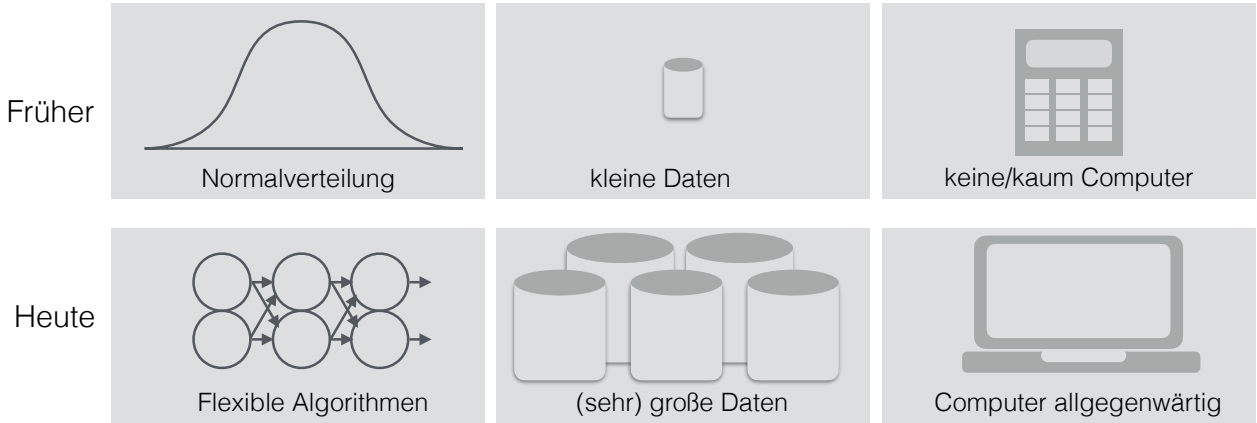
Praxis der Datenanalyse

Sebastian Sauer, Matthias Gehrke, Karsten Lübke, Oliver Gansser

Inhaltsverzeichnis

Vorwort

Statistik heute; was ist das? Sicherlich haben sich die Schwerpunkte von “gestern” zu “heute” verschoben. Wenig überraschend spielt der Computer eine immer größere Rolle; die Daten werden vielseitiger und massiger. Entsprechend sind neue Verfahren nötig - und vorhanden, in Teilen - um auf diese neue Situation einzugehen. Einige Verfahren werden daher weniger wichtig, z.B. der p-Wert oder der t-Test. Allerdings wird vielfach, zumeist, noch die Verfahren gelehrt und verwendet, die für die erste Hälfte des 20. Jahrhunderts entwickelt wurden. Eine Zeit, in der kleine Daten, ohne Hilfe von Computern und basierend auf einer kleinen Theoriefamilie im Rampenlicht standen [**?**]. Die Zeiten haben sich geändert!



Zu Themen, die heute zu den dynamischten Gebieten der Datenanalyse gehören, die aber früher keine große Rolle spielten, gehören [**?**]:

- Nutzung von Datenbanken und anderen Data Warehouses

- Daten aus dem Internet automatisch einlesen ("scraping")
- Genanalysen mit Tausenden von Variablen
- Gesichtserkennung

Sie werden in diesem Kurs einige praktische Aspekte der modernen Datenanalyse lernen. Ziel ist es, Sie - in Grundzügen - mit der Art und Weise vertraut zu machen, wie angewandte Statistik bei führenden Organisationen und Praktikern verwendet wird[1].

Es ist ein Grundlagenkurs; das didaktische Konzept beruht auf einem induktiven, intuitiven Lehr-Lern-Ansatz. Formeln und mathematische Hintergründe such man meist vergebens (tja).

Im Gegensatz zu anderen Statistik-Büchern steht hier die Umsetzung mit R stark im Vordergrund. Dies hat pragmatische Gründe: Möchte man Daten einer statistischen Analyse unterziehen, so muss man sie zumeist erst aufbereiten; oft mühselig aufbereiten. Selten kann man den Luxus genießen, einfach "nur", nach Herzenslust sozusagen, ein Feuerwerk an multivariater Statistik abzubrennen. Zuvor gilt es, die Daten aufzubereiten, umzuformen, zu prüfen und zusammenzufassen. Diesem Teil ist hier recht ausführlich Rechnung getragen.

"Statistical thinking" sollte, so eine verbreitete Idee, im Zentrum oder als Ziel einer Statistik-Ausbildung stehen [**?**]. Es ist die Hoffnung der Autoren dieses Buches, dass das praktische Arbeiten (im Gegensatz zu einer theoretischen Fokus) zur Entwicklung einer Kompetenz im statistischen Denken beiträgt.

Außerdem spielt in diesem Kurs die Visualisierung von Daten eine große Rolle. Zum einen könnte der Grund einfach sein, dass Diagramme ansprechen und gefallen (einigen Menschen). Zum anderen bieten Diagramme bei umfangreichen Daten Einsichten, die sonst leicht wortwörtlich übererseheen würden.

> Dieser Kurs zielt auf die praktischen Aspekte der Analyse von Daten ab: "wie mache ich es?"; mathematische und philosophische Hintergründe werden vernachlässigt bzw. auf einschlägige Literatur verwiesen.

R-Pseudo-Syntax: R ist (momentan) die führende Umgebung für Datenanalyse. Entsprechend zentral ist R in diesem Kurs. Zugebenermaßen braucht es etwas Zeit, bis man ein paar Brocken "Errisch" spricht. Um den Einstieg zu erleichern, ist Errisch auf Deutsch übersetzt an einigen Stellen, wo mir dies besonders hilfreich erschien. Diese Stellen sind mit diesem Symbol  gekennzeichnet (für R-Pseudo-Syntax).

Achtung, Falle: Schwierige oder fehlerträchtige Stellen sind mit diesem Symbol  markiert.

Übungsaufgaben: Das Skript beinhaltet in jedem Kapitel Übungsaufgaben oder/und Testfragen. Auf diese wird mit diesem Icon  verwiesen oder die Übungen sind in einem Abschnitt

[1] Statistiker, die dabei als Vorbild Pate standen sind: Roger D. Peng: http://www.biostat.jhsph.edu/~rpeng/, Hadley Wickham: http://hadley.nz, Jennifer Bryan: https://github.com/jennybc

mit einsichtigem Titel zu finden.

Love: Wenn Ihnen R diesen Smiley präsentiert, dann sind Sie am Ziel Ihrer Träume:  .

Dieses Buch hat einige *Voraussetzungen*, was das Vorwissen der Leser angeht; folgende Themengebiete werden vorausgsetzt:

- Deskriptive Statistik
- Grundlagen der Inferenzstatistik
- Grundagen der Regressionsanalyse
- Skalenniveaus
- Grundlagen von R

Dieses Skript wurde mit dem Paket `bookdown` [**?**] erstellt, welches wiederum stark auf den Paketen `knitr` [**?**] und `rmarkdown` [**?**] beruht. Diese Pakete stellen verblüffende Funktionalität zur Verfügung als freie Software (frei wie in Bier und frei wie in Freiheit).

Aus Gründen des Lesbarkeit wird das männliche Generikum verwendet, welches Frauen und Männer in gleichen Maßen ansprechen soll.

Die Bildnachweise sind in folgenden Muster aufgebaut: Nummer, Verweis zum Bild, Names des Autors, Titel, Quelle (URL), Lizenz, Abrufdatum.

1. Abb. **??**, Sebastian Unrau, ohne Titel, https://unsplash.com/photos/CoD2Q92UaEg, CC0, 2017-02-12

2. Abb. **??**, Lothar Spurzem, VW 1303 von Wiking in 1:87; Größe des Modells: 47,5 mm, https://de.wikipedia.org/wiki/Modellautomobil#/media/File:Wiking-Modell_VW_1303_(um_1975).JPG, CC-BY-SA 2.0, de.

Alle verwendeten Datensätze und R-Pakete finden sich im Literaturverzeichnis; im Text werden Pakete nicht zitiert.

Ein Teil dieses Skripts basiert auf Arbeiten von meinen Kollegen Oliver Gansser, Matthias Gehrke und Karsten Lübke. Ohne deren Unterstützung, Ermutigung und Kritik gäbe es diesen Kurs nicht. Gerade von Karsten Lübke habe ich einiges gelernt.

Sebastian Sauer

Kapitel 1

Organisatorisches

1.1 Modulziele

Die Studierenden können nach erfolgreichem Abschluss des Moduls:

- den Ablauf eines Projekts aus der Datenanalyse in wesentlichen Schritten nachvollziehen,
- Daten aufbereiten und ansprechend visualisieren,
- Inferenzstatistik anwenden und kritisch hinterfragen,
- klassische Vorhersagemethoden (Regression) anwenden,
- moderne Methoden der angewandten Datenanalyse anwenden (z.B. Textmining),
- betriebswirtschaftliche Fragestellungen mittels datengetriebener Vorhersagemodellen beantworten.

1.2 Themen pro Termin

Für dieses Modul sind 44 UE für Lehre eingeplant, aufgeteilt in 11 Termine (vgl. 1.1).

Tabelle 1.1: Themen pro Termin.

Termin	Thema/ Kapitel
1	Organisatorisches
	Einführung
	Rahmen
	Daten einlesen
2	Datenjudo
3	Daten visualisieren
4	Fallstudie
5	Daten modellieren

Termin	Thema/ Kapitel
	Der p-Wert
6	Lineare Regression - metrisch
7	Lineare Regression - kategorial
8	Fallstudie
9	Vertiefung: Textmining und Clusteranalyse
10	Vertiefung: Baumbasierte Verfahren
11	Wiederholung

1.3 Vorerfahrung

Bei den Studierenden werden folgende Themen als bekannt vorausgesetzt:

- Deskriptive Statistik
- Inferenzstatistik
- Grundlagen R
- Grundlagen der Datenvisualisierung

1.4 Prüfung

1.4.1 Prüfungshinweise

- Die Prüfung besteht aus zwei Teilen
  - einer Klausur (50% der Teilnote)
  - einer Datenanalyse (50% der Teilnote).

Prüfungsrelevant ist der gesamte Stoff aus dem Skript und dem Unterricht mit folgenden Ausnahmen:

- Inhalte/Abschnitte, die als "nicht klausurrelevant" gekennzeichnet sind,
- Inhalte/Abschnitte, die als "Vertiefung" gekennzeichnet sind,
- Fallstudien (nur für Klausuren nicht prüfungslevant),
- die Inhalte von Links,
- die Inhalte von Fußnoten,
- die Kpaitel *Vorwort*, *Organisatorisches* und *Anhang*.

1.4.2 Klausur

- Die Klausur besteht fast oder komplett aus Multiple-Choice (MC-)-Aufgaben mit mehreren Antwortoptionen; zumeist ist eine Antwort aus vieren auszuwählen.

- Die (maximale) Anzahl der richtigen Aussagen ist pro Aufgabe angegeben. Werden mehr Aussagen als "richtig" angekreuzt als angegeben, so wird die Aufgabe mit 0 Punkten beurteilt. Ansonsten werden Teilpunkte für jede Aufgabe vergeben.
- Jede Aussage gilt ceteris paribus (unter sonst gleichen Umständen). Aussagen der Art "A ist B" (z.B. "Menschen sind sterblich") sind *nur* dann als richtig auszuwählen, wenn die Aussage *immer* richtig ist.
- Im Zweifel ist eine Aussage auf den Stoff, so wie im Unterricht behandelt, zu beziehen. Werden in Aussagen Zahlen abgefragt, so sind Antworten auch dann richtig, wenn die vorgeschlagene Antwort ab der 1. Dezimale von der wahren Antwort abweicht (einigermaßen genaue Aussagen werden als richtig akzeptiert). Bei Fragen zu R-Syntax spielen Aspekte wie Enter-Taste o.ä. bei der Beantwortung der Frage keine Rolle; diese Aspekte dürfen zu ignorieren.
- Jede Aussage einer MC-Aufgabe ist entweder richtig oder falsch (aber nicht beides oder keines).
- Die MC-Aufgaben sind nur mit Kreuzen zu beantworten; Text wird bei der Korrektur nicht berücksichtigt.
- Bei Nachholklausuren gelten die selben Inhalte (inkl. Schwerpunkte) wie bei der Standard-Klausur, sofern nicht anderweitig angegeben.
- I.d.R. sind nur Klausurpapier und ein nicht-programmierbarer Taschenrechner als Hilfsmittel zulässig.
- Die Musterlösungen zu offenen Fragen sind elektronisch hinterlegt.

1.4.3 Datenanalyse

- Wenden Sie die passenden, im Modul eingeführten statistischen Verfahren an.
- Werten Sie die Daten mit R aus; R-Syntax soll verwendet und im Hauptteil dokumentiert werden.
- In der Wahl des Datensatzes sind Sie frei, mit folgender Ausnahme: Im Unterricht besprochene Datensätze dürfen nicht als Prüfungsleistung eingereicht werden (vgl. Abschnitt **??**).
- Der (Original-)Name des Datensatzes (sowie ggf. Link) ist bei der Anmeldung anzugeben.
- Gruppenarbeiten sind nicht zulässig.
- Hat sich jemand schon für einen Datensatz angemeldet, so darf dieser Datensatz nicht mehr gewählt werden ("first come, first serve").

- Fundorte für Datensätze sind z.B. hier, hier und hier; im Internet finden sich viele Datensätze[1].
- Schreiben Sie Ihre Ergebnisse in einer Ausarbeitung zusammen; der Umfang der Ausarbeitung umfasst ca. *1000-1500 Wörter* (nur Hauptteil; d.h. exklusive Deckblatt, Verzeichnisse, Anhang etc.).
- Untersuchen Sie 2-3 Hypothesen.
- Denken Sie daran, Name, Matrikelnummer, Modulname etc. anzugeben (Deckblatt). Bei der Gestaltung des Layout entscheiden Sie selbständig bitte nach Zweckmäßigkeit (und Ästhetik).
- Fügen Sie keine Erklärungen oder Definitionen von statistischen Verfahren an.

1.4.4 Gliederungsvorschlag zur Datenanalyse

1. Datensatz
    1. Beschreibung
        - Name
        - Hintergrund (Themengebiet, Theorien, Relevanz), ca. 100 Wörter
        - Dimension (Zeilen*Spalten)
        - Zitation (wenn vorhanden)
        - sonstige Hinweise (z.B. Datenqualität, Entstehung des Datensatzes)
    2. Variablendeskription (nur für Variablen der Hypothese)
        - Skalenniveaus
        - Kontinuität (nur bei metrischen Variablen)
        - R-Datentyp
        - Anzahl Fälle und fehlende Werte
        - Erläuterung der Variablen
2. Deduktive Analyse
    1. Hypothese(n) Beschreiben Sie die Vermutung(en), die Sie prüfen möchten, möglichst exakt.
    2. Deskriptive Statistiken
        - Berichten Sie deskriptive Statistiken für alle Variablen der Hypothesen.
        - Berichten Sie aber nur univariate Statistiken sowie Subgruppenanalysen dazu.
        - Berichten Sie ggf. Effektstärken.
    3. Diagramme

[1] Googeln Sie mal nach “open datasets” o.ä.

        - Visualisieren Sie Ihre Hypothese(n) bzw. die Daten dazu, gerne aus mehreren Blickwinkeln.

    4. Signifikanztest

3. Explorative Analyse

- Eörtern Sie interessante Einblicke, die über Ihre vorab getroffenen Hypothesen hinausgehen.
- Diagramme können hier eine zentrale Rolle spielen.

4. Diskussion

    1. Zentrale Ergebnisse Fassen Sie das zentrale Ergebnisse zusammen.

    2. Interpretation Interpretieren Sie die Ergebnisse: Was bedeuten die Zahlen/Fakten, die die Rechnungen ergeben haben?

    3. Grenzen der Analyse

        - Schildern Sie etwaige Schwachpunkte oder Einschränkungen der Analyse.
        - Geben Sie Anregungen für weiterführende Analysen dieses Datensatzes.

1.5 Literatur

Zum Bestehen der Prüfung ist keine weitere Literatur fomal notwendig; allerdings ist es hilfreich, den Stoff aus unterschiedlichen Blickwinkeln aufzuarbeiten. Dazu ist am ehesten das Buch von Wickham und Grolemund [**?**] hilreich, obwohl es deutlich tiefer geht als dieses Skript.