stats-nutshell

Sebastian Sauer

9/04/2022

Table of contents

Preface

Figure 1: A nutshell of (statistics) stars

Welcome!

This is an introductory course on statistical modelling. Welcome!

The focus of this course is on how to specify a theoretical idea (possibly vague) in a testable statistical model.

PDF-Version

There's a PDF version of this book available. Note that the HTML is the more recent one.

Course description

Analyzing research data can broadly be classified in three parts: explorative data analysis, modeling (including inference), and visualization. Either part is pivotal in its own right, but it can be argued that modeling is at the core of the scientific endeavor. However, in practice, modeling, visualization, and data exploration is heavily intertwined, so that three parts may be recognized (as individual entities) but not usefully separated from each other. This idea provides the rationale of this course: Data exploration, data visualization and data modeling is discussed as an integrated framework.

The focus is on practical data analysis; theoretical concepts are, where mentioned, second class citizens due to time constraints and the didactic aims of the course.

For example, statistical inference – such as p-values and confidence intervals – are not more than touched briefly, as the instructor believes that modeling, not inference, is of prime importance for the auditorium.

We will use the R environment for all computations (freely available). Please bring your own Laptop with R and RStudio installed (installation guides are provided). Data and R code will be provided.

We're on a crash course

The course is set-up as a "crash course" which indicates that we'll rather try to cover a breadth of steps rather than digging deep at certain particular points. The rationale of this approach is that before digging deep, it is necessary to gain an overview of the territory. In addition, if one particular topic is not of interest to a given student (perhaps to difficult/simple), not much time is lost.

Be warned! Compare this crash course to a dancing crash course right before your wedding: A lot can be achieved by such a course in some instances, or rather, the worst consequences (of not knowing how to dance) may be fenced off, but one should not expect to be a dancing queen (king) thereafter.

More on modelling

Models and modeling are of pivotal importance in many sciences, not only for providing an explanation of nature en miniature (theoretical models), but also for gauging how closely the empirical data at hand match the theoretical model. Translating a theoretical model into statistical language is called statistical modeling and provides the guiding principle in this introductory course. Regression models will be presented as a lingua franca of statistical modeling, and we will learn that many empirical questions can (comfortably) be analyzed using a regression framework. Depending on the background and aims of the participants (and time permitting), we will shed light on some standard topics such as model comparison, classification models, and typical pitfalls. Given a more advanced auditorium, we may want to explore how causal and non-causal associations can be translated and tested using simple linear statistical models. Foundational ideas of statistical modeling will be accompanied by short examples and case studies to facilitate transfer and practical application after the course.

Course prerequisites

Basic computer usage knowledge is needed (downloading materials from the internet, operating a PC, etc). Basic R knowledge is needed. Basic knowledge of statistical concepts (such as descriptive statistics) is needed. Willingness to learn is essential.

Learning objectives

Upon successful completion of this course, students should be able to:

- select the right statistical visualization for a variety of data contexts
- "crunch" or "wrangle" data
- explain what statistical modeling means
- formulate basic statistical models
- differentiate between predictive and explanatory modeling
- apply the methods to own datasets

Course Literature

This course builds on the freely available e-book ModernDive. Each topic is paralleled by an ackompagnying chapter from ModernDive. A hard copy can be purchased here. The book is for sale in print here.

Course logistics

This course can be presented as a one-day seminar or split-up in four blocks.

The course can be held in English or German.

Please *bring your own computer* and *read the notes* regarding course logistics in advance. Note that some *upfront preparation is needed* from the learners.

R and RStudio[1] will be needed throughout the course. Please make sure that the IT is running. In case of technical difficulties with R feel free to use RStudio Cloud; free plans are available.

All learning materials (such as literature, code, data) will be provided in electronic format.

[1] Desktop version, not the server

UPFRONT student preparation

- *Install R and RStudio*, see ModernDive Chap. 1.1. In case you have your R running on your system, please make sure that you're uptodate. If outdated, download and install the most recent versions of the software. Similarly, hit the "Update" button in RStudio's "Packages" tab to update your packages if you have not done so for a couple of months.
- Sign-in at RStudio Cloud. It's super helpful because I as the techer can provide you with an environment where all R stuff is ready to use (packages installed etc).
- Install the necessary R packages as used in the book chapters covered in this course (see the sections on "Needed packages" in each chapter). If in doubt, see here the instructions on how to install R packages. Here's the actual list on the R packages we'll need.
- Students new to R are advised to learn the basics, see ModernDive, Chap 1.2 - 1.5
- Bring your own laptop
- Make sure your internet connection is stable and your loudspeaker/headset is working; a webcam is helpful.
- Students are advised to review the course materials after each session.
- I recommend that you carefully check the course description to make sure the course fits your needs (not too advanced/basic).

Didactic outline

This course can rather be considered a workshop in the sense that the instructor uses a dialogue-based approach to teaching and that there are numerous exercises during the course. Instead of providing long talks to the students, the instructor feels obligated to engage students in back-and-forth conversations. Similarly, the presentation of a large number of Powerpoint slide is avoided. Instead, a thorough course literature is available (free online), so that students will have no barrier in diving deeply into the materials and ideas presented. However, during class it is more important to transmit the pivotal ideas; details need to be read and worked by the students individually after (and before) the course. As an alternative to presenting a lot of text on slides, in this course there will be a (electronic) whiteboard where concepts are developed dynamically and in pace of the teaching conversation thereby adjusting the "dose" of new thoughts to the actual pace of the instruction.

Schedule

Overview on topics covered

- Data Visualization using the grammar of graphics and ggplot2
- Data Wrangling based on the tidyverse in R
- Basic concepts of statistical modelling
- Primer on causal inference
- Introduction to regression analysis

Block 1: Explorative Data Analysis

Visualization

- Data *visualization*, see ModernDive Chap. 2, and get the R code here
  - Exploring common types of statistical diagrams, the "5NG"
  - Discussing when (not) to use diagrams see Anscombe's Quartett, and when to use which one
  - Building elegant graphics in R

Data Wrangling

- Data *wrangling*, see ModernDive Chap. 3, and get the R code here
  - A taxonomy of typical data operations
  - How to perform common data operations with R
  - Summarizing data (aka computing descriptive statistics)

Exercises / Case study

- Exercises
  - Exercises on life expectancy.
  - Case study on the visualization of flight delays
  - Advanced case study on one hit wonders
  - Visualization covid cases
  - Case study on nominal data: Survival on the Titanic
  - Inspiration for own project: Visualize Covid-19 cases from this source.

Block 2: Statistical Modelling: Basic

Theory

- Basics of *modelling*, see ModernDive Chap. 5.0, and get the R code here
  - What is modelling?
  - Basic terminology
  - Prediction vs. explanation
- Some thoughts on *causal inference*, see ModernDive Chap. 5.3.1
- *Regression* with one numerical predictor, see ModernDive Chap. 5.1
- Regression with one categorical predictor, see ModernDive Chap. 5.2
- Assessing *model fit* (using (adjusted) R^2), see ModernDive Chap. 5.3.2
- For some tips and tricks on typical issues, see ModernDive tips and tricks

Case study

- Exercises/Case studies:
  - Prices of Boston houses, first part
  - Modeling movie succes, first part

Block 3: Statistical Modelling: Multiple Regression and interaction

Theory

- Slightly more advanced topics on linear regression such as *multiple regression and interaction*, see ModernDive Chap. 6, and get the R code here
- One numerical and one categorical predictor, see ModernDive Chap. 6.1
- Two numerical predictors, see ModernDive Chap. 6.2
- Simpson's paradox and more on causal inference, see ModernDive Chap. 6.3.3

Case study

- Exercises/Case studies:
  - Prices of Boston houses, second part
  - Modeling movie succes, second part
  - Modeling flight delays

Block 4: Project coaching

- This session is dedicated to work on real projects brought in by the students.
- In addition, open questions regarding the presented concepts are being discussed.

Instructor

Sebastian Sauer works as a professor at Ansbach university, teaching statistics and related stuff. Analyzing data to answer questions related to social phenomena is one of his major interests. He is trying to help raising the methological (and particularly statistical) skills in the sciences (ie., scientists). The programming language "R" is one of his favorite tools. He sees himself as a learner, and is particularly interested learning more on quantitative approaches to understand nature. Open Science is a hot topic to him. He hopes to contribute to pressing social problems such as populism by bringing in his statistical and psychological know-how. He writes a blog which serves as a sketchpad for stuff in his mind (not immune to thought updates) at https://data-se.netlify.app/. Sebastian is the author of "Moderne Datenanalyse mit R" (Sauer 2019). His publication list is available on Google Scholar.

Contact me

Feel free to contact me via email at `sebastiansauer1@gmail.com`.

Assessment and grades

There is no assessment, there are no grades!

Talk to me

It's my goal to make this an excellent course and a stimulating and enjoyable experience for all of us. So that I can find out if this is happening, I encourage feedback—be it positive or negative—on all aspects of the course at any time. For example, if something I'm doing is making it difficult for you to learn, then let me know before it's too late; if you particularly enjoyed something we did in class, say so so that we can do it again.

Course materials

Most of the materials as presented below is made available through the course book Modern-Dive. Please check the relevant chapters of the book before the course to make sure you have all materials available.

Licence

Check out the Github repo of this project.

Resources

Recommendations

- RStudio Cheatsheets, particularly on data wrangling, and data vizualization
- Book R for Data Science as a handy reference or a serious text book.
- Tidy Tuesday video series
- Post your open question on Stack Overflow.
- Follow `#rstats` hashtag on Twitter.

For students willing to learn more and go deeper (than the concepts explored in the present course), this book on regression modelling, and this book on statistical learning are recommended. For German folks, check out my book on modern data analysis.

Suggested literature for deepening the analytic skills include Statistical Rethinking. For an introduction to graphical causal models, check out Julia Rohrer's paper. For a more in-depth journey, consider reading this book. While I wholeheartedly recommend such books, we will not be able to discuss many of the ideas presented therein in class (in this course) due to time constraints.

R Packages

All R packages are accessible through the course book; please consult the relevant chapters. Please install all R packages used before the course. Here's a tutorial on how to install R packages.

The most important R packages for this course are:

- tidyverse
- easystats

The following packages are useful for data access (but not strictly mandatory):

- gapminder
- nycflights13
- fivethirtyeight
- skimr
- ISLR

For the Bayes models you'll need some extra software (free, save and stable), but somewhat more hassle to install. Using Bayes in this course is *optional*. You don't miss a lot if you don't use it.

- rstanarm

For the R package `{rstanarm}` to run, you'll need to install RStan. On Windows, this amounts to installing RTools. On Mac, you'll need to install the XCode CLI[2].

In sum, follow the instructions on the RStan website. It's unfortunately a bit complicated.

Data

All data are accessible through the course book; please consult the relevant chapters.

[2]possibly you need also a Fortran compiler, but maybe that's optional

Labs (case studies)

Practical data analysis skills can be practiced using these labs; in addition Chapter 11 provides two cases studies. Note that such content may be used as homework.

There are a lot of case studies scattered on the internet.

Sketching causal models

Dagitty is great tool for sketching causal graphs (DAGs), it can be usd in your browser or as R package. Here's an example of a collider bias. Check out this post for an intuitive explanation.

German introductary course

Readers who speak German may check out this Blitzkurs into data analysis using R.

Where are the slides?

There are none. I feel that slides are not optimal for learning. In class, slides can be detrimental if they are too wordy because that distracts from that the dialogue with the instructor, and I hold this very dialogue as essential. Outside of class, slides are neither helpful. Instead, a good book is much more beneficial, because in a book, there's enough room to patiently explain in sufficient details, an endeavor which is impossible for a slide deck.

To underline my messages to you, dear learners, I will use some sketches on a virtual white-board, some interactive apps, live coding, and some (pre-prepared) diagrams. That's a bit similar to what happens at Khan Academy. You might have noticed that many courses at Coursera follow a similar approach.

I readily confess that this approach is novel to many learners in these days, learners who are accustomed to hundreds of Powerpoint slides. Please be open and I think you will appreciate this didactic style.

Technical Details

Last update: 2022-10-09 19:19:08

```
sessioninfo::session_info()
```

```
- 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     2022-09-16
 pandoc   2.19.2 @ /usr/local/bin/ (via rmarkdown)

- Packages -----------------------------------------------------------------
 package     * version date (UTC) lib source
 cli           3.4.0   2022-09-08 [1] CRAN (R 4.2.0)
 colorout    * 1.2-2   2022-06-13 [1] local
 digest        0.6.29  2021-12-01 [1] CRAN (R 4.2.0)
 evaluate      0.16    2022-08-09 [1] CRAN (R 4.2.0)
 fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
 htmltools     0.5.3   2022-07-18 [1] CRAN (R 4.2.0)
 jsonlite      1.8.0   2022-02-22 [1] CRAN (R 4.2.0)
 knitr         1.40    2022-08-24 [1] CRAN (R 4.2.0)
 magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
 rlang         1.0.5   2022-08-31 [1] CRAN (R 4.2.0)
 rmarkdown     2.16    2022-08-24 [1] CRAN (R 4.2.0)
 rstudioapi    0.14    2022-08-22 [1] CRAN (R 4.2.0)
 sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
 stringi       1.7.8   2022-07-11 [1] CRAN (R 4.2.0)
 stringr       1.4.1   2022-08-20 [1] CRAN (R 4.2.0)
 xfun          0.33    2022-09-12 [1] CRAN (R 4.2.0)

 [1] /Users/sebastiansaueruser/Rlibs
 [2] /Library/Frameworks/R.framework/Versions/4.2/Resources/library

---------------------------------------------------------------------------
```

1 Goals in statistics

1.1 Overview

Many stories to be told. Here's one, on the goals pursued in statistics (and related fields), see Figure Figure **??**.

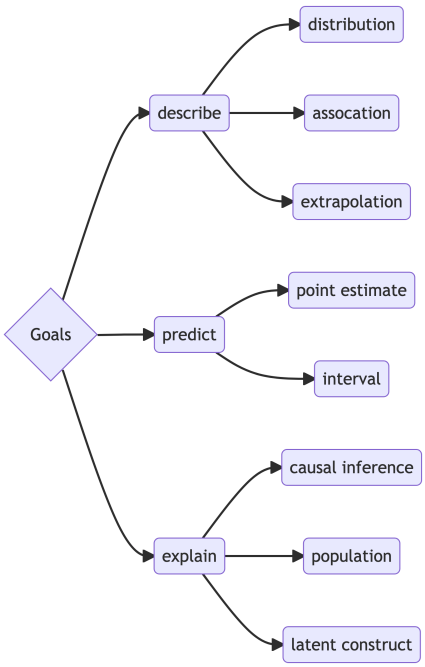

Figure 1.1: A taxonomy of statistical goals

> **Note**
>
> Note that "goals" do not exist in the world. We make them up in our heads. Hence, they have no ontological existence, they are epistemological beasts. This entails that we are free to devise goals as we wish, provided we can convince ourselves and other souls of the utility of our creativity.

1.2 Further reading

Hernán, Hsu, and Healy (2019) distinguish:

Hernán et al. (2019) distinguish:

- *Description*: "How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?"
- *Prediction*: "What is the probability of having a stroke next year for women with certain characteristics?"
- *Causal inference*: "Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?"

Gelman, Hill, and Vehtari (2021), chap. 1.1 proposes three "challenges" of statistical inference.

1.3 If nothing else helps

Stay calm and behold the infinity.