Statistik_21

Sebastian Sauer

2016-11-21

Contents


```
source("./source/libs.R")
#> Loading tidyverse: ggplot2
#> Loading tidyverse: tibble
#> Loading tidyverse: tidyr
#> Loading tidyverse: readr
#> Loading tidyverse: purrr
#> Loading tidyverse: dplyr
#> Conflicts with tidy packages ----------------------------------------------
#> filter(): dplyr, stats
#> lag():    dplyr, stats
```

Vorwort

- Worum geht es in diesem Buch
  - Einführung in moderne Verfahren der Statistik
  - Für Praktiker
  - Betonung liegt auf “modern” und “Praktiker”
- Ziel des Buches
  - Intuitives, grundlegendes Verständnis zu zentralen Konzepten
  - Handwerkszeug zum selber Anwenden
- Unterschied zu anderen Büchern
  - Wenig Formeln
  - Keine/weniger “typischen” klassischen Methoden wie ANOVA, Poweranalyse etc.
  - Aufzeigen von Problemen mit klassischen Verfahren
  - Kritik am Status-Quo
- Didaktik
  - Hands-on
  - R
  - Lernfragen
  - Fallstudien
  - Aktuelle Entwicklungen ausgerichtet

Chapter 1

Einführung

1.1 Rahmen

Der "Rahmen" dieses Buches ist der Überblick über wesentliche Schritte der Datenanalyse (aus meiner Sicht). Es gibt viele Ansätze, mit denen der Ablauf von Datenanalyse dargestellt wird. Der im Moment populärste oder bekannteste ist wohl der von Hadley Wickham und Garret Grolemund (Wickham and Grolemund, 2016). Hadley und Garrett haben einen "technischeren" Fokus als der dieses Buches. Ihr Buch "R for Data Science" ist hervorragend (und frei online verfügbar); nur ist der Schwerpunkt ein anderer; es baut ein viel tieferes Verständnis von R auf. Hier spielen aber statistisch-praktisch und statistisch-philosophische[1] Aspekte eine größere Rolle.

Das Diagramm **??** stellt den Rahmen dieses Buch dar: Die drei Hauptaspekte sind *Umformen*, *Visualisieren* und *Modellieren*. Dies ist vor dem Hintergrund der *Reproduzierbarkeit* eingebettet.

Mit *Umformen* ist gemeint, dass Daten in der Praxis häufig nicht so sind, wie man sie gerne hätte. Mal fehlt eine Variable, die den Mittelwert anderer ausdrückt, oder es gibt unschöne "Löcher", wo starrsinnige Versuchspersonen standhaft keine Antwort geben wollten. Die Zahl an Problemen und (Arten von) Fehlern übersteigt sicherlich die Anzahl der Datensätze. Kurz: Wir sehen uns gezwungen, den Daten einige Einblick abzuringen, und dafür müssen wir sie erst in Form bringen, was man als eine Mischung zwischen Artistik und Judo verstehen kann.

Ein Bild sagt mehr als 1000 Worte, weiß der Volksmund. Für die Datenanalyse gilt dies auch. Ein gutes Diagramm vermittelt eine Fülle an Informationen "auf einen Blick" und erzielt damit eine Syntheseleistung, die digitalen Darbietungsformen, sprich: Zahlen, verwehrt ist. Nebenbei sind Diagramme, mit Geschick erstellt, ein Genuss für das Auge, daher kommt der *Visualisierung* großen Wert zu.

Als letzten, aber wesentlichen Punkt führen wir das *Modellieren* auf. Es gibt mehr Definitionen von "Modell" als ich glauben wollte, aber hier ist damit gemeint, dass wir uns eine Geschichte ausdenken, wie die Daten entstanden sind, oder präziser gesagt: welcher Mechanismus hinter den Daten steht. So könnten wir Klausurnoten und Lernzeit von einigen Studenten[2] anschauen, und verkünden, wer mehr lerne, habe auch bessere Noten (ein typischer Dozentenauspruch). Unser Modell postuliert damit einen linearen Anstieg des Klausurerfolgs bei steigender Vorbereitungszeit. Das schönste an solchen Modellen ist, dass wir Vohersagen treffen können. Zum Beispiel: "Joachim, du hast 928 Stunden auf die Klausur gelernt; damit solltest du 93% der Punkte erzielen".

Was ist dann mit dem *Reproduzierbarkeitshintergrund* gemeint? Ihre Arbeiten von Umformen, Visualisieren und Modellieren sollten sich nicht ausschließlich im Arbeitsspeicher Ihres Gehirns stattfinden, auch wenn

[1] Ich hörte mal, jede Formel halbiert die Leserzahl eines Buches. Wahrsheinlich gilt das gleiche für jede Erwähnung von "philosophisch" und zugehörigen Begriffen. Zumindest zwei Leser hat dieses Buch (den Autor mitgezählt) …

[2] hier und überall sind Frauen und Männer gleichermaßen angesprochen; aus Gründen der Lesbarkeit das generische Maskulinum bevorzugt.

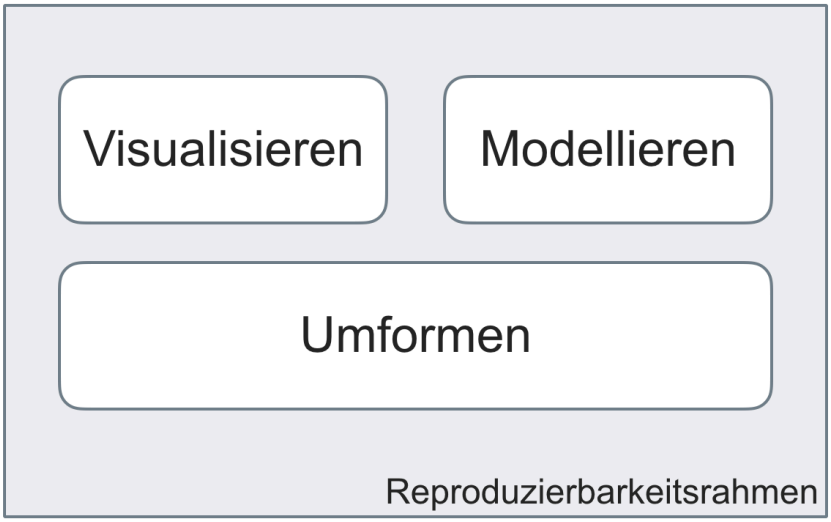

Figure 1.1: Rahmen

das bei Ihnen, lieber Leser, vielleicht schneller ginge. Stattdessen soll der Mensch sich Mühe machen, seine Gedanken aufzuschreiben, hier insbesondere die Rechnungen bzw. alles, was den Daten angetan wurde, soll protokolliert werden (auch die Ergebnisse, aber wenn der Weg dorthin klar protokolliert ist, kann man die Ergebnisse ja einfach "nachkochen"). Ein Vorteil dieses Vorgehens ist, dass andere (inklusive Ihres zukünftigen Ich) die Ergebnisse bzw. das Vorgehen einfacher nachvollziehen können.

1.2 Was ist Statistik? Wozu ist sie gut?

Diese zwei Fragen sollte man sich am Anfang der Beschäftigung mit jedem Thema stellen. Was ist Stististik? *Eine* Antwort dazu ist, dass Statistik die Wissenschaft von Sammlung, Analyse, Interpretation und Kommunikation mithilfe mathematischer Verfahren ist und zur Entscheidungshilfe beitragen solle (oxf, 2006; Romeijn, 2016). Damit hätten wir auch den Unterschied zur schnöden Datenanalyse (ein Teil der Statistik) herausgemeiselt. Statistik wird häufig in die zwei Gebiete *deskriptive* und *inferierende* Statistik eingeteilt. Erstere fasst viele Zahlen zusammen, so dass wir den Wald statt vieler Bäume sehen. Letztere verallgemeinert von den vorliegenden (sog. "Stichproben-")Daten auf eine zugrunde liegende Grundmenge (Population). Dabei spielt die Wahrscheinlichkeitsrechnung und Zufallsvariablen eine große Rolle.

Auch wenn die gerade genannte Diskussion die häufigste oder eine typische ist, mehren sich doch Stimmen, die Statstik anders akzentuieren. So schreibt Briggs in einem aktuellen Buch (Briggs, 2016), dass es in der Statistik darum ginge, die Wahrscheinlichkeit zukünftiger Ereignisse vorherzusagen: "Wie wahrscheinlichi ist es, dass - gegeben einem statistischen Modell, allerlei Annahmen und einem Haufen Daten - Kandidat X der neue Präsident wird"[3]? Das schöne an dieser Idee ist, dass das "Endprodukt" etwas sehr Weltliches und damit praktisches ist: Die Wahrscheinlichkeit einer interessanten (und unbekannten) Aussage. Nebenbei ruht diese Idee auf dem sicheren Fundament der Wahrscheinlichkeitstheorie.

Abgesehen von philosophischen Überlegungen zum Wesen der Statistik kann man sagen, dass Vorhersagen etwas sehr praktisches sind. Sie nehmen daher aus praktischen Überlegungen einen zentralen Platz in diesem Buch an. Die philosophische Relevanz des prädiktiven Ansatzes ist gut bei Briggs (Briggs, 2016; **?**) nachzulesen.

- Statistik meint Methoden, die das Ziel haben, Ereignisse präzise vorherzusagen

[3]Eine Vorhersage, die bei vielen Vorhersagemodellen komplett in die Grütze ging, wenn man sich die US-Präsidentenwahl 2016 anschaut.

- Statistik soll sich um Dinge dieser Welt drehen, nicht um Parameter
- Statt einer Frage “ist µ_1 größer als µ_2?” besser “Wie viel Umsatz erwarte ich von diesem Kunden?”, “Wie viele Saitensprünge hatte er wohl?”, “Wie groß ist die Wahrscheinlichkeit für sie zu überleben?” und dergleichen.
- Der Nutzen von Vorhersagen liegt auf der Hand: Vorhersagen sind praktisch; eine nützliche Angelegenheit (wenn auch schwierig).

Chapter 2

Trends

- Big Data
- Open Science
- Computerisierung
- Neue Methoden zur numerischen Vorhersage
- Textmining

Chapter 3

Unbehagen

- p-Werte
- Theorie der Wahrscheinlichkeit im Frequentismus
- Reproduzierbarkeitskrise
- Parameter
- Kausalität
- Übersicherheit

Chapter 4

Datenjudo

Daten umformen.

- dplyr
- Normalform

Chapter 5

Visualisierung

- Nutzen (Anscombe)
- Prinzipien nach Tufte
- Cleveland
- ggplot2

Chapter 6

Statistisches Modellieren

- Was sind Modelle?
- Überanpassung
- Prädiktion vs. Explanation
- Numerische vs. klassifizierende Modelle
- Geleitete vs. ungeleitete Modelle
- Parametrische vs. nichtparametrische Modelle
- Fehler- vs. Varianzreduktion
- Modellgüte

Chapter 7

Numerische Modelle

- Lineare Regression
  - Grundlagen
  - Multiple Regression
  - Interaktion
  - Eisberge
- Logistische Regression
- Penalisierende Regression
- Baumbasierte Verfahre
- Ausblick

Chapter 8

Klassifizierende Modelle

- Clusteranalyse
- Nächste-Nachbarn-Analyse

Chapter 9

Literaturverzeichnis

Bibliography

(2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Briggs, W. (2016). *Uncertainty: The Soul of Modeling, Probability & Statistics*. Springer.

Romeijn, J.-W. (2016). Philosophy of statistics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition.

Wickham, H. and Grolemund, G. (2016). *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. O'Reilly Media.