StatPREP

MAA MATHEMATICAL ASSOCIATION OF AMERICA

AMATYC OPENING DOORS THROUGH MATHEMATICS

ASA

November 1, 2018

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 also have more live webinars coming up, so keep your eyes out for the next one!

WHO'S WHO?

Are you looking for specialized help, but don't know who to contact? Check out our comprehensive list below that shows "who's who" on the StatPREP Leadership team!

- ◊ Need help with R Studio? Contact Danny Kaplan (kaplan@macalester.edu)
- ◊ Need help with Little Apps? Contact Danny Kaplan (kaplan@macalester.edu)
- ◊ Need help finding a data set? Contact Danny Kaplan (kaplan@macalester.edu)
- ◊ Need help incorporating StatPREP materials into a lesson? Contact Kate Kozak (kathryn.kozak@coconino.edu)
- ◊ Need help assessing a StatPREP lesson? Contact Ambika Silva (Ambika.silva@canyons.edu)
- ◊ Have a question about the StatPREP website? Contact Jenna Carpenter (carpenter@campbell.edu)
- ◊ Want to know more about StatPREP webinars? Contact Kate Kozak (kathryn.kozak@coconino.edu)
- ◊ Want to know more about an upcoming StatPREP Workshop? Contact Mike Brilleslyper (mike.brilleslyper@usafa.edu)

WHO'S WHO LEADERSHIP TEAM

Mike Brilleslyper, Air Force Academy

Jenna Carpenter, Campbell University

Danny Kaplan, Macalester College

Kathryn Kozak Coconino College

Donna LaLonde, ASA

Ambika Silva College of the Canyons

Rachel Levy MAA

HUB LEADERS

Joe Roith, St. Catherine's University, Minneapolis, MN (2017-18)

Ambika Silva, College of the Canyons, Santa Clarita, CA (2017-18)

Helen Burn, Highline College, Seattle, WA (2018-19)

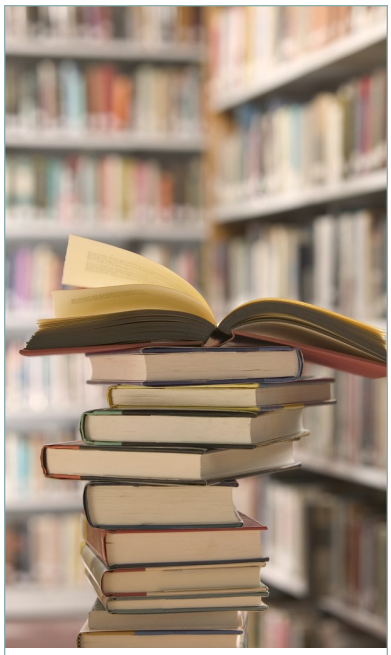
Hwayeon Ryu, University of Hartford, Hartford, CT (2018-19)

Carol Howald, Howard Community College, Columbia, MD (2019-2020)

Thomas Kinzeler, Tarrant County College, Fort Worth, TX (2019-2010)

Rona Axelrod, Florida SW State College, Fort Myers, FL (2020-2021)

Brooke Orosz, Essex Community College, Newark, NJ (2020-2021)



COOL DATA RESOURCES

Did you know that people tend to be local optimists and national pessimists (no matter which country they live in)? That people in richer countries work less, yet productivity increases as working hours decrease? That the number of US households that use social media exceeds the number of households that have a land line? For an enormous amount of data about everything from culture and politics to health and food (and lots in between), check out **Our World in Data** at: https://ourworldindata.org

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.**

WORDS FROM DANNY

At October's StatPREP meeting at the Mathematical Association of America's DC headquarters the new MAA deputy executive director, Rachel Levy asked a simple question: What's real data?

A core recommendation of the American Statistical Association's GAISE report is to "use real data" when teaching statistics. Prof. Levy wasn't looking to prompt a philosophical discussion of the nature of reality, but to define a benchmark. If a widely lauded, consensus report from the world's leading organization of statisticians calls for every introductory course to use real data, we need a way for instructors to know, for sure, whether they are in compliance. And so we discussed what is "real" when it comes to teaching statistics using real data. Our conclusion:

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.

How did we come up with this definition? In part, we looked at the examples of "real data" in the GAISE report, for instance a dataset on housing with 2930 rows and 80 variables, or a dataset on 53,940 diamonds with 10 variables. But mainly, we looked at the reasons motivating the recommendation to teach with real data: which practices are encouraged and which discouraged. These are: teach statistics as an investigative process, foster active learning, give students experience with multivariate thinking, use technology but focus on concepts.

Why 1000 rows? Working with data on this scale requires using appropriate technology, the sort used in the data workplace. Graphics with 1000 points can be rich enough to see relationships, even when there are multiple variables. And with 1000 rows, a central concept in statistical reasoning, sampling variation, can be shown directly using random selection.

Why 5 variables? "Multivariate" is at least three and there are three basic roles played by variables in data analysis: response variable, explanatory variable(s), covariate(s). But we need more than 3 because both categorical and quantitative variables can star in any of these roles. And we need room for students to explore actively which can be as simple as letting them choose which variables to relate to which.

Of course we understand that there is no hard statistical boundary between n=999 and n=1000, just as there is no hard boundary at p=0.05.

Now that you have precise criteria for the "real" in "use real data," our next task will be to define "use."

A REVIEW OF *THE PROCESS OF POLLING*

In 1996 my department chair handed me the first statistics textbook I had ever seen. That single gesture constituted my college's faculty development program for teaching statistics. One of the earliest examples in the book was about the importance of random sampling. It included a picture of President Truman holding up the Chicago Tribune's infamous "Dewey Defeats Truman" headline. It's a good story, but hardly timely, having taken place 48 years earlier. Few of my students knew who Truman was and none of them knew anything about Dewey.

Our students have grown up in an era of "scientific" polling. Being scientific, the results are reported with a margin of error, often ±3 percentage points, to help us know when conclusions are warranted and when not. Many of our statistics courses feature units on constructing a margin of error on a sample proportion, often with explicit reference to political polls. But, like Dewey defeating Truman, the story is no longer timely. The "error" in the "margin of error" is now only a small part of the unreliability of polls. Why?

In an unprecedented opening up of the process of polling, The New York Times is letting us observe, live, their polling for the 2018 mid-term elections. You'll find a description of the project in a September 2018 column and the live action here. It's worth watching.

For those of you reading this after the polling ends, I'll describe the action. As I write this, 2,070,469 telephone calls have been made. In each Congressional district, the results from the past calls are laid out in a long line of circles, filled red or blue depending on the the recipient's response. But only 1 or 2% of the dots are filled. The large majority are empty: no response. Each new call generates a wiggling box at the head of the line of dots. It wiggles until the end of the call. Almost always, the box turns into an unfilled circle.

The poll I'm watching now, New Jersey 3rd district, is in its early stage. 4250 calls producing 62 responses. The margin of error? There's a simple but meaningful statement laid right on top of the grayed-out tally so far: "Don't take this poll seriously until we reach at least 250 people. We're at 62."

The calls are made based on a random selection from the phone numbers known to be in the district. But the random selection hardly generates a random sample when the response rate is 2%. To get something that resembles the population, pollsters weight their results. The New York Times is weighting "by age, party registration, gender, likelihood of voting, race, education and region, mainly using data from voting records files compiled by L2, a nonpartisan voter file vendor." And then there's the "likely voter" model, an informed guess about what fraction of people in each weighting strata will actually vote. There's a detailed explanation in this article on the site, where the faulty results from the 2016 presidential election are attributed to a failure to weight by education level.

Seeing the polling process in such detail reveals our misconceptions about what's important in statistics. The so-called "margin of error" is not an adequate indicator of the reliability of the poll. Instead, we need to be thinking about the factors used in weighting and the extent to which they capture the current configuration of political schisms. Polls are now about big, multivariable data (the "voting records compiled by L2") and building models of turnout based on previous elections.