Stats for Data Science

Daniel Kaplan

2020-01-03

Contents

Orientation

These are notes for the MAA sponsored minicourse *Stats for Data Science* at the 2020 Joint Mathematics Meetings in Denver, Colorado.

- Part A, Thursday, 1:00 –3:00 pm
- Part B, Saturday, 1:00–3:00 pm.

Abstract

As universities and colleges rush to offer courses and even degree programs in data science, it's fair to wonder whether data science is genuinely new or is merely a rebranding of statistics. This mini-course will introduce participants to important and substantial ways that a statistics course that genuinely engages data science differs from traditional statistics.These include an emphasis on prediction, classification and causality rather than the traditional focus on estimation and significance. During the mini-course, we'll work through both theoretical and computational exercises from a new book, Stats for Data Science (available at https://dtkaplan.github.io/SDS-book/preface.html). The workshop is appropriate for anyone from a newcomer to statistical computing to experts. Some small groups in the mini-course will choose to use mouse-driven "Little Apps" to display data-science oriented statistical concepts. Others will choose to work with interactive R tutorials based on modern modeling and graphics packages in R. Participants should bring a laptop or tablet. All work will be browser based; there's no need to install new software.

Resources

- Participant notes, comments, suggestions, questions, etc.
- Participant introductions
- Books
  - *Stats for Data Science* textbook. The link is to the current draft of a textbook I am writing to explore how statistics can be taught

in a way that genuinely embraces the typical goals of data-science practice.

- *A Compact Guide to Classical Inference* by Daniel Kaplan. This book deals with one small but important part making a mental and teaching transition from a conventional intro stat course into a course suitable for data science. In particular, the *Compact Guide* approaches statistical description using model functions and, with this basis, unifies and simplifies the inferential settings typically covered in inferential stats. All those traditional settings–difference of means and of proportions, simple regression, inference on contingency tables, one-way analysis of variance, two-way analysis of variance, multiple regression–are translated into a single test statistic, F, with a simple formula and simple interpretations. For instance, statistical "significance"[1] is addressed by the simple question, is $F > 4$. Confidence intervals on differences and slopes are all shown to have the same form, proportional to $1 \pm \sqrt{4/F}$.
- *Statistical Inference via Data Science* by Chester Ismay and Albert Y. Kim. For the introductory-course instructor who is not shy of using R with her class and who wants to touch on non-statistical aspects of data science such as data wrangling, this can be good choice for a textbook. The statistical topics are conventional, but the book wisely leaves the mean-median-mode stuff for an appendix. In the sense that the presentation of statistics is based on regression, I see this book as a kind updating for data science and recent developments in R of *Statistical Modeling.* (See next entry). *Statistical Inference* is not as radical as *Stats for Data Science*, but for many instructors that's probably a good thing. There are exercises ("learning checks") and solutions.
- *Statistical Modeling: A Fresh Approach* by Daniel Kaplan was my attempt, circa 2010, to re-imagine what can be done in an introductory statistics course to make the course more relevant to genuine practice, take confounding seriously, and provide room for student creativity in framing statistical questions. So, instead of "do I use t or chi-squared?" the question becomes "what covariates are relevant and what are the implications of including them in a statistical analysis?"
- *Computer Age Statistical Inference by Bradley Efron and Trevor Hastie. This is a concise review of classical statistical inference that is much broader than the* Compact Guide* and particularly oriented to deep theoretical limitations of classical inference and a couple of generations of work to overcome those limitations.
- *The Book of Why* by Judea Pearl and Dana Mackenzie. This is a

[1] Recently, Jeff Witmer at Oberlin College suggested replacing the misleading technical use of an everyday word with an utterly different meaning: significance. His suggestion is "discernible" and "discernibility", as in "the difference is statistically *discernible*" or "one part of inference is statistical *discernibility*".

fantastic introduction to causal inference which, yes, does go beyond the pat "correlation is not causation" or "no causation without experimentation."
  - *The Theory that Would not Die* by Sharon Bertsch McGrayne. The subtitle is "How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy" which aptly describe the book's historical approach. This isn't a textbook, but it is a good way to see why Bayes is important.
  - *Modern Data Science with R* by Ben Baumer, Daniel Kaplan, and Nicholas Horton. This book covers a wide range of data science techniques, but wouldn't be suitable for a *statistics* course.
  - *R for Data Science* by Garrett Golemund and Hadley Wickham. Like *Modern Data Science with R*, it's not a suitable book for a statistics course. But it's an excellent (even canonical) choice to make sense of the recent generation of R data-science tools.
- Little Apps. These are web-based apps that provide statistical computing capabilities *without coding*. There are, of course, many other apps provided to the stat-ed community such as the Lock[5] StatKey collections and Dan Adrian's Happy Apps.
  1. Functions and F statistics. This is the Little App written specifically for the *Compact Guide*. It also happens to be the prototype for the next generation of Little Apps that are mobile-device ready.
  2. Regression models is a pre-cursor to the *Functions and F statistics* Little App. It explains the idea of *model values*, which are simply the values of a statistical model evaluated using the training data as input.
  3. Resampling and Bootstrapping demonstrates these ideas graphically.
  4. A few other Little Apps, developed as part of <StatPREP.org>, cover topics of the traditional intro course such as t-tests, the normal distribution and center and spread.
- Computing tutorials
  1. GET THE LIST FROM STATPREP.org. MAYBE WRITE ONE FOR THE COMPACT GUIDE?

Chapter 1

Stats and Data Science

1.1 The parable of the highway.

Pictures of Denver in 1910 and a highway exchange in 1980+.

Car pictures.

1.2 What's different about data science?

- Prediction
- Large data sets
- Multiple "tests" – e.g. batting averages
- Causality
- **Decision making**
  - integrate the information from data into a broader framework
  - Examples:
    - Screening versus diagnostic tests
    - Fuel economy. Not "Is fuel economy different at different speeds?" but "How different is it and what are the implications of this for my decision?"

Chapter 2

Core statistical tools

This will be paper-and-pencil introductions to the tools.

Note that I'm being much more mathy here than I would in teaching a typical class. The audience here is professional mathematicians, hence likely not too scared by algebraic notation.

2.1 Graphics

- Data graphics
- Density graphics
- Interval graphics

2.2 Models and effect sizes

Summarizing a relationship with a function

We'll start with models with a single *degree of flexibility*, that is $^\circ\mathcal{F} = 1$. This includes all the settings covered in most introductory stats courses.

2.3 Variance

Average pairwise square differences between values.

$$\frac{1}{n(n-1)} \sum_{i \neq j} |x_i - x_j|^2 = 2 \text{ Var}(x)$$

2.4 Basic discernibility

1. Is there any discernible relationship between the response and explanatory variables revealed by the model?
   - Inputs from the model: v_r, v_m, n, and degrees of flexibility $^\circ\mathcal{F}$
   - Output:
   $$\mathrm{F} = \frac{n - (^\circ\mathcal{F} + 1)}{^\circ\mathcal{F}} \frac{v_m}{v_r - v_m}$$
   - Interpretation: Is $\mathrm{F} \gtrapprox 4$?. Then a relationship is discernible.
2. Given a *base* model and a proposed *elaboration* of that model, does the elaboration reveal new aspects of the relationship between the response and explanatory variables?
   - Inputs from the model:
     - v_r and n
     - v_m^{base} and v_m^{elab},
     - degrees of flexibility $^\circ\mathcal{F}^{base}$ and $^\circ\mathcal{F}^{elab}$
   - Output:
   $$\Delta\mathrm{F} = \frac{n - (^\circ\mathcal{F}^{elab} + 1)}{^\circ\mathcal{F}^{elab} - ^\circ\mathcal{F}^{base}} \cdot \frac{v_m^{elab} - v_m^{base}}{v_r - v_m^{elab}}$$
   - Interpretation: Is $\Delta\mathrm{F} \gtrapprox 4$? Then a relationship is *discernible*.[1]
   - Notes:
     - The special case of a model with $^\circ\mathcal{F} = 0$ is called the *Null Model* and corresponds to the claim that there is no relationship between the explanatory variables and the response variable. In this special case, $\mathrm{F} = \Delta\mathrm{F}$.
     - $\Delta\mathrm{F} \neq \mathrm{F}^{elab} - \mathrm{F}^{base}$

2.5 Confidence intervals (when $^\circ\mathcal{F} = 1$)

When $^\circ\mathcal{F} = 1$, there is only one explanatory variable and the modeling situation is one of these:

- difference between two groups
- slope of a regression line

Either way, there is only one effect size: the difference or slope.

- Inputs:
  - Effect size B
  - F
- Output:

[1] Recall that I'm using *discernible* as a replacement for *significant*, as proposed by Jeff Witmer.

  - Margin of error is $\pm B\sqrt{4/F}$
- Interpretation:
  - We wouldn't be at all surprised if a much, much bigger study revealed an effect size within the confidence interval. - If we are comparing our study to another study, we're only justified in claiming a contradiction when the two confidence intervals don't overlap.

  - Do we really need to refer to populations?

Note that when $^\circ\mathcal{F} \geq 2$, there is either more than one explanatory variable or more than one group in that explanatory variable or a non-straight-line regression (e.g. a polynomial). In none of these cases can the margin of error be deduced directly from F due to one or more of:

- effect size not constant
- multiple effect sizes
- collinearity among explanatory variables

Instead of the simple formula based on F, confidence intervals can be based on a regression table or bootstrapping.

Chapter 3

Computation

3.1 No coding

Infrastructure: a browser. Can work on a smart phone.

The Little Apps

3.2 Coding

Infrastructure: a browser. Need a tablet+-sized screen and a keyboard.

RStudio tutorials

3.3 Blank canvas

Reproducible tools: Rmd

Chapter 4

Bayes (and decision making)

Chapter 5

Machine learning

Cross validation

Bootstrapping

Chapter 6

Causality (and decision making)

Chapter 7

More aspects to data science

Data science is not merely a *rebranding* of statistics.

The scenario where statisticians would have come to lead the development of data science is plausible, but historically computer scientists and people from fields such as genetics, marketing, public health, medicine, remote sensing, etc. have played crucial roles.

Whether or not data science ought to be considered part of the mathematical sciences, any genuine approach should be fundamentally based in realistic applications and the actual kinds of problems–especially decision making–that data science is used to address.

For concise introductions to wrangling and visualization, see *Statistical Inference via Data Science* by Chester Ismay and Albert Y. Kim or *Data Computing* by Daniel Kaplan and Matthew Beckman.

I don't know of a concise introduction to *decision making* from the statistics, mathematics, or computer science perspectives. (Please tell me if you do know of one!) But if you are willing to wade into the business literature, you would do well with *How to Measure Anything* by Douglas Hubbard. It even has a workbook.