Letter to a Young Statistician: On 'Student' and the Lanarkshire Milk Experiment

Stephen T. Ziliak

The history of science is a laboratory for learning experimental science and statistics. Imagine Pasteur without a Priestley counterpoint, or Ronald Fisher without William S. Gosset, aka "Student."

I discovered the value of history at school—the school of hard knocks. Many years ago, when I was a college student at Indiana University taking Introduction to Statistics, I imagined Student's *t*-table was designed for students to use; the faculty, the professors of economics and biometrics and statistics, I figured, were probably using a different table of probabilities for testing hypotheses.

Pasteur is a fascinating case in point for why we need history of science to inform the present. As Gerald L. Geison documents in *The Private Science of Louis Pasteur*, Pasteur kept two contradictory notebooks on his vaccination work treating rabies in dogs by inoculation. One of Pasteur's notebooks contained notes on his immunological treatment of dogs according to expectations Pasteur knew to be held by fellow scientists and a scrutinizing French public. The other notebook,which Pasteur kept hidden, contained details on the 'experiments' as they in fact occurred—or did not occur—ironically. His dog data was fudged. Yet Pasteur proceeded to inoculate multiple times a now famous Alsatian schoolboy, Joseph Meister, bitten by a "mad dog."

The history of science does more than shine a light on present ideas. When uncertainty about hypotheses is both large and important, what we can glean from the context and details of previous experiments leading to "present knowledge" is a path to enlightenment in the future. In *The Foundation of Statistics*, Leonard J. Savage sums up the false distinction between "experimental" and "observational" statistics:

> Finally, experiments as opposed to observations are commonly supposed to be characterized by reproducibility and repeatability. But the observation of the angle between two stars is easily repeatable and with highly reproducible results in double contrast to an experiment to determine the effect of exploding an atomic bomb near a battleship. All in all, however useful the distinction between observation and experiment may be in ordinary practice, I do not yet see that it admits of any solid analysis.

Astronomers know this. Einstein started a book club in college. His first book of choice to discuss with colleagues was Karl Pearson's *The Grammar of Science*. But this laboratory of ideas—in our case, the history of experimental statistics and methods—gets shoved aside by most graduate programs. Einstein's book club is laughing from the grave because, eclipsed from public view—like Pasteur's notebook—there is often a relevant myth or falsehood about the content and context of actual discovery, its value and credit, and—unless exposed and scrutinized—the whole society loses. For example, until 2008, the truth about the methodological and philosophical differences between William Gosset and Ronald Fisher was hiding for more than 70 years in the basement of the Guinness Storehouse in Dublin. Controversies over statistical methods are like pickles in the icebox: We keep those 'in the back of the back' and rarely speak of them.

Louis Pasteur

Cold dark storage works for pickling cucumbers and science,

William Sealy Gosset (1876-1937), aka "Student," in 1908, the year he published Student's table and test of significance.

too. Neglecting and censoring alternative experimental philosophies, past and present, impedes scientific progress and stunts imagination. Journals now look for a one-size-fits-all method; grantors do, too, confusing students and frustrating natural economists such as Student.

The problematic method is randomization plus statistical significance. Search the bookshelves of the average economist or other social scientist and one is likely to spot (besides cute animal photos) lots of frequentist textbooks, all recommending the same basic methodological cure: randomization plus significance equals validity. Though based on a false equation and several false elements—such as statistical significance being a necessary condition—the usual cure has become a disease, and especially so in my home field of economics. Other frequently superior methods, such as Guinnessometrics, are neglected. The same can be said of a host of related methods, Bayesian and other, that have proven to be more precise, profitable, and ethical than randomization plus significance.

Recently, three development economists (two at MIT and one at Harvard) were awarded the Nobel Prize in Economic Sciences for their contribution to "methodology" and poverty eradication. Their methodological contribution is neocolonial econometrics. They were rewarded for traveling overseas to conduct large, one-off, randomized controlled trials on thousands upon thousands of impoverished and largely disenfranchised people of color living in the tropics in an elaborate attempt to prove the obvious. They are financed in multiples of millions of dollars by, for example, the International Monetary Fund, World Bank, and many private foundations such as Bill and Melinda Gates. Claiming equipoise, they use blank placebos for controls (even when studying mosquito-borne illness and intestinal worms in African school children). They make a big show of statistically significant results ($p < .05$) to assert, for instance, that wearing corrective eyeglasses can help children with refractive conditions do better at school, as one study did. Funded partially by J-PAL Poverty Lab at MIT, a large, randomized controlled trial was conducted on 19,000 impoverished school children with refractive conditions in rural China, and the results were recorded in "Visualizing Development: Eyeglasses and Academic Performance in Rural Primary Schools in China."

Helen Keller said, "It is a terrible thing to see, and have no vision." Some of the data were dropped, however, representing three townships. Local officials and compassionate teachers took pity upon the children struggling to read and gave surplus eyeglasses to some of the 'controls' in those townships. But by design, the children randomly selected to be in the control group were to be denied corrective lenses of any kind, yet still followed and coded. That's not experimental science or economic development striving to eradicate poverty.

Cementing a methodology into science while burying the diamonds and disputes of the past, or trying to, is a social problem. "Inefficient," "imprecise," and "unethical" are not commendable virtues. Statistical methods should shine a light on well-being, not dim the lights and sabotage well-being, as they did in the Chinese eyeglass experiment.

Bringing history back won't save us all. History contains some, not all, of the answers to scientific and ethical questions. But by systematically ignoring history in our teaching and scholarship, valuable trials and answers are eclipsed from view, creating waste, redundancy, screwball ethics, and the illusion of a gold standard methodology when no such standard exists. Ancient Babylonian astronomers used more than one instrument to study the stars. But a graduate student in economics or statistics is upbraided for suggesting we add a "study of the past" to better advance the present.

Just do it. That is my first point. Be entrepreneurial, take a chance, and study the history of scientific method as the greats did and do. Emulate statistical scientists like my friend and epidemiologist Ken Rothman, for example, who eats books for breakfast, or my late friend, Arnold Zellner, the founder of Bayesian econometrics. Zellner was a close student of science biography, and he allowed it to affect his work.

Study the history of fields near to and seemingly uncorrelated with your own. Reverend Joseph Priestley was an experimental chemist and political economist who, as an amateur home brewer, was rather successful in his late-night hobby by accidentally discovering oxygen and carbon dioxide.

Priestley put whiskers on a small sample, paired test. He made his discoveries by placing live mice inside little glass vials and stringing the

vials by wire from a ceiling to dangle over the home brewer's mash tun, one mouse per vial. (No one is recommending that now, thankfully, but stay with me.) He dangled a control group of live mice nearby from the same ceiling, but a distance away from the wide open tun of fresh, fermenting beer. Next morning, all the mice that had dangled between the ceiling and the tun were deceased. "Fixed air," Priestley causally inferred.

Inspired by Priestley, Antoine Lavoisier, a French chemist and economist, dipped his nose deeply into small sample wine experiments and emerged with the balanced equation of chemistry and principle of the conservation of mass. Randomization, large numbers, and a test of statistical significance were neither necessary nor sufficient to make these fundamental scientific discoveries. (The lesson instead seems to be if you wish to make fundamental scientific discoveries, stay close to beer, wine, and spirits!)

Gosset would not be surprised. Unlike Priestley, Gosset—who is much better known by his pen name, Student of Student's *t*-table and test of statistical significance—was a professional beer brewer and a lifer at Guinness' Brewery in Dublin. Educated as a scholar at Winchester College and Oxford, New College, in chemistry and the natural sciences, and taking a minor at Oxford with honors in math mods., the co-inventor of modern statistics would not see any statistical theory at all until he got to Guinness in 1899.

Gosset was hired as apprentice brewer to bring quantitative methods to bear on the work of the rapidly expanding brewery. From 1899 to 1907, he was essentially an advanced graduate student at Guinness working on full scholarship and excellent salary whose charge was to quantify whatever it was the largest brewery in the world had been doing since 1759 and to advise on what to do next.

He read on his own about estimation and the errors of observation in a book by George Airy, the royal astronomer. More importantly, for about 30 years, Gosset gained practical knowledge and experimental sophistication traveling between Dublin and Reading to work on the farm with Edwin S. Beaven.

Beaven, author of *Barley: Fifty Years of Observation and Experiment*, was a barley farmer and experimental maltster who worked on commission with Guinness to test and supply new cereal varieties. Gosset learned first-hand and early on that his practical knowledge of the soil mattered. He learned from Beaven that perceptible differences in soil fertility affecting barley yield were found to exist in one yard "cage" experiments as much as in large plot uniformity trials; literally speaking, Beaven and Gosset found the soil changes inch by inch. That insight, and some comparisons balanced with random experiments, led to the ABBA design: If A=Irish Archer and B= English Archer variety, the layout of the field is Irish–English–English–Irish/Irish–English–English–Irish, etc. The most precise allocation is not random, Student found. Random designs had been tried and rejected by Student in 1905 in favor of systematic balancing.

The humble Student insisted Beaven co-invented the "maximum contiguity," "pairing," and "twinning" designs as used now in field experiments in economics and biometrics, together with Student's table and test. When Fisher asked Gosset in a letter about the origins of small sample pairing in statistics, Gosset replied that credit goes to "Old Noah," captain and leader of Noah's Ark! (Beaven refused to accept credit; it was Gosset all the way, Beaven insisted.) Regardless, the Gosset–Beaven, Gosset–Hunter, and Irish Department of Agriculture experiments from the early 1900s to the 1940s helped turn Irish and English barley into Europe's highest yielding. Gosset was certain about one source of that success as he wrote in "On Testing Varieties of Cereals"—the secret to the art of good design begins with highly correlated material:

> The art of designing all experiments lies even more in arranging matters so that ρ [the correlation coefficient] is as large as possible than in reducing σ^2_x and σ^2_y [the variance]. The peculiar difficulties of the problem lie in the fact that the soil in which the experiments are carried out is nowhere really uniform; however little it may vary from eye to eye, it is found to vary not only from acre to acre but from yard to yard, and even from inch to inch. This variation is anything but random [Student noted], so the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

From history, in other words, one can imagine and consider completely different designs, contexts, and approaches to experimental statistics. After all, sometimes the experimental method itself, however popular and ubiquitous, is the problem, as it was with the infamous Lanarkshire Milk Experiment.

In 1930, a report was published on "milk consumption and the growth of schoolchildren" by Gerald Leighton and Peter L. McKinlay. Few readers were as prepared as Student was to comment. The report concerned a nutritional experiment on 20,000 school children in Lanarkshire, which was one of the most impoverished regions of Scotland during the Depression. The children involved in the experiment were found in 67 schools. For four months, 5,000 children received three-fourths a pint of raw milk, 5,000 received the same amount of pasteurized milk,

and 10,000 acted as controls, blank placebo; they got no milk at all. Some schools got only raw milk, some got pasteurized, but no school got both. The allocation was selected by alphabet or ballot and, as Student found, teachers showing pity on the smallest, poorest, or most nutritionally challenged children (as we saw in the Chinese eyeglass experiment). The milk experiment, in other words, was neither stratified and balanced nor randomized and balanced.

Authors of the report and many commentators, including Fisher and S. Bartlett, concluded from the experiment that, however imperfect the experimental design, raw milk contributes more to the growth of school children. Student disagreed. He wrote to Fisher and Pearson, arguing back and forth on various points. Student, Beaven, Guinness, and Irish and English agriculture had all reaped the benefits of small sample "pairing" and "twinning" ever since the early 1900s. Student expressed admiration for the intent of the ambitious nutritional study, but he sharply disagreed with the design, analysis, and interpretation of results. He observed that in a sample of 20,000 students, it would be easy to find 50 pairs of identical twins.

Lanarkshire students represented a wide range of social and economic backgrounds, but that heterogeneity played no part in the formal design or analysis. There was no stratification. Students were weighed with their clothes on, but they were measured in winter and again in warmer months, making no adjustment for the weight of clothing worn nor the magnitude of relative poverty. Gosset's main point was that much time and money could be saved, and much precision could be gained, by choosing pairs of identical twins and then "flipping" (a coin, say) to determine who gets raw and who pasteurized. And, if possible, weigh them with their clothes off.

Using Gosset's method, the Scotland Ministry of Health could run from one to 200 repeated experiments in Lanarkshire rather than the large, expensive, and imprecise one-off they conducted. Few researchers would bother doing even 10 replications using Gosset's method, and there'd be no need to.

Gosset's economical twinning method would also help answer Pearson's second objection—that underfed children at birth and in youth will sometimes later accelerate in growth, moving them closer to the average weight and height of their cohort. Gosset's flexible and economical approach would also answer Pearson's third objection—how do we know we can extrapolate from twins to nontwins? Using Student's method, one can afford to run parallel "twin" and "pairing" experiments over time, yielding intergenerational panel data on child nutrition, growth, and such for each new class or cohort.

Gosset was puzzled by the government's decision to let 10,000 children go with no milk at all. We already agree milk is beneficial, so why deny school children their daily glass? Or Chinese school children their daily vision? My cowriter, Edward Teather-Posadas, and I make the same point about penicillin and the tragic, decades-long Tuskegee syphilis experiment in *The Unprincipled Randomization Principle in Economics and Medicine*.

Pearson was impressed, a bit embarrassed, and deeply challenged. His ego was raw, and he barked at Gosset in a series of letters. He could not bear to lack a good, definitive answer, once again, to the brewer's attack on large sample biometrics. Pearson could not admit that with a single flash of insight, Gosset had shored up the lion's share of errors caused by heterogeneity, lack of randomness, and other moving variables in the controllable environment. The rest of the variance, the rest of the noise, Gosset figured, was random, as best as current knowledge could say. This emphasis placed by the brewer on reducing "real error" and not merely random was fundamental and easy to understand. It was also practical and economical. Therefore, Pearson and Fisher refused to accept it.

Gosset's Guinnessometrics remains a valid approach to experimental philosophy and decisions. COVID researchers and some in angiogenesis are starting to catch on, and there is a small revival of Guinnessometrics sprouting up in agriculture and pharmacy. Three decades of profit-driven work on small sample, stratified, representative, balanced, economic, and independently repeated experiments had uniquely prepared Student to solve the spoiled milk experiment in 1931.

In running the largest brewery in the world, Student needed a million-dollar method, so to speak, to help him judge. He needed a simple toolkit he could use quickly and effectively to make a profitable judgment on the spot, or nearly so. At the same time, he sought to develop a set of tools precise and robust enough for external validity in the farmer's and brewer's sense; that is, for making prudent gambles on a big, new harvest of barley and hops. Small sample analysis and twinning have an economic origin and purpose. Pearson struggled to understand the point, and Fisher refused to accept it. The best tools did not yet exist. Gosset had to invent them. C

About the Author

Stephen T. Ziliak is professor of economics at Roosevelt University and lead author of The Cult of Statistical Significance, a foundational text for the reform of significance testing and decisions in science, law, and society. He discovered Guinnessometrics in 2008, when he helped establish the Gosset Archive at the Guinness Storehouse in Dublin, Ireland. He is, with ASA Executive Director Ron Wasserstein and others, a major contributor to the ASA "Statement on Statistical Significance and *P*-Values," and he co-edited a special issue of *The American Statistician*, "Scientific Inference in the 21st Century: A World Beyond $p < 0.05$."