Statistik1

Sebastian Sauer

01.02.23

Inhaltsverzeichnis

Hinweise **3**
Lernziele . . . 3
Was lerne ich hier und wozu ist das gut? . . . 3
Motivieren Sie mich! . . . 4
Voraussetzungen . . . 4
Überblick . . . 4
Software . . . 4
PDF-Version . . . 5
Lernhilfen . . . 5
Videos . . . 5
Online-Zusammenarbeit . . . 6
Selbstlernkontrolle . . . 6
Forum . . . 6
Modulzeitplan . . . 6
Literatur . . . 7
Organisatorische Hinweise . . . 7
Sicherheitsunterweisung . . . 7
Organisatorisches . . . 7
Präsenzunterricht . . . 8
Rechtliche Hinweise . . . 8
Datenschutz . . . 8
IT . . . 9
Ich brauche Hilfe. Was soll ich tun? . . . 10
Videokonferenzen . . . 10
Streaming . . . 11
Umgangsformen . . . 11
Wenn der Unterricht ausfällt . . . 12
Salvatorische Klausel . . . 12
Technische Details . . . 12

1 Fragen stellen **14**
1.1 Lernsteuerung . . . 14
1.1.1 Standort im Lernpfad . . . 14
1.1.2 Lernziele . . . 14
1.2 Was ist Statistik und wozu ist sie gut? . . . 15
1.3 Was ist das Ziel Ihrer Analyse? . . . 16
1.3.1 Arten von Zielen . . . 16
1.3.2 Forschungsfrage . . . 17

1.4 Was sind Daten? 18
1.4.1 Was ist eine Variable? 18
1.4.2 Beobachtungseinheit 19
1.4.3 Wert 19
1.4.4 Tidy-Data 20
1.4.5 Arten von Variablen 21
1.5 Beispiele für Skalenniveaus 23
1.6 Modelle 25
1.7 Praxisbezug 27
1.8 Fazit 28
1.9 Aufgaben 28
1.10 Vertiefung 28
1.11 Literatur 28

Hinweise

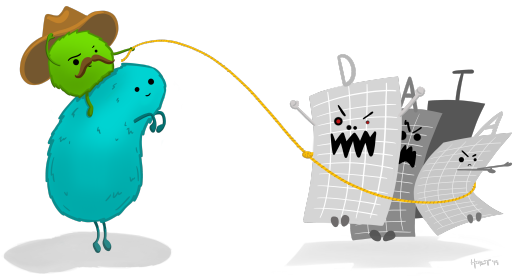

Abbildung 1: Daten zähmen

Bildquelle: Allison Horst, CC-BY

Lernziele

- Die Studentis sind mit wesentlichen Methoden der explorativen Datenanalyse vertraut und können diese selbständig anwenden.
- Die Studentis können gängige Forschungsfragen in lineare Modelle übersetzen, diese auf echte Datensätze anwenden und die Ergebnisse interpretieren.

Was lerne ich hier und wozu ist das gut?

Warum ist das wichtig?

Wir wollen nicht auf Leuten vertrauen, die behaupten, sie wüssten, was für uns richtig und gut ist. Wir wollen selber die Fakten prüfen können.

Wozu brauche ich das im Job?

Datenanalyse spielt bereits heute in vielen Berufen eine Rolle. Tendenz stark zunehmend.

Wozu brauche ich das im weiterem Studium?

In Forschungsarbeiten (wie in empirischen Forschungsprojekten, etwa in der Abschlussarbeit) ist es üblich, statistische Ergebnisse hinsichtlich quantitativ zu analysieren.

Gibt es auch gute Jobs, wenn man sich mit Daten auskennt?

Das World Economic Forum (2020) berichtet zu den "Top 20 job roles in increasing and decreasing demand across industries" (S. 30, Abb. 22):

1. Data Analysts und Scientists
2. AI and Machine Learning Specialists
3. Big Data Specialists

Motivieren Sie mich!

Ansprache zur Motivation

Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie Folgendes mitbringen:

- Bereitschaft, Neues zu lernen
- Bereitschaft, nicht gleich aufzugeben
- Kenntnis grundlegender Methoden wissenschaftlichen Arbeitens

Überblick

Abb. Abbildung 2 gibt einen Überblick über den Verlauf und die Inhalte des Buches.

Software

Installieren Sie R und seine Freunde. Für die Bayes-Inferenz brauchen Sie[1] zusätzliche Software, was leider etwas Zusatzaufwand erfordert. Lesen Sie hier die Hinweise dazu. Installieren Sie die folgende R-Pakete[2]:

- tidyverse
- easystats
- weitere Pakete werden im Unterricht bekannt gegeben (es schadet aber nichts, jetzt schon Pakete nach eigenem Ermessen zu installieren)

R Syntax aus dem Unterricht findet sich im Github-Repo bzw. Ordner zum jeweiligen Semester.

[1] nicht gleich zu Beginn, aber nach 2-3 Wochen
[2] falls Sie die Pakete schon installiert haben, könnten Sie mal in RStudio auf "update.packages" klicken

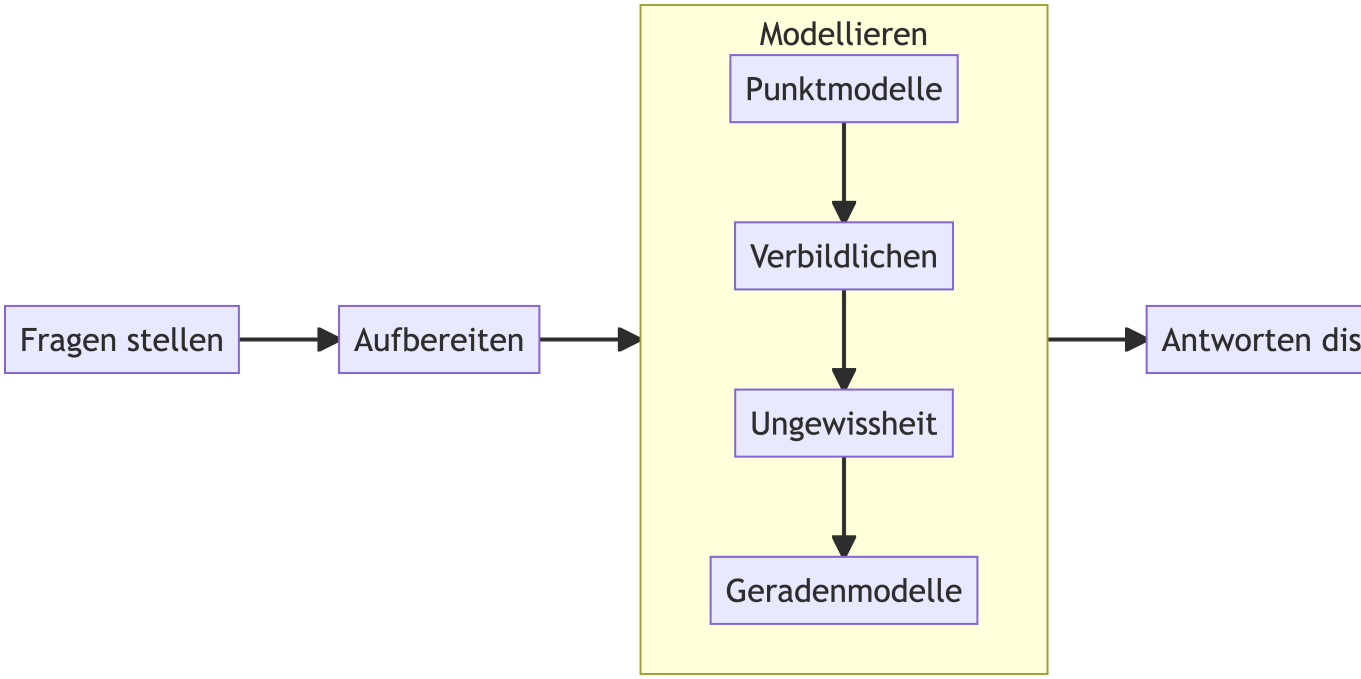

Abbildung 2: Überblick über den Inhalt und Verlauf des Buches

PDF-Version

– EXPERIMENTAL – EXPERIMENTAL

Von diesem "Webbuch" (HTML-Format) gibt es hier eine PDF-Version. Die PDF-Version eignet sich zum Ausdrucken und zur Offline-Nutzung.

Allerdings wurden die Inhalte *in erster Linie für ein Webbuch-Format* formatiert, die PDF-Ausgabe ist daher nicht ideal. Es ist empfehlenswert, mit der Webbuch-Version zu arbeiten. Außerdem wird die PDF-Version nicht ganz aktuell gehalten - die aktuelle Version ist immer die Webbuch-Variante. Prüfen Sie im Zweifel das Datum der Erstellung des Dokuments.

Lernhilfen

Videos

Auf dem YouTube-Kanal des Autors finden sich eine Reihe von Videos mit Bezug zum Inhalt dieses Buchs. Besonders diese Playlist passt zu den Inhalten dieses Buchs.

Online-Zusammenarbeit

Hier finden Sie einige Werkzeuge, die das Online-Zusammenarbeiten vereinfachen:

- Frag-Jetzt-Raum zum anonymen Fragen stellen während des Unterrichts. Der Keycode wird Ihnen bei Bedarf vom Dozenten bereitgestellt.
- Padlet zum einfachen (und anonymen) Hochladen von Arbeitsergebnissen der Studentis im Unterricht. Wir nutzen es als eine Art Pinwand zum Sammeln von Arbeitsbeiträgen. Die Zugangsdaten stellt Ihnen der Dozent bereit.

Selbstlernkontrolle

Für jedes Kapitel sind (am Kapitelende) Aufgaben eingestellt, jeweils mit Lösung. Ein Teil dieser Aufgaben hat eine kurze, eindeutige Lösung (z.B. "42" oder "Antwort C"); ein (kleiner) Teil der Aufgaben verlangen komplexere Antworten (z.B. "Welche Arten von Prioris gibt es bei `stan_glm()`?). Nutzen Sie die Fragen mit eindeutiger, kurzer Lösung um sich selber zu prüfen. Nutzen Sie die Fragen mit komplexerer, längerer Lösung, um ein Themengebiet tiefer zu erarbeiten.

> **Hinweis**
>
> Fortwährendes Feedback zu Ihrem Lernfortschritt ist wichtig, damit Sie Ihre Lernbemühungen steuern können. Bearbeiten Sie daher die bereitgestellten Arbeiten ernsthaft.

Forum

Nutzen Sie das vom Dozenten bereitgestelle Forum, um Fragen zu stellen und Fragen zu beantworten.

Modulzeitplan

Nr	Thema	Datum	Kommentar
1	Fragen stellen	13.3. - 19.3.	Lehrbeginn ist am Mi., 15.3.23
2	Daten aufbereiten	20.3. - 26.3.	NA
3	Daten aufbereiten	27.3. - 2.4.	NA
4	Daten verbildlichen	3.4. - 9.4	Karwoche (kein Unterricht am Do. und Fr.)
5	Daten verbildlichen	10.4. - 16.4.	Osterwoche (kein Unterricht am Mo. und Di.)
6	Punktmodelle	17.4. - 23.4.	NA
7	Punktmodelle	24.4. - 30.4.	NA
8	Ungewissheit quantifizieren	1.5. - 7.5.	Maifeiertag (kein Unterricht am Mo.)
9	Ungewissheit quantifizieren	8.5. - 14.5.	NA

10	Fallstudien	15.5. - 21.5.	NA
11	-	22.5. - 28.5.	Blockwocke - kein regulärer Unterricht
12	Geradenmodell	29.6. - 4.6.	Pfingstwoche (kein Unterricht am Mo. und Di.)
13	Geradenmodell	5.6. - 11.6.	Fronleichnam (kein Unterricht am Do. und Fr.)
14	Antworten diskutieren	12.6. - 18.6.	NA
15	Fallstudien	19.6. - 25.6.	NA
16	Abschluss	26.6. - 2.7.	Letzter Lehrtag ist Fr., 30.6.

Literatur

Pro Thema wird Literatur ausgewiesen.

Organisatorische Hinweise

Sicherheitsunterweisung

- Der Dozent weist auf Fluchtwege und auf die Sammelstelle hin (abhängig vom Seminarraum).
- Flucht- und Rettungspläne hängen im Gebäude aus; bitte betrachten Sie sie sorgfältig. Prägen Sie sich die Rettungswege, den Ort der Sammelstelle sowie von Feuerlöschern und Brandmeldeeinrichtungen gut ein.
- Bei Alarm: Ruhe bewahren. Der Dozent weist an, den Raum sofort zu verlassen und geordnet über den kürzesten Weg zur Sammelstelle zu gehen. Tun Sie das.
- Prüfen Sie, ob niemand zurückgelassen ist (z. B. auf der Toilette).
- Fluchtwege (z. B. Fluren und Treppen) dürfen nicht versperrt sein.
- Notausgänge müssen ausgeschildert und frei zugänglich sein.
- Aufzüge dürfen im Brandfall nicht verwendet werden. Lebensgefahr!
- Im Brandfall sind Fenster und Türen zu schließen (aber nicht zu verriegeln!).
- Hilflose Personen sind mitzunehmen.
- Brände sind zu bekämpfen, aber die sichere Evakuierung hat Vorrang.
- Bei Bränden ist sofort die Feuerwehr zu alarmieren.
- Bei einem Hupton ist das Gebäude sofort zu verlassen und die Sammelstelle aufzusuchen. Brandschutztüren sind stets geschlossen zu halten.
- Prägen Sie sich die Position von Feuerlöschern und Erste-Hilfe-Material gut ein.

Organisatorisches

- Bitte laden Sie sich rechtzeitig die Materialien herunter.
- Beachten Sie die Vorbereitungshinweise zur ersten Unterrichtsstunde und für die einzelnen Termine.

- Ein Foliensatz kann kein Lehrbuch ersetzen; falls Sie bei einem Termin gefehlt haben oder Ihre Aufzeichnung lückenhaft ist, lesen Sie bitte in der Literatur nach oder bitten Sie eine/n Kommiliton/in um Hilfe.
- Beachten Sie die Hinweise der Hochschule (s. Moodle) wie die Orientierungshilfe, die Klausur- und Studiengangsordnungen. Dort finden Sie verbindliche Hinweise zu vielen organisatorischen Fragen.
- Bitte prüfen Sie jetzt schon und in regelmäßigen Abständen die Modulseite auf neue Materialien.
- Nachdem eine Unterrichtseinheit abgeschlossen ist, ändert der Dozent grundsätzlich nichts mehr an den Materialien (in Ausnahmefällen wie etwa der Korrektur eines Fehler informiert der Dozent schriftlich).
- Es können sich Aktualisierungen des Unterrichtsmaterials ergeben. Bitte prüfen Sie regelmäßig, ob neues Material eingestellt ist. Nachdem der Unterricht stattgefunden hat, wird das Material aber nicht mehr geändert, so dass Sie Sicherheit für das Lernen haben.

Präsenzunterricht

- Beachten Sie aktuell geltenden Hygienevorschriften.
- Bitte meiden Sie datenhungrige Applikationen wie das Streamen von Filmen – Sie behindern den Unterricht und verärgern damit andere Studierende (in diesem oder anderen Kursen).
- Bitte bringen Sie einen Computer (mit Internetanschluss) in alle Präsenztermine mit. Bringen Sie ggf. Ersatzbatterien und/oder Verlängerungskabel mit - leider sind nicht überall ausreichend Steckdosen vorhanden.

Rechtliche Hinweise

- Dieser Kurs ist lizensiert unter der MIT Lizenz. Das ist eine permissive Lizenz, die erlaubt, dass Sie diesen Kurs frei verwenden können. Sie haben (nur) die Verpflichtung, zu zitieren und auf die Lizenzart hinzuweisen.
- Mitarbeit oder Verbesserungsvorschläge: am besten als Github Issue auf der Github-Seite dieses Projekts einstellen.
- Für die Inhalte von Links kann keine Haftung übernommen werden.
- Aus Datenschutzgründen dürfen Teilnehmis den Unterricht (Video, Ton, sonstige) nicht aufnehmen.

Datenschutz

- Einige Teile des Unterrichts werden u.U. aufgenommen.
- Die Aufnahme erfolgt nur nach vorheriger Information durch den Dozenten.

- Bei den Bild-Aufnahmen wird nur der Bildschirm des Dozenten aufgenommen; außerdem wird der Ton aufgenommen.
- Bitte verzichten Sie daher während der Aufnahme auf Zoom-Reaktionen wie Hand heben, da diese u.U. mit Ihrem Namen auf dem Bildschirm des Dozenten und damit auf der Aufnahme zu sehen sind.
- Während einer Aufnahme dürfen aus Datenschutzgründen keine Wortbeiträge zu hören sein, die Personen identifizieren.
- Bitte verzichten Sie während der Aufnahme daher auf Wortbeiträge. - Der Dozent informiert umgehend, wenn er die Aufnahme beendet.
- Die Aufnahmen werden im Internet (z.B. YouTube) frei veröffentlicht. Solche offenen Beiträge im Internet können u.U. in Zukunft nicht kontrolliert werden, insbesondere können sie unkontrolliert verbreitet und kommentiert werden.
- Der Dozent hat u.U. keinen Einfluss auf Verbreitung, Kommentare oder sonstige Weiterverwendung.
- Alle Rechte an den Aufnahmen liegen beim Dozenten.
- Aufnahmen des Unterrichts durch die Studentis sind nicht erlaubt.

IT

- Sie benötigen einen Computer für diesen Kurs. Ein Laptop ist ideal (ein günstiges Modell ist vollkommen ausreichend); ein Tablett ist nicht ideal, reicht aber zur Not, wenn Sie eine Tastatur für das Gerät haben.
- Bereiten Sie bitte R und seine Freunden vor vor der ersten Unterrichtsstunde. Sie können entweder R auf Ihrem Computer installieren oder den Cloud-Dienst *RStudio Cloud* nutzen.
- Bitte beachten Sie die Installationshinweise für Software und stellen Sie sicher, dass die Software vor Beginn des ersten Termins einsatzbereit ist.
- Sorgen Sie dafür, dass Sie in jeder Stunde über eine gute Internetverbindung verfügen.
- Bitte stellen Sie sicher, dass Sie Zugriff auf Moodle, Zoom und weitere Dienste haben, die wir im Unterricht nutzen (Passwörter dabei haben etc.).
- Kontaktieren Sie die IT-Abteilung bei technischen Probleme.
- Für Präsenztermine: Bitte bringen Sie folgende IT-Geräte in jede Stunde mit: Laptop mit Stromkabel, Smartphone oder Tablett, Kopfhörer inkl. Mikrofon.
- Für Präsenztermine: Ein langer Uni-Tag zehrt an der Batterie; nicht nur an Ihrer, sondern auch in Ihrem Laptop und Smartphone. Ggf. kann eine Zusatzbatterie (Akku-Pack) hilfreich sein. Bei schlechter WLAN-Verbindung kann ein Hotspot über das Handy eine Lösung sein.
- Bitte lesen Sie auch die Hinweise zur Software, die wir in diesem Modul benötigen.
- Stellen Sie sicher, dass Sie die jeweils aktuelle Version für die in diesem Kurs verwendete Software verwenden (z.B. bei Zoom gibt es viele Aktualisierungen).
- Updaten Sie ggf. R, RStudio und Ihre R-Pakete.

Ich brauche Hilfe. Was soll ich tun?

- Versuchen Sie erst selber, das Problem zu lösen. Manchmal hilft etwas Abstand (Pause, drüber schlafen), um ein Problem klarer zu sehen (oder es als nicht mehr so wichtig zu sehen).
- Fragen Sie Kommilitonis.
- Recherchen Sie (online) nach einer Lösung.
- Posten Sie das Problem im Forum oder einem Online-Forum. Wichtig: Stellen Sie Ihrem Post ein Erbie bei.
- Besuchen Sie die Sprechstunde des Dozenten.
- Schreiben Sie möglichst keine Email, sondern posten Sie Ihre Frage im Forum o.Ä., da die meisten Fragen für Ihre Kommilitonis auch nützlich sein können (eine Ausnahme sind natürlich individuelle, persönliche Angelegenheiten, die nicht den Stoff betreffen).

Videokonferenzen

- Ihr Dozent informiert Sie, ob Videokonferenzen in diesem Modul angeboten werden.
- Für Videokonferenzen/Webinare in diesem Modul wird die Software Zoom verwendet. Zur Einwahl benötigen Sie eine Zoom-Meeting-ID bzw. die dazu gehörige URL (Link); außerdem benötigen Sie ggf. ein zugehöriges Passwort. Diese Informationen werden Ihnen vom Dozenten zugestellt-.
- Für die Teilnahme an der Videokonferenz können Sie einen internetfähigen Rechner oder Tablet/Smartphone verwenden. Alternativ können Sie den Zoom-Client auf Ihrem Computer installieren (die Installation auf dem Rechner ist zu empfehlen).
- Ein paar Minuten vor Unterrichtsbeginn da sein, hilft beim Ankommen und um etwaige technische Probleme auszuräumen.
- -Wenn jede(r) Hintergrundgeräusche vermeidet (auch die Tastatur kann ziemlich laute Geräusche verursachen), verstehen wir uns alle gut.
- Das Mikro auszuschalten, wenn man gerade nichts sagen möchte, gibt Freiheit, doch nicht ganz leise zu sein, ohne die anderen zu stören. - Ein Headset ist hilfreich, um Rückkopplungsgeräusche (Echo) zu vermeiden.
- Eine einigermaßen schnelle Internetverbindung ist nötig.
- Ich empfehle, eine Webcam anzuschließen (bzw. freigeben), so dass wir uns sehen können. Das ist der Qualität des Unterrichts zuträglich.
- Hilfe zu Zoom findet sich hier: https://support.zoom.us/hc/de.
- Zu bestimmten Seiten kann der Dienst (Zoom) überlastet sein, so dass das Webinar nur eingeschränkt oder nicht möglich ist. In diesem Fall erstellt der Dozent ein Video und stellt es im Nachgang über Moodle zur Verfügung.
- Die Datenschutzhinweise von Zoom finden Sie hier.
- Prüfen Sie regelmäßig, ob Ihr Zoom-Client aktuell ist (sonst updaten).
- Aus Gründen der Sicherheit der Teilnehmis kann der Dozent anonyme Teilnahme an Videokonferenzen untersagen.
- Der Chat sollte nur für Themen des Unterrichts verwendet werden (nicht für private Themen).

- Aus Sicherheitsgründen kann eine Zoom-Konferenz auf Ihre “offizielle” E-Mail-Adresse (der Hochschule) begrenzt sein. Mit einer anderen, potenziell anonymen E-Mail-Adresse, können Sie also u.U. nicht teilnehmen.

Streaming

In einigen Kursen wird Präsenzunterricht plus Streaming angeboten. In diesem Fall gelten folgende Hinweise:

- Wenn die Teilnehmerzahl am Kurs geringer ist als die Raumkapazität (z.B. weniger als 80 Teilnehmis), so haben Sie stets freie Wahl, ob Sie in Präsenz beiwohnen oder den Unterricht online verfolgen.
- Übersteigt die Teilnehmerzahl am Kurs die Raumkapazität bis zum Doppelten (z.B. 160 Teilnehmis bei 80 Plätzen), so findet der Präsenzunterricht zweizügig statt. In der 1. Woche können alle Studentis des Zugs A zum Präsenunterricht kommen; in Woche 2 können alle Studentis des Zugs B zum Präsenzunterricht kommen. Wer nicht für Präsenz eingeteilt ist, kann dem Unterricht online folgen. Auch wer in Präsenz eingeteilt ist, kann dem Unterricht jederzeit online folgen. Auch wer nicht für Präsenz eingeteilt ist, kann zum Präsenzunterricht kommen, FALLS es freie Plätze gibt: Kommen Sie einfach zum Hörsaal, wenn es freie Plätze gibt, sind Sie herzlich willkommen (es gilt: first come, first serve).
- Sie müssen sich nicht anmelden für Präsenz oder Online.
- Falls die Teilnehmerzahl größer ist als das Doppelte oder es Probleme mit dem Prozedere gibt, erfolgt eine neue Planung.
- Ihr Dozent informiert Sie zu den Details, sobald die Teilnehmerzahl (grob) feststeht.
- Alle Angaben sind vorbehaltlich neuer Corona-Entwicklungen bzw. der Vorgaben der Hochschulleitung dazu.

Umgangsformen

- Freundlicher Umgang miteinander ist selbstverständlich.
- Selbstverständlich sind auch die grundlegenden Formen der höflichen Miteinanders (Begrüßung, Ansprache mit Namen, bitte/danke sagen, Pünktlichkeit, Verbindlichkeit, …).
- Bei Nichtwissen oder einem inhaltlichen Fehler wird niemand bloßgestellt oder lächerlich gemacht.
- Um eine Wortmeldung in einer Videokonferenz anzuzeigen, bietet sich die Funktion des Handhebens an.
- In Breakout-Sessions (in Zoom) sollte aktiv mitgearbeitet werden, wer das nicht möchte, betritt den Breakout-Room nicht.
- Spricht man eine Person in einer Videokonferenz an, sollte die Kamera angeschaltet sein.

Wenn der Unterricht ausfällt

- Im Krankheitsfall werden Sie möglichst frühzeitig informiert (via Moodle); das kann bedeuten, dass Sie am Morgen eine Nachricht bekommen, dass der Unterricht des jeweiligen Tages ausfällt.
- Leider muss der Unterricht in einigen Fällen aufgrund anderer Dienstverpflichtungen des Dozenten entfallen; ein Beispiel sind Berungskomissionen (Auswahlgespräche für neu einzustellende Professor:innen).
- Aufgrund gesetzlicher Feiertage sowie hochschulweit geltenden lehrfreien Tagen fallen jedes Semester einige Tage für den Unterricht aus. Bitte informieren Sie sich auf der zentralen Hochschulseite dazu.
- Falls der Unterricht ausfällt, heißt das *nicht*, dass der für die jeweilige Woche geplante Stoff obsolet ist. Viel mehr ist Ihre Pflicht, den Stoff selbständig zu erarbeiten. Natürlich können Sie bei Fragen jederzeit Ihren Dozenten kontaktieren.

Salvatorische Klausel

- Diese Hinweise hier gelten nur insofern Ihre Dozentis Ihnen nicht andere Hinweise geben :-)

Technische Details

Dieses Dokument wurde erzeugt am/um 2023-02-01 23:03:43.

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.2.1 (2022-06-23)
##  os       macOS Big Sur ... 10.16
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Berlin
##  date     2023-02-01
##  pandoc   2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (
##
## - Packages -------------------------------------------------------------------
##  package     * version date (UTC) lib source
##  assertthat    0.2.1   2019-03-21 [1] CRAN (R 4.2.0)
##  cellranger    1.1.0   2016-07-27 [1] CRAN (R 4.2.0)
##  cli           3.6.0   2023-01-09 [1] CRAN (R 4.2.0)
```

```
##  codetools     0.2-18  2020-11-04 [2] CRAN (R 4.2.1)
##  colorout    * 1.2-2   2022-06-13 [1] local
##  colorspace    2.0-3   2022-02-21 [1] CRAN (R 4.2.0)
##  DBI           1.1.3   2022-06-18 [1] CRAN (R 4.2.0)
##  digest        0.6.31  2022-12-11 [1] CRAN (R 4.2.0)
##  dplyr         1.0.10  2022-09-01 [1] CRAN (R 4.2.0)
##  evaluate      0.19    2022-12-13 [1] CRAN (R 4.2.0)
##  fansi         1.0.3   2022-03-24 [1] CRAN (R 4.2.0)
##  fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
##  generics      0.1.3   2022-07-05 [1] CRAN (R 4.2.0)
##  ggplot2       3.4.0   2022-11-04 [1] CRAN (R 4.2.0)
##  glue          1.6.2   2022-02-24 [1] CRAN (R 4.2.0)
##  gt            0.8.0   2022-11-16 [1] CRAN (R 4.2.0)
##  gtable        0.3.1   2022-09-01 [1] CRAN (R 4.2.0)
##  htmltools     0.5.4   2022-12-07 [1] CRAN (R 4.2.0)
##  jsonlite      1.8.4   2022-12-06 [1] CRAN (R 4.2.0)
##  knitr         1.41    2022-11-18 [1] CRAN (R 4.2.0)
##  lifecycle     1.0.3   2022-10-07 [1] CRAN (R 4.2.0)
##  magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
##  munsell       0.5.0   2018-06-12 [1] CRAN (R 4.2.0)
##  pillar        1.8.1   2022-08-19 [1] CRAN (R 4.2.0)
##  pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.2.0)
##  R6            2.5.1   2021-08-19 [1] CRAN (R 4.2.0)
##  readxl        1.4.1   2022-08-17 [1] CRAN (R 4.2.0)
##  rlang         1.0.6   2022-09-24 [1] CRAN (R 4.2.0)
##  rmarkdown     2.19    2022-12-15 [1] CRAN (R 4.2.0)
##  rstudioapi    0.14    2022-08-22 [1] CRAN (R 4.2.0)
##  scales        1.2.1   2022-08-20 [1] CRAN (R 4.2.0)
##  sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
##  stringi       1.7.12  2023-01-11 [1] CRAN (R 4.2.0)
##  stringr       1.5.0   2022-12-02 [1] CRAN (R 4.2.0)
##  tibble        3.1.8   2022-07-22 [1] CRAN (R 4.2.0)
##  tidyselect    1.2.0   2022-10-10 [1] CRAN (R 4.2.0)
##  utf8          1.2.2   2021-07-24 [1] CRAN (R 4.2.0)
##  vctrs         0.5.1   2022-11-16 [1] CRAN (R 4.2.0)
##  withr         2.5.0   2022-03-03 [1] CRAN (R 4.2.0)
##  xfun          0.36    2022-12-21 [1] CRAN (R 4.2.0)
##  yaml          2.3.6   2022-10-18 [1] CRAN (R 4.2.0)
##
##  [1] /Users/sebastiansaueruser/Rlibs
##  [2] /Library/Frameworks/R.framework/Versions/4.2/Resources/library
##
## ------------------------------------------------------------------------------
```

1 Fragen stellen

1.1 Lernsteuerung

1.1.1 Standort im Lernpfad

Abb. Abbildung 1.9 zeigt den Standort dieses Kapitels im Lernpfad und gibt damit einen Überblick über das Thema dieses Kapitels im Kontext aller Kapitel.

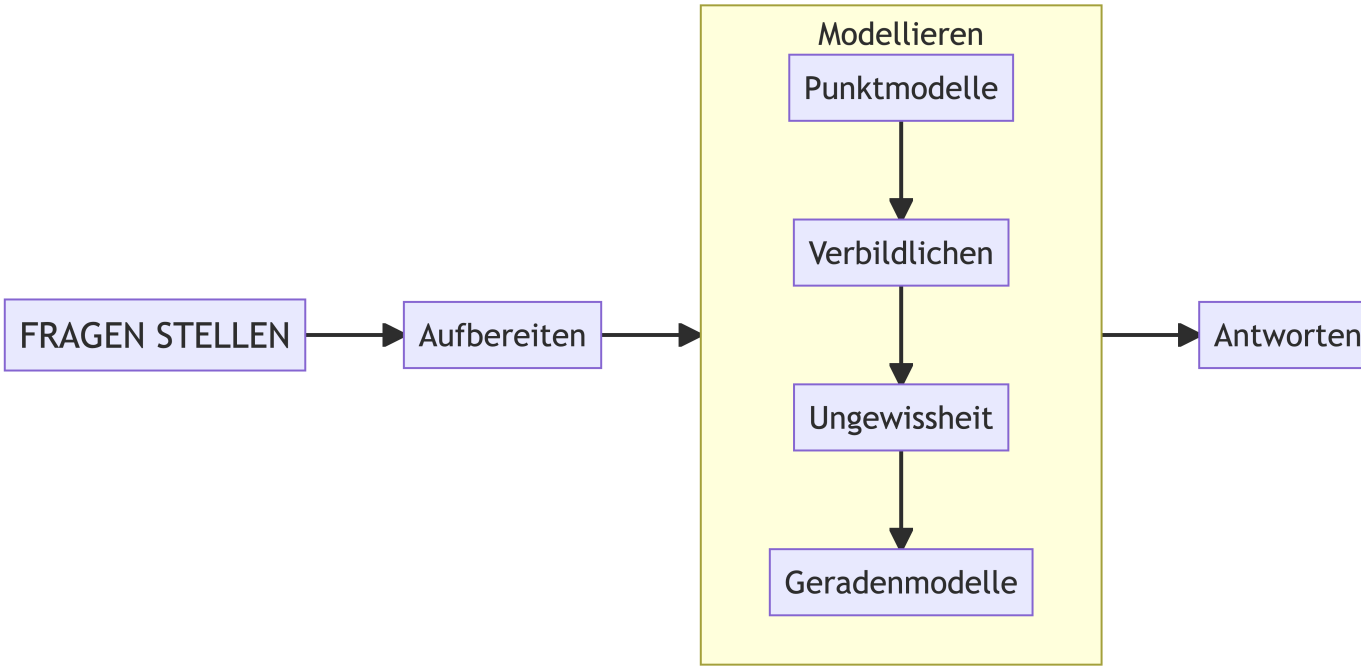

Abbildung 1.1: Überblick über den Inhalt und Verlauf des Buches

1.1.2 Lernziele

- Sie können eine Definition von Statistik wiedergeben.
- Sie können eine Definition von Daten wiedergeben.
- Sie können den Begriff Tidy-Daten erläutern.
- Sie können Beispiele für verschiedene Skalenniveaus nennen.

1.2 Was ist Statistik und wozu ist sie gut?

Es gibt mehrere Definition von Statistik; hier ist eine.

Definition 1.1 (Statistik)**.** Statistik fasst Daten zusammen, um wesentliche Informationen den Daten zu entnehmen und beschreibt die Ungewissheit unserer Schlüsse Kaplan (2009), Poldrack (2023).

Ein zentrales Vorgehen bei statistischen Analysen ist es, die *Unterschiedlichkeit der Dinge* zu beschreiben, präziser gesagt: die *Variation zu quantifizieren*. Betrachten wir dazu das Beispiel in s. Abbildung 1.2.

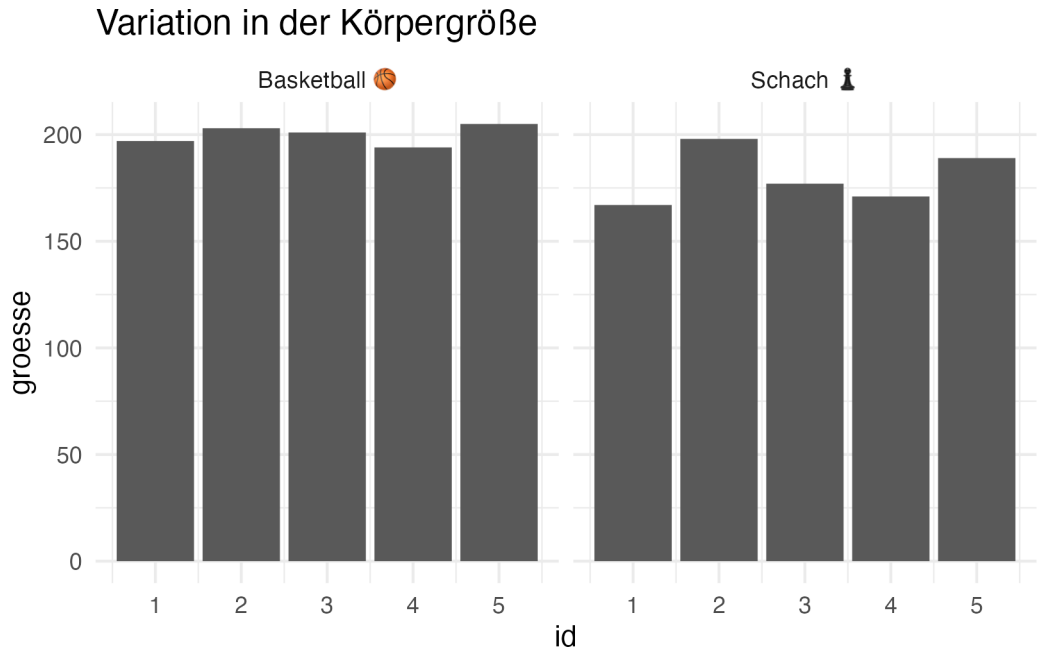

Abbildung 1.2: Wenig Variation in der Körpergröße bei den Basketballern. Alles lange Kerle. Viel Variation bei den Schachspielern: Manche sind klein, ander groß.

Bei den Basketballern gibt es *wenig* Variation in der Körpergröße - alle sind groß, ähnlich groß. Bei den Schachspielenr gibt es (im Verhältnis) *viel* Variation: Einige Personen sind groß, andere klein.

Die Variation (auch "Variabilität" genannt) kann man auch gut so darstellen wie in s. Abbildung 1.3 gezeigt.

Eine *Abweichung* (auch *Residium*) genannt, zeigt die Differenz von Mittelwert und dem Wert der Körpergröße bei der jeweiligen Person. Wenn wir allgemein von einer Person i sprechen, die Körpergröße mit x bezeichnen und den Mittelwert der Körpergröße als $\bar{x}$ ("x quer"), dann können wir knapp und präzise das Residuum r so definieren: $r = x_i - \bar{x}$

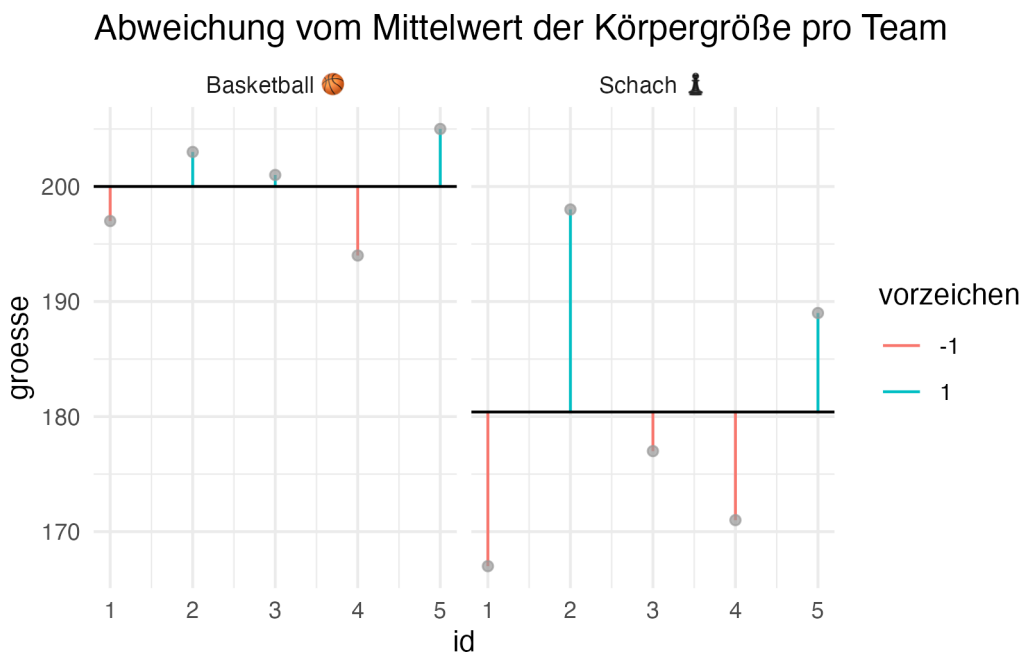

Abbildung 1.3: Die Abweichungen der einzelnen Personen von der mittleren Körpergröße ihres Teams

1.3 Was ist das Ziel Ihrer Analyse?

1.3.1 Arten von Zielen

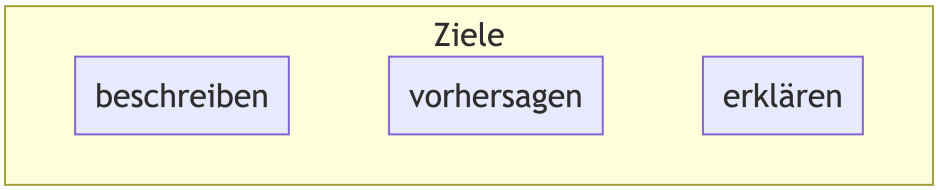

Abbildung 1.4: Zielarten einer Datenanalyse

Beispiele für die einzelnen Zielarten der Datenanalyse:

- *Beschreiben*: “Wie groß ist der Gender-Paygap in der Branche X im Zeitraum Y?”
- *Vorhersagen*: Wenn ich 100 Stunden auf die Statistikklausur lernen, welche Note kann ich dann erwarten?
- *Erklären*: Wie viel bringt mir das Lernen auf die Statistikklausur?

1.3.2 Forschungsfrage

Eine Forschungsfrage ist die Leitfrage Ihrer Analyse. Sie definiert, was Sie herausfinden wollen. Häufig sind Forschungsfragen so aufgebaut:

> Hat X einen Einfluss auf Y?

Beispiel 1.1 (Forschungsfrage 1)**.**

> Hat Lernen einen Einfluss auf den Prüfungserfolg? Verringert Joggen die Menge des Hüftgolds? Um welchen Betrag erhöht sich der Umsatz, wenn wir 1000€ mehr Werbung ausgeben?

Beispiel 1.2 (Forschungsfrage 2)**.** Nach dem Studium haben Sie bei einem großen Online-Auktionshaus angeheuert. Da Sie angaben, sich im Studium ~~intensiv~~ etwas mit Statistik beschäftigt zu haben, hat man Sie in die F&E-Abteilung[1] gesteckt. Heute ist es Ihre Aufgabe, Auktionen zur Spielekonsole Wii zu untersuchen, genauer gesagt, geht es um das Spiel Mariokart. Ihre Forschungsfrage lautet:

> Welche Produktmerkmale stehen mit einem hohen Verkaufserlös in Zusammenhang?

Eine Forschungsfrage weist häufig folgende Struktur auf, s. Abbildung 1.5.

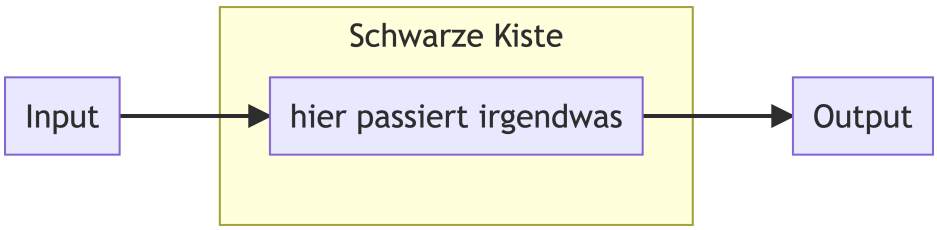

Abbildung 1.5: Struktur eine Forschungsfrage

[1] Forschung und Entwicklung

1.4 Was sind Daten?

Definition 1.2 (Daten)**.** Daten kann man als eine geordnete Folge von Zeichen definieren.

Daten kommen häufig in Tabellenform vor; so sind sie am besten zu untersuchen, s. Tabelle 1.1.

Tabelle 1.1: So sehen Daten aus.

id	name	note
1	Anna	1.3
2	Berta	2.3
3	Carla	3.0

Die erste Spalte `id` ist nur eine laufende Nummer. Sie dient dazu, die einzelnen Beobachtungen (hier Studentis) identifizieren zu können und birgt ansonsten keine Information. Beispiele für ID-Variabeln sind z.B. Matrikulationsnummer, Personalausweisnummern oder Bestellnumern.

Beispiel 1.3 (Daten zur Forschungsfrage 2)**.** Hier ist ein Auszug der Daten zur Tabelle `mariokart`:

id	duration	n_bids	cond	start_pr	ship_pr	total_pr	ship_sp	seller_rate	stock_photo	wheels
150377422259	3	20	new	0.99	4.00	51.55	standard	1580	yes	1
260483376854	7	13	used	0.99	3.99	37.04	firstClass	365	yes	1
320432342985	3	16	new	0.99	3.50	45.50	firstClass	998	no	1
280405224677	3	18	new	0.99	0.00	44.00	standard	7	yes	1
170392227765	1	20	new	0.01	0.00	71.00	media	820	yes	2
360195157625	3	19	new	0.99	4.00	45.00	standard	270144	yes	0

Eine Erklärung aller Variablen findet sich hier. Eine Erklärung, was die Namen einer Datentabelle bedeuten, nennt man *Code Book* or *Data Dictionary*.

1.4.1 Was ist eine Variable?

Definition 1.3 (Variable)**.** Eine Variable ist ein Platzhalter, der für ein Merkmal steht, das verschiedene Werte annehmen kann.

Man kann sich eine Variable wie einen Behälter vorstellen, auf dem mit einem Stift geschrieben steht, was für eine Art Inhalt darin ist, s. Abbildung 1.6.

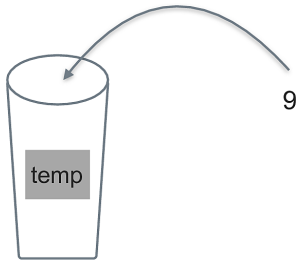

Abbildung 1.6: Wir definieren eine Variable “temp” mit dem Inhalt “9”

1.4.2 Beobachtungseinheit

Definition 1.4 (Beobachtungseinheit)**.** Beobachtungseinheiten sind die Dinge, die wir untersuchen (beobachten). Beobachtungseinheiten sind die Träger von Variablen.

In Tabelle 1.1 gibt es drei Variablen: `id`, `Name` und `Note`. Es gibt auch drei Beobachtungseinheiten: *Anna*, *Berta* und *Carla.*

1.4.3 Wert

Definition 1.5. Ein *Wert* ist der Inhalt einer Variablen.

In Abbildung 1.6 ist der Wert von `temp` 9.

In Tabelle 1.1 hat die Variable `name` drei Elemente: Anna, Berta, Carla. Der Wert des 2. Elements ist Berta.

Als *Ausprägungen* bezeichnet man die verschiedenen Werte einer Variablen.

Beispiel 1.4. In einer Studie wurden zehn Probanden untersucht. Die Variable `geschlecht` dokumentiert die Geschlechter der Personen:

```
geschlecht <- c("Mann", "Frau", "Frau", "Frau", "Mann",
                "Frau", "Mann", "Mann", "divers", "Frau")
geschlecht
##  [1] "Mann"   "Frau"   "Frau"   "Frau"   "Mann"   "Frau"   "Mann"   "Mann"
```

```
##  [9] "divers" "Frau"
```

In dieser Variable (die aus 10 Werten besteht) finden sich drei Ausprägungen: divers, Frau, Mann.

1.4.4 Tidy-Data

Definition 1.6. Unter *Tidy-Data* (tidy data) versteht man eine Tabelle, in der jede Zeile eine Beobachtungseinheit darstellt, jede Spalte eine Variable und jede Zelle der Tabelle einen Wert, s. Abbildung 1.7. (Zusätzlich ist noch eine "Kopfzeile" erlaubt, in der die Namen der Variablen stehen.)

Tabelle 1.1 ist ein Beispiel für Tidy-Data.

Abbildung 1.7 zeigt ein Sinnbild für Tidy-Data (Wickham und Grolemund 2018).

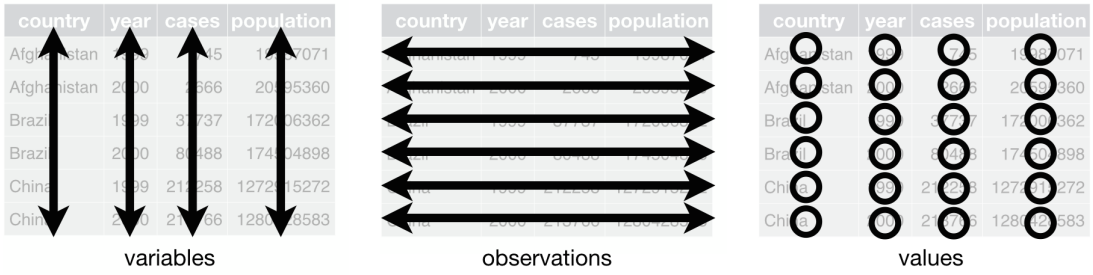

Abbildung 1.7: Tidy-Data

> **!** Wichtig
>
> Für eine statistische Analyse ist es fast immer nötig, dass die Daten im Tidy-Format vorliegen.

Der Vorteil des Tidy-Formats ist es, dass man weiß, wie die Daten aufgebaut sind. Außerdem können Statistikprogramme oft mit dieser Form am besten umgehen.

Das Tidy-Format wird auch als "langes" Format bezeichnet.

Abbildung 1.8 zeigt einen Datensatz in der "langen" Form, also tidy, und den gleichen Datensatz, umformatiert in der "breiten" Form, nicht-tidy.

In fast allen Organisationen werden Exceltabellen zur Datenverarbeitung verwendet. Dabei wiederholen sich immer wieder die gleichen Fehler bzw. suboptimalen Vorgehensweise zum Aufbau einer Exceltabelle.

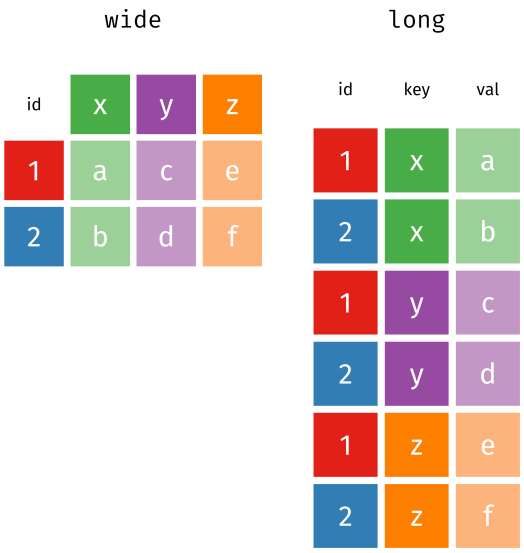

Abbildung 1.8: Links: Eine Tabelle mit Format “wide” - nicht “tidy”. Rechts: Das “Langformat” ist tidy.

Dieser Artikel von Broman und Woo (2018) zeigt anhand einiger praktischer Tipps, wie man Exceltabellen so aufbaut, dass Fehler minimiert werden.

1.4.5 Arten von Variablen

1.4.5.1 Nach Position in der Forschungsfrage

Angenommen, Ihre Forschungsfrage lautet:

> Hat Lernen einen Einfluss auf den Prüfungserfolg?

In dem Fall gilt:

- *Lernen* ist die Inputvariable/X-Variable/Ursache/UV
- *Prüfungserfolg* ist die Outputvariable/Y-Variable/Wirkung/AV

Abbildung 1.9 stellt diese beiden “Positionen” einer Variable dar. Die erste Position ist vor dem Pfeil. Die zweite Position ist nach dem Pfeil.

1.4.5.2 Nach dem Skalenniveau

Abbildung 1.10 gibt einen Überblick über typisch verwendete Skalenniveaus.

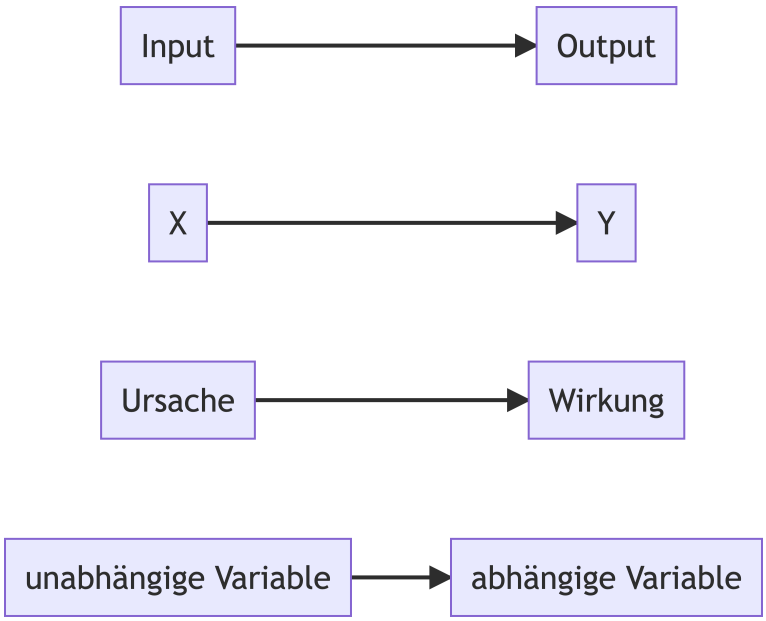

Abbildung 1.9: Überblick über den Inhalt und Verlauf des Buches

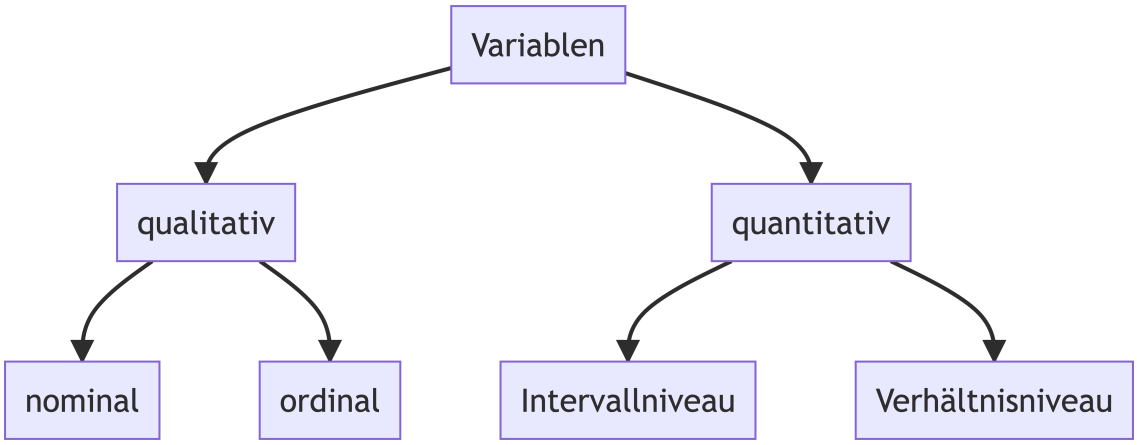

Abbildung 1.10: Skalenniveaus

1.5 Beispiele für Skalenniveaus

Beispiele zu den Skalenniveaus sind in Tabelle 1.3 aufgeführt.

Tabelle 1.3: Beispiele für Skalenniveaus

Variable	Skalenniveau
Haarfarbe	Nominalskala
Augenfarbe	Nominalskala
Geschlecht	Nominalskala
Automarke	Nominalskala
Partei	Nominalskala
Lieblingsessen	Ordinalskala
Medaillen beim 100-Meter-Lauf	Ordinalskala
Uniranking	Ordinalskala
IQ	Intervallskala
Extraversion	Intervallskala
Temperatur in Celcius	Intervallskala
Temperatur in Fahrenheit	Intervallskala
Temperatur in Kelvin	Verhältnisskala
Körpergröße	Verhältnisskala
Geschwindigkeit	Verhältnisskala
Länge	Verhältnisskala

Je nach dem, über welches Skalenniveau eine Variable verfügt, sind verschiedenen Rechenoperationen erlaubt, s. Tabelle 1.4.

Tabelle 1.4: Erlaubte Rechenoperationen nach Skalenniveau

Skalenniveau	Quantitativ	$+$	$\times$
Nominalniveau	nein		
Ordinalniveau	nein		
Intervallniveau	ja		
Verhältnisniveau	ja		

Was soll das bedeuten, Rechenoperationen? M

Schauen wir uns für jedes Skalenniveau ein "Rechenbeispiel" an.

Nominalskala: Die Variable *Geschlecht* ist nominalskaliert. Das bedeutet, dass ihre Ausprägungen *Frau* und *Mann* z.B. nicht (sinnvoll) addiert oder sonswie "verrechnet" werden können. Man könnte, z.B. um das Eintippen zu erleichtern, Frauen mit `1` kodieren und Männer mit `2`. Damit darf man aber nicht rechnen! Es macht keinen Sinn zu sagen: "Ich habe eine Frau und einen Mann in meiner Tabelle, das ist im Schnitt ein diverses Geschlecht, weil der Mittelwert von 1 und 2 ist 1,5!"

Die *einzige* “Rechenoperation”, die man auf der Nominalskala machen darf, ist die Prüfung auf Gleichheit: Mann kann feststellen, ob ein Objekt gleich zu einem anderen ist oder unterschiedlich. Also ob zwei Personen das gleiche Geschlecht haben oder von unterschiedlichem Geschlecht sind. Etwas formaler ausgedrückt:

- $\neq$
- $=$
- $=$

Ordinalskala: Diese Skala entspricht einer Rangordnung. Eine Rangordnung ist etwa die geordnete Abfolge Ihres Leibgerichte[2]. Etwas formaler ausgedrückt:

- $\succ$ $\succ$

Das komische Zeichen $\succ$ soll heißen: “Ist auf meiner Liste von Leibgerichten weiter oben, mag ich lieber”. Man kann aber *nicht* sagen, “Ich mag aber Pizza um 42% mehr als die Spagetthi und die wieder um 73% mehr als ein Schnitzel!”. Zumindest kann man das nicht ohne weitere Informationen und Annahmen. Es gibt also Dinge auf der Welt, die man leicht in eine Rangordnung bringen kann, aber die man nur schwer in der Größe der Unterschiede bemessen kann. Das ist die Ordinalskala, s. Abbildung 1.11.

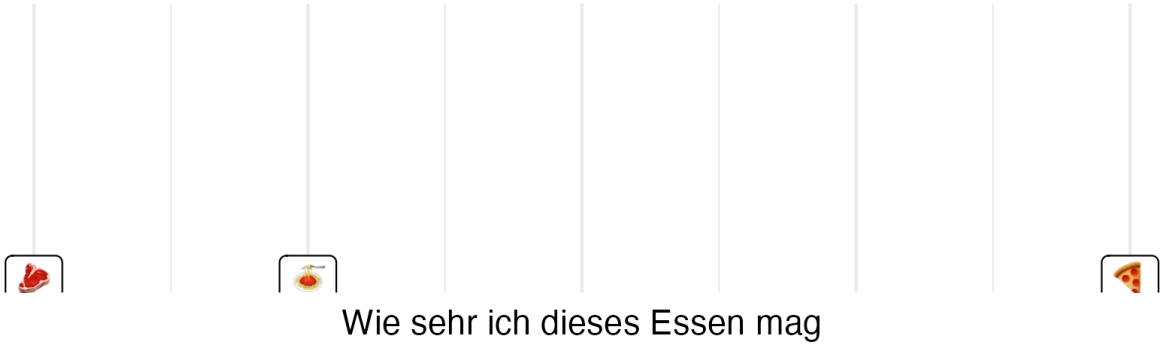

Abbildung 1.11: Die Ordinalskala: Je weiter rechts auf der X-Achse, desto höher lieber esse ich das Gericht.

Intervallskala: Das ist vielleicht eine Überraschung für Sie: Wenn es heute 10°C hat und morgen 5°C – dann ist es heute *nicht* doppelt so warm wie morgen. Ja, 10 ist das Doppelte von 5. Aber *10° Celcius* ist nicht doppelt so warm wie 20° Celcius. Wenn Sie das verwundert: Das ist normal, so geht es den meisten, wenn sie das zum ersten Mal hören. Der Grund, dass es nicht erlaubt ist, Verhältnisse (wie doppelt/halb so viel etc.) auf der Celcius-Skala zu bilden, ist, dass der Nullpunkt der Skala, 0° C, kein echter, physikalischer Nullpunkt ist. Bei 0° C liegt eben nicht Null Wärmeenergie vor. Stattdessen wurde eine Wärmenergiemenge gewählt, die für uns Menschen ganz praktisch, da augenfällig ist: der Gefrierpunkt von Wasser. Was bei der Intervallskala erlaubt ist, ist das Addieren (und Subtrahieren): heute 10°C, morgen 5°C, das ist ein Unterschied von 5°C. Oder: Im Schnitt waren es 7,5°C, das ist genau in der Mitte von 5 und 10°C. Abbildung 1.12 versinnbildlicht die Intervallskala.

[2] 1. Pizza, 2. Spagetthi, 3. Schnitzel

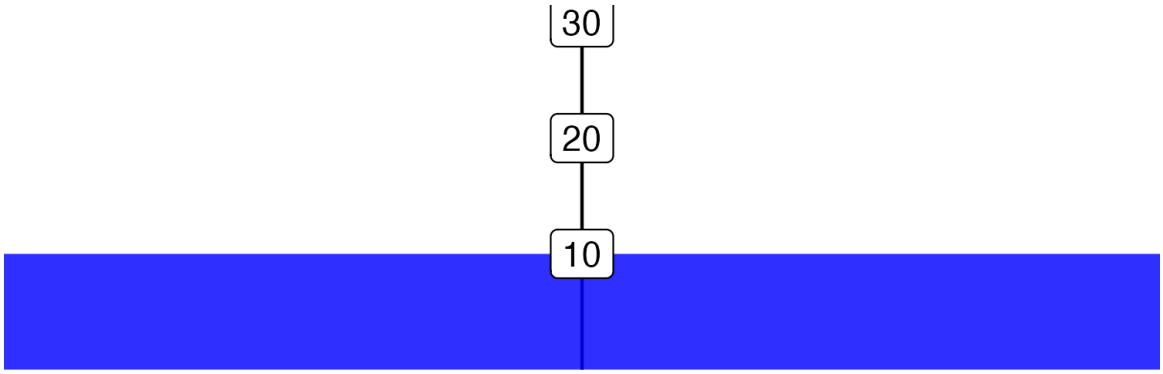

Abbildung 1.12: Ein Metermaß steckt im Wasser. Auf dem Metermaß können wir die aufgedruckten Zahlen ablesen. Aber wir wissen nicht, ob der Metermaß auf dem Boden steht. Wir wissen demnach nicht, ob der vom Metermaß angegebene Nullpunkt der wahre Nullpunkt (Meeresboden) ist.

Verhältnisskala: Eine Verhältnisskala ist das, was man sich gemeinhin unter einer metrische Variable vorstellt: Man kann "normal" rechnen, alle Rechenoperationen sind erlaubt. Zuzüglich zu denen, die auch in anderen, "niedrigeren" Skalenniveaus erlaubt sind, ist das das Bilden von Verhältnissen - Multiplizieren, s. Abbildung 1.13.

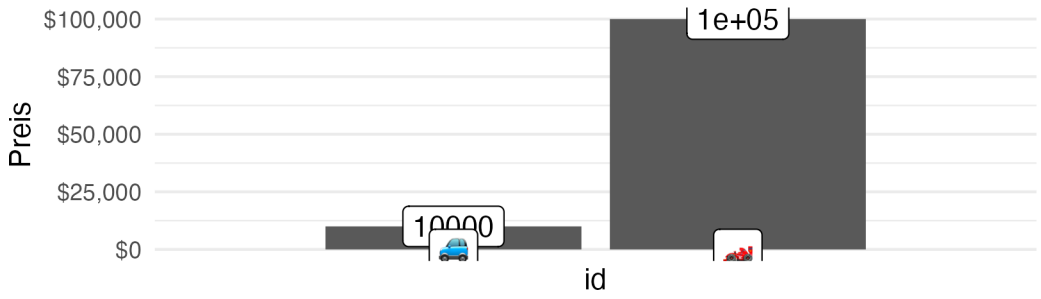

Abbildung 1.13: Puh! Der rote Flitzer ist 10 Mal so teuer wie die blaue Möhre. Kohlen zusammenkratzen."

Außerdem können quantitative Variablen untergliedert werden in:

- *stetige* Variablen, das sind Variablen, bei denen man zwischen zwei Ausprägungen immer noch eine weitere quetschen kann. So gibt es eine Wert für die Köpergröße zwischen 1.60m und 1.61. Und einen Wert zwischen 1.601m und 1.602m, etc.
- diskrete Variablen, das sind metrische Variablen, die nur bestimmte Ausprägungen haben, häufig sind das die natürlichen Zahlen: $1, 2, ...$ Ein Beispiel wäre die Anzahl der Kinder in einer Familie.

1.6 Modelle

Woran denken Sie beim Wort "Modell"? Vielleicht an Spielzeugautos, s. Abbildung 1.14.

Abbildung 1.14: Matchbox-Autos sind Modelle für Autos

Definition 1.7 (Modelle)**.** Modelle sind ein vereinfachtes Abbild der Realität eine *Repräsentation* (Kaplan 2009).

Beispiel 1.5 (Beispiele für Modelle)**.** Puppen sind Modelle für Babies, Landkarten für Landstriche und das Atommodell von Nils Bohr ist ein Modell für Atome.

Auch in der Statistik nutzen wir Modelle. Helfen Sie Prof. Weiss-Ois: Er blickt nicht durch. Gerne würde er wissen, wie viele Stunden seine Studentis auf die Prüfung lernen. Aber mit so vielen Zahlen kann er nicht umgehen … Geben Sie ihm ein Modell: Sagen Sie ihm, wie lang die Studis typischerweise lernen (sagen Sie ihm ein einfach den Mittelwert der Lernzeiten).

12, 8, 10, 11, 10, 9, 13, 9, 14, 9, 12, 14, 7, 9, 9, 11, 9, 4, 5, 12, 9, 6, 9, 12, 13, 9, 9, 6, 10, 8

Abbildung 1.15: Oh jeh, so viele Zahlen! Ich check nix! Wie viel lernen denn jetzt meine Studis?!

Flaticon licence, Autor: iconixar

9.6

Flaticon licence, Autor: iconixar

Abbildung 1.16: Yeah, jetzt weiß ich, wie viel die Studis so typischerweise lernen. Viel zu wenig natürlich!

Der Nutzen von Modellen ist, dass sie komplexe Sachverhalte vereinfachen und damit oft überhaupt erst einer Untersuchung zugänglich. In der Datenanalyse bzw. Statistik[3] fassen Sie oft viele Daten prägnant zusammen, z.B. zu einer einzelnen Kennzahl.

1.7 Praxisbezug

Wir leben im Datenzeitalter; Daten durchdringen alle Bereiche des beruflichen, gesellschaftlichen und privaten Lebens. Die Datenanalyse hat sich in den letzten Jahren massiv verändert, s. Abbildung 1.17.

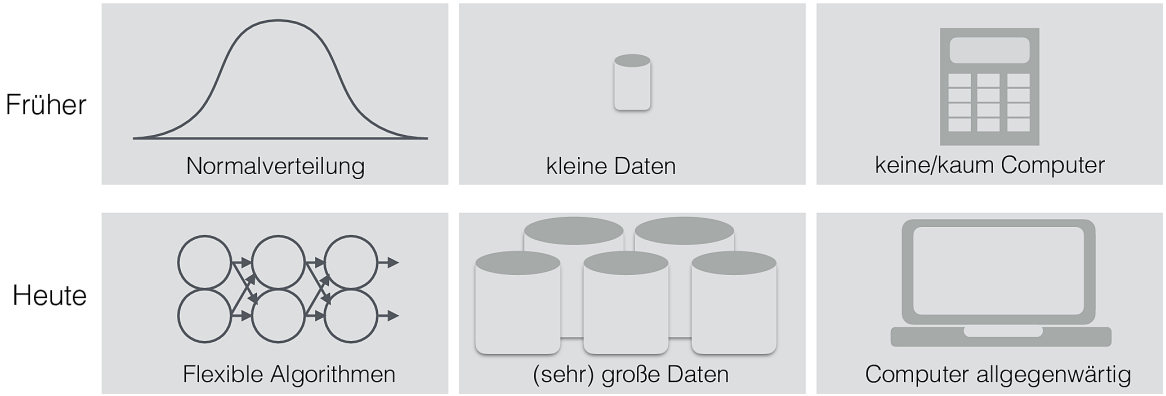

Abbildung 1.17: Forschung früher und heute

Diese Entwicklung ist durchaus auch kritisch zu betrachten. Mit der wachsenden Bedeutung von Daten wächst in gleichem Maße die Bedeutung von Datenanalyse. Denn Daten ohne Sinn sind nutzlos. Aus diesem Grund kann man sagen, dass Datenanalyse (und damit auch Statistik als eine spezielle Art von Datenanalyse) zu stark nachgefragten Jobs gehören.

Laut dem Entgeltatlas der Bundesagentur für Arbeit liegt ein typisches Gehalt von Data Scientisten bei knapp 6000 € pro Monat (in der Altersgruppe von 25 bis 54)[4]. Laut dem

[3] die beiden Begriffe werden hier weitgehend synonym gebraucht
[4] Abrufdatum: 1.2.23

Gehaltsreporter liegt das Einstiegsgehalt dieser Berufsgruppe bei knapp 50.000€ pro Jahr.

1.8 Fazit

Die Aufgabe von Statistik ist es, durch Zusammenfassen von Daten Modelle zu bilden, die es uns einfacher machen, schwierige Sachverhalte zu verstehen. Zentral ist dabei, die Analyse von Variabilität der Daten. Daten kommen in verschiedenen Varianten vor, typischerweise in Tabellenform, möglichst im Tidy-Format.

1.9 Aufgaben

1. variation01
2. Def-Statistik01
3. tidy1
4. Skalenniveau1a
5. Ziele-Statistik
6. variation02
7. Skalenniveau1b

1.10 Vertiefung

nicht prüfungsrelevant

1. Fassen Sie den Artikel von Broman und Woo (2018) zusammen.

1.11 Literatur

2

Broman, Karl W., und Kara H. Woo. 2018. „Data Organization in Spreadsheets“. *The American Statistician* 72 (1): 2–10. https://doi.org/10.1080/00031305.2017.1375989.

Kaplan, Daniel T. 2009. *Statistical modeling: a fresh approach*. Scotts Valley, Calif.: CreateSpace. https://dtkaplan.github.io/SM2-bookdown/.

Poldrack, Russell A. 2023. *Statistical thinking: analyzing data in an uncertain world*. Princeton: Princeton University Press. https://statsthinking21.github.io/statsthinking21-core-site/.

Wickham, Hadley, und Garrett Grolemund. 2018. *R für Data Science: Daten importieren, bereinigen, umformen, modellieren und visualisieren*. Übersetzt von Frank Langenau. 1. Auflage. Heidelberg: O'Reilly. https://r4ds.had.co.nz/index.html.

World Economic Forum. 2020. „The Future of Jobs Report 2020“. CH-1223 Cologny/Geneva Switzerland: World Economic Forum. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf.