SDSS 2019: Bellevue, WA

Data Science & the pedagogical reform of introductory statistics

Brendan Patrick Purdy, PhD
Professor of Mathematics
Moorpark College, California
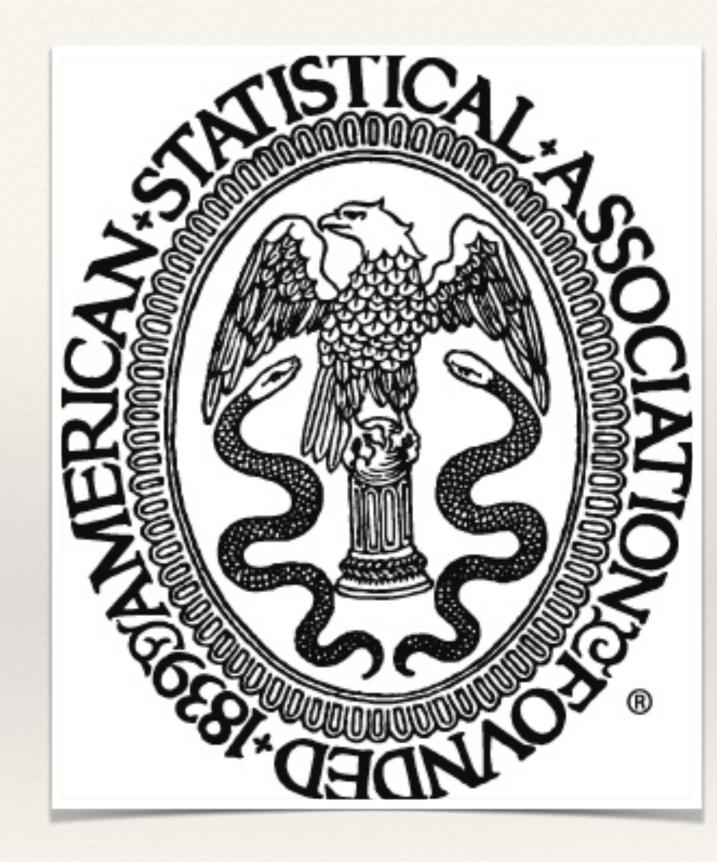MS-ACMS: Data Science, Class of 2021
University of Notre Dame

Overview

❖ American Statistical Association the last half-decade

❖ Data Science & Teaching Introductory Statistics

❖ The Pedagogical Reform of Introductory Statistics

ASA the past half-decade

- ❖ GAISE College Report (2016)

- ❖ *p*-values (2016-17)

- ❖ TYCDSS (2018)

- ❖ SDSS (2018, 2019)

GAISE

Guidelines for Assessment and Instruction
in Statistics Education (GAISE)
College Report 2016

Committee:

Robert Carver (Stonehill College), Michelle Everson, co-chair (The Ohio State University), John
Gabrosek (Grand Valley State University), Nicholas Horton (Amherst College), Robin Lock (St.
Lawrence University), Megan Mocko, co-chair (University of Florida), Allan Rossman (Cal Poly
– San Luis Obispo), Ginger Holmes Rowell (Middle Tennessee State University), Paul Velleman
(Cornell University), Jeffrey Witmer (Oberlin College), and Beverly Wood (Embry Riddle
Aeronautical University)

Citation: GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in
Statistics Education College Report 2016," http://www.amstat.org/education/gaise.

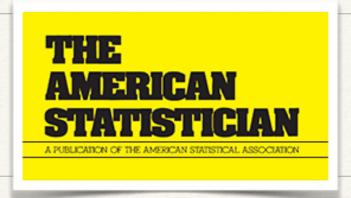Endorsed by the American Statistical Association

❖ *Guidelines and Instruction in Statistics Education (GAISE) College Report*, 2016

1. Teach statistical thinking.

   a. Teach statistics as an investigative process of problem-solving and decision making.

   b. Give students experience with multivariable thinking.

2. Focus on conceptual understanding.

3. Integrate real data with a context and purpose.

4. Foster active learning.

5. Use technology to explore concepts and analyze data.

6. Use assessments to improve and evaluate student learning.

p-values

- ❖ ASA's Symposium on Statistical Inference (SSI) in 2017.

- ❖ *The American Statistician,* Volume 73 (2019)

p-values

❖ American Statistical Association's (ASA's) <u>Statement on p-values</u> in 2016.

1. *p*-values can indicate how incompatible the data are with a specified statistical model.

2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

TYCDSS

- Two-Year College Data Science Summit (TYCDSS), 2018
  - Career Transfer Education (CTE)
  - Associate Degree
  - Transfer Degree
- At Moorpark College, we are in the beginning process of starting a Data Science CTE.

SDSS

- Symposia on Data Science & Statistics (2018-2019)

  - The interface of statistics, computing science, and data visualization.

    - SDSS 2018: "Beyond Big Data: Leading the Way"

    - SDSS 2019: "Beyond Big Data: Building Data Tools"

An autobiography

Data Science & Teaching Introductory Statistics

- ❖ How is a "Classical" Introductory Statistics at the College level taught (in opposition to the GAISE recommendations)?

- ❖ Two exogenous pressures from Data Science

- ❖ CA AB 705

"Classic" Introductory Statistics

If the GAISE recommendations (and the ASA's statement on p-values) represent the right pedagogy, then this is the wrong pedagogy:

- Artificial data

- Pointless computations

- Archaic tables

- Ridiculous "bright-lines"

- Privation of theory

- Ignorance of context (to include history)

- Treating statistics like a calculus class

- No data science

This is how I taught introductory statistics *circa* 2006, before I put away childish things…

Two exogenous pressures from Data Science

- Two exogenous pressures on the Math Department at my Community College from Data Science that are forcing a pedagogical reform of our introductory statistics course:

  - Data Science CTE

  - The needs of other departments that we serve

Data Science CTE

- Moorpark College is in the process of creating a Data Science CTE (based on the attendance of my colleague and myself at TYCDSS).

  - Currently, our standard Introductory Statistics course will serve both the Data Science students and everyone else.

  - In order to properly prepare Data Science students for the Data Science CTE curriculum, we need to follow the GAISE recommendations.

Needs of other departments

- Albeit the Mathematics Department is very territorial with respect to our Department teaching the Introductory Statistics class on campus, some of the disciplines that we serve (e.g. psychology and sociology) have expressed a strong desire to create and teach an introductory statistics for their students.

  - While the Mathematics Department is correctly worried about rigor, the other Departments are correct that we are not teaching statistics a way that benefits their students as much as it should.

    - Emphasizing aspects of data science, viz. large data sets, the visualization of data, and multivariate data.

Another pressure: AB 705

- California AB 705: ends the use of placement tests for transfer level English and mathematics courses, but instead uses "multiple-measure;" and in particular high school transcripts.

  - As a result, at our College we are drastically reducing the number of pre-transfer level classes.

- In our Department, we are creating supplementary and bridge classes to help under-prepared students succeed in their transfer level mathematics courses.

  - This is leading to the creation of new curriculum for these Supplementary and Bridge courses, and as a result allowing us to modify the curriculum for our Introductory Statistics (C-ID).

Pragmatic concerns

Pedagogical Reform of Statistics

- ❖ Challenges we're currently facing

- ❖ What we are doing

Challenges we're currently facing

Ch. 9: Confidence Intervals (two populations)

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

$$\text{where } E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \quad (\text{Indep.})$$

$$\text{where } E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (df = \text{smaller of } n_1 - 1, n_2 - 1)$$

σ_1 and σ_2 unknown and not assumed equal)

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \quad (df = n_1 + n_2 - 2)$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

σ_1 and σ_2 unknown but assumed equal)

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$(\sigma_1, \sigma_2 \text{ known})$

- Personnel
  - Uncompromising full-time faculty
  - Uninvolved part-time faculty
  - Transient administration
- Material
  - Antediluvian textbooks
  - Privation of technology resources

What we are doing

- At Moorpark College, we are doing the following to succeed in spite of the challenges faced

  - Statistics Syndicate

  - Curriculum

  - Professional Development

  - Two introductory textbooks, and the changing of the "default" text.

My Introductory Statistics Course

❖ Theory and concepts

❖ Simulations

❖ Interpreting output (from statistical software)

❖ Messy, large data

❖ Multivariate thinking

❖ No bright lines with the p-value

❖ Historical context as appropriate, to include the use of historical data

❖ Understanding the relationship between confidence intervals and hypothesis tests, and the assumptions of each type of statistical inference

Last Slide

❖ Thank you for your time.

❖ Questions, commets, &c.?

❖ Thanks to my colleague Professor Tom Ogimachi

  ❖ Support from the California Community Colleges Chancellor's Office

    ❖ Strong Workforce Grant

    ❖ Guided Pathways Grant

❖ Contact Brendan:

  ❖ brendan_purdy@vcccd.edu