StatPREP

MAA MATHEMATICAL ASSOCIATION OF AMERICA

AMATYC OPENING DOORS THROUGH MATHEMATICS

ASA

February 11, 2019

News You Can Use

WHAT'S COMING UP?

We have a number of webinars planned to help you with using StatPREP materials or to better understand some of the technological tools available to help you run a more data-centric course. Several previous webinars are archived on our website and can be viewed here.

We are pleased to announce the first StatPREP webinar of the 2019 academic year!

Title: MAA StatPREP Webinar 4: Using the "What's Normal?" Class Lesson to Introduce the Normal Distribution

Date: Feb. 22, 2019

Time: 1:00 pm EST, 12:00 pm CST, 11:00 am MST, 10:00 am PST

Duration: 1 hour

How to join: Please click the link below to register for the webinar:

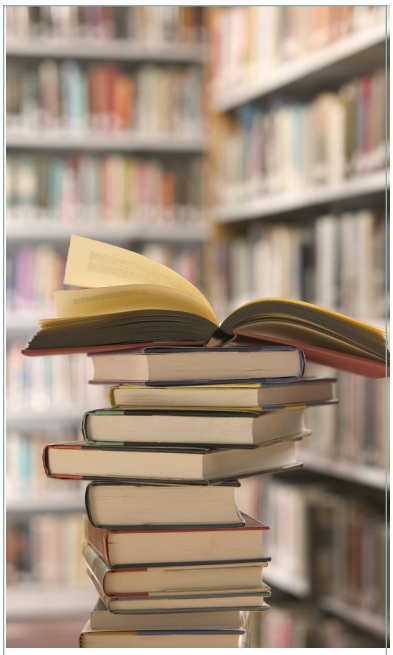
https://zoom.us/webinar/register/WN_BB4SnC3bT52lQ3-Gq8MUqw

- iPhone one-tap US: US: +16468769923,,139906737# or +16699006833,,139906737#
- Telephone: US: Dial (for higher quality, dial a number based on your current location):
  US: +1 646 876 9923 or +1 669 900 6833 or +1 408 638 0968

  Webinar ID: 532 717 173

- **Webinar Summary:**

This webinar will present a classroom lesson called "What's Normal?" that models data which can be used to explain the normal distribution to your class. The lesson shows the connection between data and the empirical rule of the normal curve.



QUOTE:

"Data is real when it has at least 1000 rows, at least 5 variables, and was not initially collected with a primary purpose of teaching statistics."

RESOURCES:

The ASA statistics education guidelines are published at:

http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

WORDS FROM DANNY

Use real data. That's a simple but profound recommendation from the American Statistical Association's guidelines for statistics instruction. In the November 2018 issue of this newsletter, I presented the operational definition of "real data" developed at StatPREP's Oct. 2018 meeting at the American Mathematical Association:

Data is real when it has at least 1000 rows, at least 5 variables, and was not initially collected with a primary purpose of teaching statistics.

As appropriate for an operational definition, it is readily applied to any data set and provides a yes-or-no answer. There's no call to be fanatical about the numbers 1000 and 5; anything close will do. But the common textbook practice of using on the order of 10 rows and only one or two variables is not close.

Now it's time to talk about the meaning of the word *"use"* in the phrase "use real data." Much of the data that appears in textbooks is intended to provide numbers for statistical calculations. Any other numbers could be used instead. But there are far richer ways to make data central to a lesson or exercise, ways that involve the actual handling of data, the search for patterns or anomalies, data display, and the interpretation of data in its actual context.

- The data should be "raw." That is, students should work directly with the data themselves rather than with a statistical summary. Statistics like means and proportions are not data but *summaries of data.*
- The data should be in a professional form. It's important for students to learn about the conventional ways in which data are organized, the most important of which is as a spreadsheet like table, with a columns for each variable and a row for each "unit of observation." They should see examples of good practice, for instance putting meta-data such as measurement units and descriptions in a separate codebook file.
- The data should be rich enough to reward exploration and encourage hypothesis *formation*.

As an example, consider polling data. I'm not talking about the summaries of polling data that we see in the news and textbooks, such as p-hat = 48% and n = 893. Instead, think about the raw data that lies beneath such summaries. As an example, this CSV file contains the 578 survey responses collected in a poll prior to a 2017 special election for a vacated congressional seat in Montana. The political choice of the person surveyed is just one of the variables. Others include the sex and age-group of the respondent, his or her location in the state, the number of milliseconds it took to give a response to the question, the time-of-day of the response. There's also a demographic "weight" for the respondent, which can support a conversation about sampling bias. (How was the weight calculated? How might the weight be used?) You can do all the usual calculations of confidence intervals on proportion, but there are also questions to be answered such as how age groups differ or even whether Republicans or Democrats respond to the survey at different times of day.

A data skeptic might ask, "In the end, doesn't it all boil down to

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

Not insofar as you are answering questions about the real world. I found the Montana data in a news commentary about fake polls (https://fivethirtyeight.com/features/fake-polls-are-a-real-problem/) where the issue was how to recognize fraud by examining closely the raw data for inconsistencies or unrealistic clumping. In today's world,

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

is only a small part of the story.

The ASA statistics education guidelines are published at:

http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

WHERE DO I FIND IT?

If you are looking for information about how to do something, there is a lot of information on http://statprep.org. In addition, many people are interested in learning about various aspects of teaching with real data. In future additions of this newsletter, we will list some valuable resources that you can consult to learn even more about teaching with data and the technology needed to do it.

WHO WANTS TO CHAT?

The best way to work on curriculum change and on incorporating new ideas and tools is to talk to other people that are doing the same thing. A significant goal of StatPREP is to establish online communities of statistics instructors who are actively engaged in sharing ideas, successes, and failures. We strongly encourage StatPREP participants to reach out to the other members of their regional hubs by email. **Simply use a reply all to any email sent from your hub leader. Feel free to include any member of the StatPREP leadership team on the email you send.**

REVIEW: STATS FOR DATA SCIENCE

BY DANIEL KAPLAN

There is a lot of excitement about the "new" field of data science. Universities are starting up degree programs, colleges are offering data-science courses, and there is a movement to bring data-science certificates and degrees to two-year colleges.

The enthusiasm for data science has many statisticians grumbling. "'Data science' is just a rebranding of statistics." I think it's much more than that. Statistics should be a central component of data science, but data science has many themes that are not well represented in statistics.

Notice that I said, "Statistics *should be* ..." rather than "statistics *is* a central component ...." The use of the present tense is not yet appropriate because the statistics we teach is not yet appropriate to play a meaningful role in data science. Data science is about using data to inform decisions and actions. Consider two basic tasks in data science: prediction and classification. In statistics, however, we focus on description and significance. For instance, just about every intro stats course introduces the confidence interval on the sample mean. That's a form of description. It's certainly not a legitimate form for a prediction since it doesn't tell us the likely range of future events. The statistics emphasis on significance makes sense when our concern is whether we have enough data to support a claim. But in data science, the issue is how to *use the data we have* to do the best job possible to guide decisions. And usually, in data science, the data we have is so large that statistical significance becomes a formality. In any event, the statistical techniques for significance taught in intro stats don't deal with the forces that drive false discovery.

In order to make statistics genuinely central to data science, we have to draw on the many elements of statistical practice and theory that relate to the goals of data science. These do not include the focus on methods for small data (e.g. n = 10) that figure so prominently in intro stats (e.g. the t test). Or consider a procedure that's ubiquitous in intro stats: the chi-squared test for a relationship between two factors. Chi-squared tells nothing about the shape of the relationship nor the practical size of the relationship. And then there's the question of causality. Intro stats is in the no-causality zone, which isn't much use when the problem is to inform a decision as best can be done with the data already at hand.

All this is by way of motivation for a new book, *Stats for Data Science*. The aim of this book is to embed statistics in the context of data science: prediction, prediction error, and the selection of useful predictors; causality; models; and so on. But it's not appropriate to review the book here: it's still in draft form and ... I wrote it. Instead, I encourage you to look through it and report your reactions and impressions to me. Undoubtedly, one of your reactions will be shock. *Stats for Data Science* is unlike any statistics textbook you have taught with and you may wonder, "Is this statistics?" Certainly it isn't the statistics that you're teaching now or are likely to be teaching in the near future. But, I think, it is where statistics education should be heading in order to stay engaged with the contemporary uses of data encompassed by data science.

The draft book is available online at https://dtkaplan.github.io/SDS/preface.html. Whether your reaction is positive or negative, it will be useful to me to hear it.