Rethinking Introductory Statistics

Today's introductory statistics course is an adequate foundation for analyzing small random samples. But it is inadequate for the needs of students in 2040. Those students need to focus more on causation, confounding and coincidence in quasi-experiments, observational studies, big data, administrative data and multivariate data. Rethinking introductory statistics is not just a change in pedagogy or assessment; it is also involves a change in audience, goals and content. Now is the time to rethink this course or to think of offering an alternative course. Once there is agreement on the direction, goals, topics and assessment in the new course(s), then the real work begins in retraining more than 5,000 faculty teaching introductory statistics.

BACKGROUND:

Today's introductory statistics course is focused on small samples involving random selection (surveys) or random assignment (clinical trials). In each case the goal is demonstrate the movement from statistical independence to p-values and statistical significance. This course provides immediate value for those in biostatistics, psychology and market research. It is an excellent foundation for those who will analyze studies based on small random samples.

Today's world is different and tomorrow's world will be more so. Big data is everywhere. Big data can be defined as data so large that all associations are statistically significant.

Traditional introductory statistics courses:

1. Deal primarily on random variation in small random samples. They ignore the fact that as the number of records increases, coincidence increases.
2. Deal primarily with random selection (surveys) and random assignment (clinical trials). There is little said about quasi-experiments or observational studies. Yet quasi-experiments and observational studies are now common in articles in JAMA and Nature.
3. Deal with one and two variables. Analyzing observational data or quasi-experiments requires a multivariate approach. Three variables are the minimum to deal with confounding. Big data and administrative data typically involve multiple variables.
4. Avoid dealing with confounding. Most intro statistics textbooks do not contain any entry for confounding, confounder or lurking variable. When dealing with big data, the focus is on forming associations and building models. In predicting, Simpson's paradox and confounding are irrelevant; in identifying causal connections, they are central.
5. Make no mention of the Cornfield conditions: necessary conditions for a confounder to nullify or reverse an association in an observational study or a quasi-experiment.
6. Are silent on any aspect of data science: acquiring, cleaning, merging and summarizing big data.
7. Avoids dealing with observational causation by saying "association is not causation." Yes, association is not sufficient for causation, but association is typically a sign of causation somewhere. Identifying associations (two-group comparisons or two-variable covariation) is the cornerstone of all scientific inquiry.

Confounding is omnipresent in observational studies. Students need to understand how to control for (to take into account) the influence of measured confounders and how to evaluate the susceptibility of an association to confounders that are unknown.

What should undergraduate STEM education look like in 2040 and beyond to meet the needs of students, science, and society?

In order to meet the needs of students, science and society, undergraduate STEM education must offer introductory statistics courses that deal with big data, coincidence, confounding and causation. This can be done by revising the existing course or by creating a companion course. Steps are being taken in fielding a confounder-based statistical literacy course. These involve the University of New Mexico approving Statistical Literacy as satisfying a mathematics requirement in the core curriculum and in General Education.

A confounder-based course is not the only possible direction. Two recent conferences in 2020 both focused on what parts of data science might or should be taught.

What should we do now to prepare?

NSF should encourage statistical educators to think broadly: to address three distinct audiences: majors in experimental and survey-based disciplines, majors in observational disciplines (epidemiology and geology as well as business, sociology and social psychology, and students in non-quantitative majors.

NSF should hold forums and fund projects on rethinking introductory statistics. What should be taught? Should it be taught in one class or two? How should it be taught? How should it be assessed? How should teachers be retrained?

By itself, retraining teachers is a monstrous task.

> Monstrous in number: In 2016-17, there were 1,956,000 new US bachelor's degrees granted[1]. Of these, an estimated 50% (one million) took introductory statistics annually.[2] If the average class size ranges from 25 to 50, then intro statistics involves 20,000 to 40,000 sections annually. If statistical educators average two to four intro statistics classes per year, then the number of teachers ranges from 5,000 to 20,000. Retraining as many as 10,000 teachers is a major challenge.
>
> Monstrous in effort: Many of these faculty are not statisticians. They are teaching the same course that they took in college. Retraining them will be extremely difficult.

Updating introductory statistics may be one of the largest tasks ever attempted by the NSF. The discipline has disregarded many if not most of the changes recommended by leading statistical educators.[3]

Though the task may seem almost overwhelming, it is well-worth the effort.

Proposal submitted by Dr. Milo Schield. PhD in Space Physics, Rice University
Professor (Tenure) in Dept of Business & MIS at Augsburg College;
Consultant to the Math-Stat Department, University of New Mexico, Albuquerque
Fellow, American Statistical Association (ASA); Elected Member: International Statistical Institute (ISI)
President of the National Numeracy Network (NNN);
US Representative of the International Statistical Literacy Project (ISLP)
Editor and webmaster of www.StatLit.org. 340,000 visits (460,000 downloads) in 2018.
Schield publications at www.StatLit.org/Schield-Pubs.htm

[1] https://nces.ed.gov/fastfacts/display.asp?id=37

[2] Schield, M. (2008). Quantitative Literacy and School Mathematics: Percentages and Fractions: a chapter in *Calculation vs. Context: Quantitative Literacy and Its Implications for Teacher Education*. Steen and Madison, Eds. Mathematical Association of America (MAA). Copy at www.statlit.org/pdf/2008SchieldMAA.pdf

[3] Schield, M. (2013). Statistics Education: Steadfast or Stubborn? ASA Proceedings of the Section on Statistical Education. Copy at www.statlit.org/pdf/2013-Schield-ASA.pdf