$ Rechnungshöhe steigt das Trinkgeld im Mittel um 0.11 \$.
- C. Mit jedem \$ Rechnungshöhe steigt das Trinkgeld im Mittel um 0.11 \$, gegeben alle anderen Faktoren bleiben konstant.
- D. In einem linearen Modell steigt mit jedem \$ Rechnungshöhe das Trinkgeld im Mittel um 0.11 \$, gegeben alle anderen Faktoren bleiben konstant.
- E. In der Stichprobe steigt in einem linearen Modell mit jedem \$ Rechnungshöhe das Trinkgeld im Mittel um 0.11 \$, gegeben alle anderen Faktoren bleiben konstant.

Wechselwirkung, Interaktion

Hängt evt. auch die Steigung mit dem Geschlecht zusammen?

```
erglm4 <- lm(tip ~  
               total_bill + sex + total_bill:sex,  
               data = tips)  
plotModel(erglm4)
```



Übung 122: Wechselwirkung

Welches Geschlecht gibt im Mittel, unter sonst gleichen Umständen, mit zunehmender Rechnungshöhe mehr zusätzliches Trinkgeld?

- A. Frauen.
- B. Männer
- C. Beide gleich.

Ergebnis Wechselwirkung

```
summary(erglm4)

##
## Call:
## lm(formula = tip ~ total_bill + sex + total_bill:sex, data = tips)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -3.2232 -0.5660 -0.0977  0.4796  3.6675 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.048020  0.272498   3.846 0.000154 ***
## total_bill   0.098878  0.013808   7.161 9.75e-12 ***
## sexMale     -0.195872  0.338954  -0.578 0.563892  
## total_bill:sexMale 0.008983  0.016417   0.547 0.584778  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.026 on 240 degrees of freedom
## Multiple R-squared:  0.4574, Adjusted R-squared:  0.4506 
## F-statistic: 67.43 on 3 and 240 DF,  p-value: < 2.2e-16
```

ANOVA Tabelle Wechselwirkung

```
anova(erglm4)
```

```
## Analysis of Variance Table
##
## Response: tip
##                               Df  Sum Sq Mean Sq F value Pr(>F)
## total_bill             1 212.424 212.424 201.9597 <2e-16 ***
## sex                   1    0.039   0.039   0.0369 0.8478
## total_bill:sex        1    0.315   0.315   0.2994 0.5848
## Residuals            240 252.435    1.052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Formeln bieten innerhalb der Modellierung in R viele Möglichkeiten:

- ▶ +: Hinzunahme von Variablen
- ▶ -: Herausnahme von Variablen (-1 für Achsenabschnitt)
- ▶ :: Wechselwirkung von Variablen
- ▶ *: Hinzunahme von Variablen und deren Wechselwirkung
- ▶ /: hierarchisch untergeordnet (engl.: nested)
- ▶ I(): Arithmetische Operationen der Variablen

Woran können Sie am ehesten erkennen, dass eine Variable x_j zur Modellierung von y beiträgt?

- A. An einem kleinen $|\hat{\beta}_j|$.
- B. An einem großen $|\hat{\beta}_j|$.
- C. An einem kleinen p-Wert.
- D. An einem großen p-Wert.

Die Wahl der *wichtigen* Variablen im Modell ist nicht trivial. Dabei wird ein Kriterium wie z. B. AIC⁸⁷ zur Modellevaluierung verwendet. Mögliche Herangehensweisen z. B.

- ▶ Vorwärts Auswahl: Fangt nur mit Achsenabschnitt an und füge schrittweise neue Variablen hinzu, bis sich die Modellgüte nicht mehr verbessert.⁸⁸
- ▶ Rückwärts Auswahl: Fangt mit allen Variablen an und eliminiere schrittweise einzelne Variablen, bis sich die Modellgüte nicht mehr verbessert.

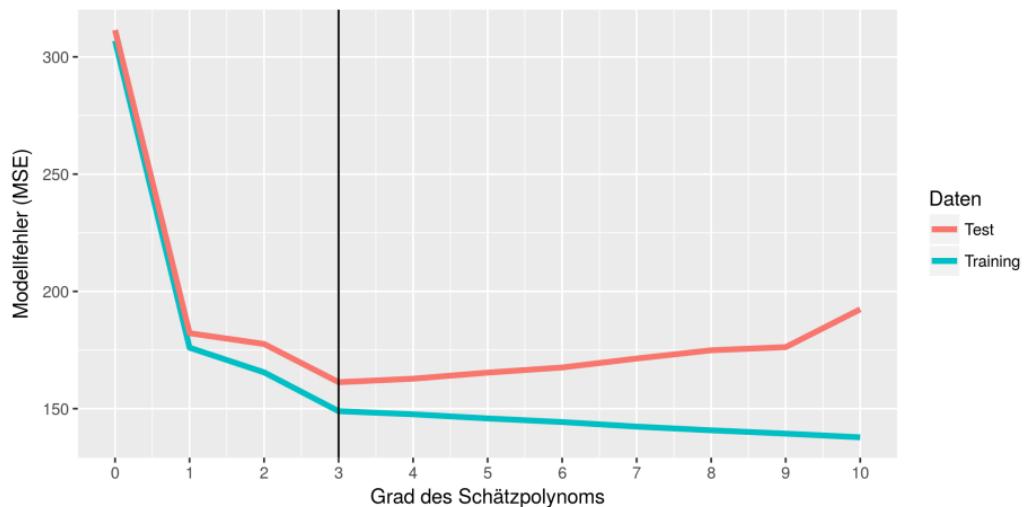
In R: z. B. `step()`

⁸⁷ Akaike Informations Kriterium, siehe z. B. <https://www.otexts.org/fpp/5/3>

⁸⁸ Das *normale* R^2 steigt mit jeder Variablen im Modell – auch wenn diese nicht mit y zusammenhängt.

Schätzen (auf Basis von $n = 100$ Beobachtungen: Training) und Testen (auf Basis von $n = 10000$: Test) des Polynoms

$$y = -x^3 + 8x^2 - 9x - 18 + \epsilon$$



Übung 124: Modellkomplexität (I / II)

Stimmt die Aussage: Je komplexer⁸⁹ ein Modell ist, desto besser erklärt es die vorhandenen Daten?

- ▶ Ja.
- ▶ Nein.

⁸⁹Hier: Grad des Polynoms.

Übung 125: Modellkomplexität (II / II)

Stimmt die Aussage: Je komplexer⁹⁰ ein Modell ist, desto besser erklärt es zukünftige Daten?

- ▶ Ja.
- ▶ Nein.

⁹⁰Hier: Grad des Polynoms.

[...] In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality – a point well made by George Box in his oft-cited remark that “all models are wrong, but some are useful” [...]⁹¹

⁹¹Hand, D. J. (2014). Wonderful Examples, but Let's not Close Our Eyes. Statistical Science 29(1), 98-100 <https://projecteuclid.org/euclid.ss/1399645735>

Offene Übung 126: Rechnungshöhe

Modellieren Sie die Rechnungshöhe als Funktion der Anzahl Personen sowie der Tageszeit.

11 Datenhandling

Häufig müssen Daten vor der eigentlichen Analyse vorverarbeitet werden, z. B.:

- ▶ Variablen auswählen: `select()`
- ▶ Beobachtungen auswählen: `filter()`
- ▶ Variablen verändern, neu erzeugen: `mutate()`
- ▶ Beobachtungen zusammenfassen: `summarise()`
- ▶ ...

Das Paket `dplyr`⁹² bietet dazu viele Möglichkeiten.

Umfangreiche Dokumentation: <http://dplyr.tidyverse.org/index.html>

⁹²wird mit `mosaic` installiert und geladen.

Einlesen der *Tipping*⁹³ Daten sowie laden des Pakets *mosaic*.

```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
tips <- read.csv("tips.csv")
# Alternativ - heruntergeladene Datei einlesen:
# tips <- read.csv2(file.choose())

library(mosaic) # Paket laden
```

⁹³Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

Variablen auswählen: select()

```
tips %>%
  select(sex, total_bill) %>%
  inspect()

## 
## categorical variables:
##   name  class levels  n missing
## 1 sex    factor     2 244      0
##                                         distribution
## 1 Male (64.3%), Female (35.7%)
## 

## quantitative variables:
##       name  class min     Q1 median      Q3 max     mean      sd
## 1 total_bill numeric 3.07 13.3475 17.795 24.1275 50.81 19.78594 8.902412
##       n missing
## 1 244      0
```

- ▶ Logisches Und (\wedge): `&`: Eine *und* Verknüpfung von zwei Aussagen ist genau dann wahr, wenn beide Aussagen wahr sind.
- ▶ Logisches Oder (\vee): `|`: Eine *oder* Verknüpfung von zwei Aussagen ist genau dann wahr, wenn mindestens eine Aussage wahr ist.
- ▶ Logische Verneinung (\neg): `!`
- ▶ Wahr: TRUE, Falsch: FALSE

Dabei wird vektorelementweise verglichen. Zusammenfassung durch Klammern.⁹⁴

```
x <- c(TRUE, TRUE)
y <- c(TRUE, FALSE)
x & y
```

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

```
x | y
```

```
## [1] TRUE TRUE
```

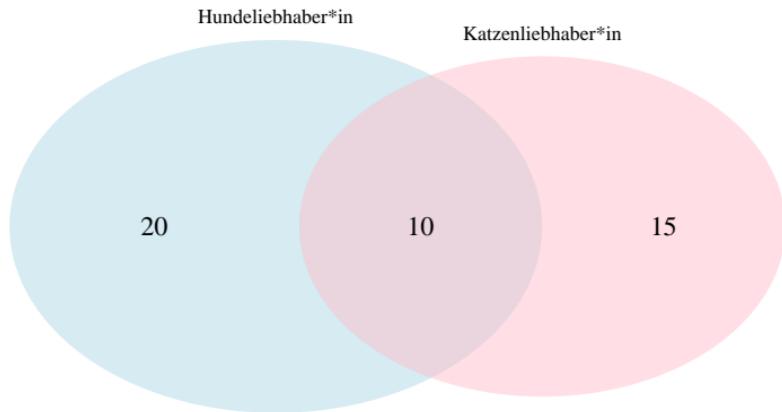
```
x | (!y)
```

```
## [1] TRUE TRUE
```

⁹⁴Über `all()` (\forall) und `any()` (\exists) lassen sich Wahrheitsvektoren zusammenfassen.

Was ergibt: (TRUE | FALSE) & (FALSE)

- A. FALSE
- B. TRUE



Stimmt die Aussage: Wenn Beobachtungen anhand einer *Und* (\wedge) Verknüpfung ausgewählt werden, so werden mindestens so viele Beobachtungen wie bei einer *Oder* (\vee) Verknüpfung ausgewählt?

- ▶ Ja.
- ▶ Nein.

- ▶ gleich, ($=$): `==`
- ▶ ungleich (\neq): `!=`
- ▶ kleiner, kleiner gleich ($<$, \leq): `<`, `<=`
- ▶ größer, größer gleich ($>$, \geq): `>`, `>=`

```
4 == 5
```

```
## [1] FALSE
```

```
4 != 5
```

```
## [1] TRUE
```

```
4 <= 5
```

```
## [1] TRUE
```

```
4 > 5
```

```
## [1] FALSE
```

Beobachtungen auswählen: filter()

```
tips %>%
  filter(sex=="Female" & total_bill>20) %>%
  inspect()

## 
## categorical variables:
##   name   class levels n missing
## 1 sex   factor     2 27      0
## 2 smoker factor     2 27      0
## 3 day   factor     4 27      0
## 4 time  factor     2 27      0
##                                     distribution
## 1 Female (100%), Male (0%)
## 2 No (63%), Yes (37%)
## 3 Sat (48.1%), Sun (25.9%), Thur (22.2%) ...
## 4 Dinner (77.8%), Lunch (22.2%)
##
## quantitative variables:
##   name   class   min    Q1 median    Q3 max     mean      sd
## 1 total_bill numeric 20.27 22.585  26.41 29.995 44.3 27.726667 6.594811
## 2 tip   numeric  1.50  2.900   3.61  5.000  6.5  3.800370 1.191278
## 3 size  integer  2.00  2.000   3.00  4.000  6.0  3.259259 1.227649
##   n missing
## 1 27      0
## 2 27      0
## 3 27      0
```

Offene Übung 129: Datensatz auswählen

Erzeugen Sie einen Datensatz, der nur die Variable tip enthält, und zwar für die Tische, an denen zum Dinner geraucht wurde.

Variablen verändern/erzeugen: mutate()

```
tips %>%  
  mutate(paid=total_bill+tip) %>%  
  select(paid) %>%  
  inspect()
```

```
##  
## quantitative variables:  
##   name    class   min    Q1 median      Q3    max    mean      sd     n  
## 1 paid numeric 4.07 15.475   20.6 27.7225 60.81 22.78422 9.890116 244  
##   missing  
## 1       0
```

Wie viele Beobachtungen haben eine relative Trinkgeldhöhe $\frac{tip}{total_bill}$ über 10%?

- A. 244
- B. 217
- C. 27

Variablen verändern: case_when()

```
tips %>%
  mutate(bill = case_when(total_bill <= 10 ~ "low",
                         total_bill <= 20 ~ "middle",
                         total_bill > 20 ~ "high")) %>%
  select(bill) %>%
  table()

## .
##   high     low middle
##     97      17    130
```

Hinweis: Anstelle der letzten Abfrage (`total_bill > 20`) hätte auch einfach `TRUE` verwendet werden können.

Übung 131: Variablen erzeugen

Welcher Befehl ist richtig, wenn die Personen die Raucher und Dinner sind eine Gruppe sein sollen, alle anderen eine andere?

A.

```
tips %>%
  mutate(party = case_when((smoker=="Yes" & time=="Dinner")
                            ~ "Party",
                           TRUE ~ "No Party"))
```

B.

```
tips %>%
  mutate(party = case_when((smoker=="Yes" | time=="Dinner")
                            ~ "No Party",
                           TRUE ~ "Party"))
```

Daten zusammenfassen: summarise()

```
tips %>%
  summarise(mean_bill=mean(total_bill), n=n())
##   mean_bill    n
## 1  19.78594 244
```

Nach Gruppen zusammenfassen: group_by()

```
tips %>%  
  group_by(sex, time) %>%  
  summarise(mean_bill=mean(total_bill), n=n())  
  
## # A tibble: 4 x 4  
## # Groups: sex [?]  
##   sex     time   mean_bill     n  
##   <fct>   <fct>     <dbl> <int>  
## 1 Female  Dinner      19.2     52  
## 2 Female  Lunch       16.3     35  
## 3 Male    Dinner      21.5    124  
## 4 Male    Lunch       18.0     33
```

Mit welchem Befehl können Beobachtungen mit bestimmten Eigenschaften ausgewählt werden?

- A. `select()`
- B. `filter()`
- C. `mutate()`
- D. `summarise()`

Die ersten n Beobachtungen: top_n()

```
tips %>%
  group_by(sex) %>%
  top_n(n=3, tip)

## # A tibble: 6 x 7
## # Groups: sex [2]
##   total_bill  tip sex   smoker day    time   size
##       <dbl> <dbl> <fct> <fct>  <fct> <fct> <int>
## 1     39.4  7.58 Male   No     Sat    Dinner     4
## 2     34.8  5.20 Female No     Sun    Dinner     4
## 3     34.8  5.17 Female No    Thur  Lunch      4
## 4     50.8 10.0  Male   Yes    Sat    Dinner     3
## 5     48.3  9.00 Male   No    Sat    Dinner     4
## 6     28.2  6.50 Female Yes   Sat    Dinner     3
```

Hinweis: Auf diese Art und Weise können auch Datensätze balanciert werden.⁹⁵

⁹⁵Vgl. geschichtete Stichprobe: group_by() %>% sample_n()

Beobachtungen sortieren: arrange()

```
tips %>%
  group_by(sex) %>%
  top_n(n=3, tip) %>%
  arrange(sex)

## # A tibble: 6 x 7
## # Groups: sex [2]
##   total_bill  tip sex   smoker day    time   size
##       <dbl> <dbl> <fct> <fct> <fct> <fct> <int>
## 1     34.8  5.20 Female No    Sun Dinner     4
## 2     34.8  5.17 Female No    Thur Lunch      4
## 3     28.2  6.50 Female Yes   Sat Dinner     3
## 4     39.4  7.58 Male   No    Sat Dinner     4
## 5     50.8 10.0  Male   Yes   Sat Dinner     3
## 6     48.3  9.00 Male   No    Sat Dinner     4
```

```
# ID (Zeilennummer, Schlüssel) erzeugen
tipsID <- tips %>%
  mutate(ID=row_number())

# Zwei (Teil-)Datensätze erzeugen
tips1 <- tipsID %>%
  select(ID, total_bill)
tips2 <- tipsID %>%
  select(ID, tip)

# Innere Verknüpfung
tips1 %>%
  inner_join(tips2, by = "ID") %>%
  inspect()

## 
## quantitative variables:
##      name   class   min    Q1   median     Q3    max    mean
## 1      ID integer 1.00 61.7500 122.500 183.2500 244.00 122.500000
## 2 total_bill numeric 3.07 13.3475 17.795 24.1275 50.81 19.785943
## 3      tip numeric 1.00  2.0000  2.900  3.5625 10.00  2.998279
##      sd    n missing
## 1 70.580923 244      0
## 2  8.902412 244      0
## 3  1.383638 244      0
```

Berechnen Sie den Mittelwert und die Standardabweichung der relativen Trinkgeldhöhe, je nachdem ob es sich um eine “Party” oder nicht gehandelt hat.

Tidy data:⁹⁶

- ▶ Jede Variable ist eine Spalte.
- ▶ Jede Beobachtung ist eine Zeile.
- ▶ Jeder Wert ist eine Zelle.

Herausforderungen:

- ▶ *Breiter* Datensatz: Eine Variable über mehrere Spalten.
- ▶ *Langer* Datensatz: Eine Beobachtung über mehrere Zeilen.

```
# Ggf. einmalig installieren
install.packages("tidyverse")
# Paket laden
library(tidyverse)
```

⁹⁶Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.

Langer Datensatz über gather()

Überführt einen *breiten* Datensatz in einen *langen*:

```
tips_long <- tips %>%  
  mutate(id = row_number()) %>%  
  select(id, tip, total_bill) %>%  
  gather(key = "Variable", value = "Wert", -id)  
  
str(tips_long)
```

```
## 'data.frame':    488 obs. of  3 variables:  
##   $ id      : int  1 2 3 4 5 6 7 8 9 10 ...  
##   $ Variable: chr  "tip" "tip" "tip" "tip" ...  
##   $ Wert     : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
```

Breiter Datensatz über spread()

Überführt einen *langen* Datensatz in einen *breiten*:

```
tips_wide <- tips_long %>%  
  spread(key = "Variable", value="Wert")
```

```
str(tips_wide)
```

```
## 'data.frame':    244 obs. of  3 variables:  
##   $ id        : int  1 2 3 4 5 6 7 8 9 10 ...  
##   $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...  
##   $ total_bill: num  17 10.3 21 23.7 24.6 ...
```

Welcher Datensatz hat mehr Beobachtungen?

- A. Ein langer.
- B. Ein breiter.
- C. Beide gleich.

12 Organisatorisches

12. Organisatorisches Literatur (Auswahl)

- ▶ David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel (2014): *Introductory Statistics with Randomization and Simulation*,
https://www.openintro.org/stat/textbook.php?stat_book=isrs
- ▶ Nicholas J. Horton, Randall Pruim, Daniel T. Kaplan (2015): Project MOSAIC Little Books *A Student's Guide to R*, <https://github.com/ProjectMOSAIC/LittleBooks/raw/master/StudentGuide/MOSAIC-StudentGuide.pdf>
- ▶ Chester Ismay, Albert Y. Kim (2017): *ModernDive – An Introduction to Statistical and Data Sciences via R*, <http://moderndive.com/>
- ▶ Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning – with Applications in R*,
[<http://www-bcf.usc.edu/~gareth/ISL/>] (<http://www-bcf.usc.edu/~gareth/ISL/>)

12. Organisatorisches Lizenz / Version

Diese Folien wurden von Karsten Lübke zusammen mit Kolleg*innen von der FOM
<https://www.fom.de/> entwickelt und stehen unter der Lizenz CC-BY-SA-NC 3.0 de:
<https://creativecommons.org/licenses/by-nc-sa/3.0/de/>

Der verwendete Code sowie das Beamer Template aus dem [NPBT-Projekt](#) von Norman Markgraf stehen unter der Lizenz [GNU General Public License v3.0](#).

- ▶ Datum erstellt: 2018-02-09
- ▶ R Version: 3.4.3
- ▶ mosaic Version: 1.1.1

Bitte melden Sie Fehler und Verbesserungsvorschläge: karsten.luebke@fom.de

Mitarbeit und Hinweise von Thomas Christiaans, Oliver Gansser, Matthias Gehrke, Jörg Horst, Bianca Krol, Norman Markgraf, Sebastian Sauer, Daniel Ziggel. **Vielen Dank!**

Die Studierenden können nach erfolgreichem Abschluss des Moduls:

- ▶ den Prozess der Erkenntnisgewinnung in der Psychologie nutzen,
- ▶ Daten gewinnen, zusammenfassen, analysieren und graphisch darstellen,
- ▶ statistische Aussagen über Zusammenhänge und Prognosen machen,
- ▶ die Gültigkeit der gefundenen Schlussfolgerungen abschätzen,
- ▶ wichtige Methoden der deskriptiven und Inferenzstatistik passgenau auswählen und anwenden,
- ▶ Auswertungen mit R durchführen,
- ▶ die Anwendung statistischer Auswertungen in Fachveröffentlichungen verstehen und einordnen,
- ▶ in der Struktur psychologischer Veröffentlichungen ihre Ergebnisse berichten,
- ▶ vorbereitend für die Projektarbeiten und ihre Abschlussarbeit angemessene empirische Methoden einsetzen.

- ▶ Datenanalyse (ca. 1500 Wörter)
- ▶ Klausur 90 Minuten

Seminararbeit und Klausur gehen jeweils zu 50 % in die Modulnote ein, beide Prüfungsleistungen müssen mit mindestens 4,0 bewertet werden.

Beachten Sie die im OC hinterlegten Fristen!

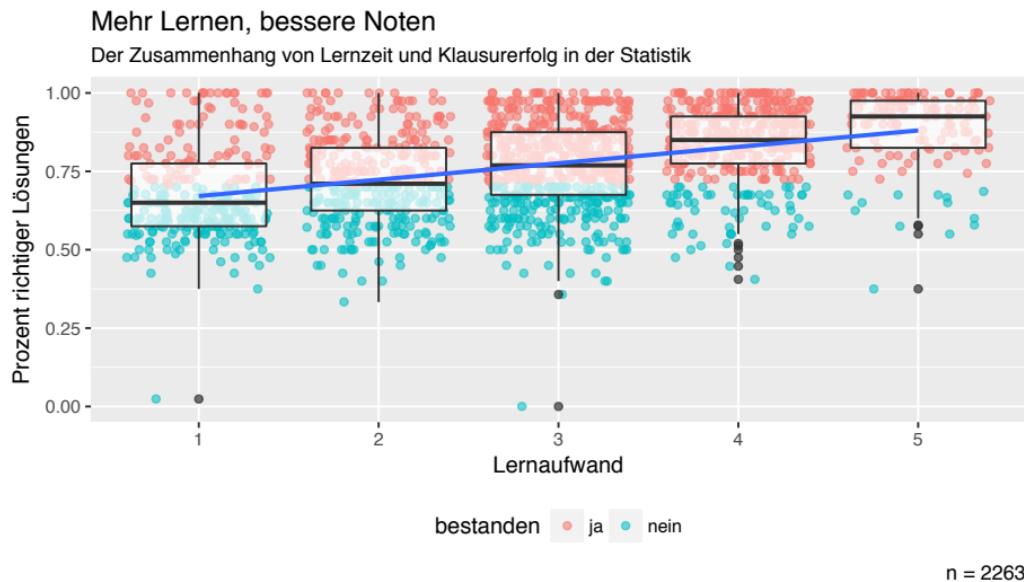
Workload:

- ▶ Präsenzstunden: 60,0 UE
- ▶ Strukturiertes Eigenstudium 130,00 ZStd
- ▶ Workload gesamt: 175,0 ZStd
- ▶ ECTS-Credit Punkte: 7

- ▶ Prüfungsrelevant ist der gesamte Stoff der Vorlesung. Für die Wiederholungsprüfung gelten die gleichen Rahmendbedingungen wie für den regulären Prüfungstermin.
- ▶ Lesen Sie sich erst die ganze Klausur in Ruhe durch und fangen Sie mit den Aufgaben an, die Sie sicher können.
- ▶ Halten Sie sich nicht zu lange mit Aufgaben auf, die wenig Punkte bringen.

Viel Erfolg!

12. Organisatorisches Des Lehrenden letzte Worte



Siehe Blogbeitrag [Sebastian Sauer \(20.12.2017\)](#) "Zusammenhang von Lernen und Noten im Statistikunterricht"