Science is heroic, with a tragic (statistical) flaw

Mindless use of statistical testing erodes confidence in research

By Tom Siegfried 11:29am, July 2, 2015

Just as Achilles' heel was an unguarded weakness that ultimately brought down a hero of Greek mythology, dogmatic devotion to traditional statistical methods threatens public confidence in the institution of science.

First of two parts

Science is heroic. It fuels the economy, it feeds the world, it fights disease. Sure, it enables some unsavory stuff as well — knowledge confers power for bad as well as good — but on the whole, science deserves credit for providing the foundation underlying modern civilization’s comforts and conveniences.

But for all its heroic accomplishments, science has a tragic flaw: It does not always live up to
the image it has created of itself. Science supposedly stands for allegiance to reason, logical rigor and the search for truth free from the dogmas of authority. Yet science in practice is largely subservient to journal-editor authority, riddled with dogma and oblivious to the logical lapses in its primary method of investigation: statistical analysis of experimental data for testing hypotheses. As a result, scientific studies are not as reliable as they pretend to be. Dogmatic devotion to traditional statistical methods is an Achilles heel that science resists acknowledging, thereby endangering its hero status in society.

And that's unfortunate. For science is still, in the long run, the superior strategy for establishing sound knowledge about nature. Over time, accumulating scientific evidence generally sorts out the sane from the inane. (In other words, climate science deniers and vaccine evaders aren't justified by statistical snafus in individual studies.) Nevertheless, too many individual papers in peer-reviewed journals are no more reliable than public opinion polls before British elections.

Examples abound. Consider the widely reported finding that typographical ugliness improves test performance compared with questions printed in a nice normal font. It’s not true, as biological economist Terry Burnham has thoroughly established.

More emphatically, an analysis of 100 results published in psychology journals shows that most of them evaporated when the same study was conducted again, as a news report in the journal Nature recently recounted. And then there's the fiasco about changing attitudes toward gay marriage, reported in a (now retracted) paper apparently based on fabricated data.

But fraud is not the most prominent problem. More often, innocent factors can conspire to make a scientific finding difficult to reproduce, as my colleague Tina Hesman Saey recently documented in Science News. And even apart from those practical problems, statistical shortcomings guarantee that many findings will turn out to be bogus. As I've mentioned on many occasions, the standard statistical methods for evaluating evidence are usually misused, almost always misinterpreted and are not very informative even when they are used and interpreted correctly.

Nobody in the scientific world has articulated these issues more insightfully than psychologist Gerd Gigerenzer of the Max Planck Institute for Human Development in Berlin. In a recent paper written with Julian Marewski of the University of Lausanne, Gigerenzer delves into some of the reasons for this lamentable situation.

Above else, their analysis suggests, the problems persist because the quest for “statistical significance” is mindless. “Determining significance has become a surrogate for good research,” Gigerenzer and Marewski write in the February issue of Journal of Management. Among multiple scientific communities, “statistical significance” has become an idol, worshiped as the path to truth. “Advocated as the only game in town, it is practiced in a compulsive, mechanical way — without judging whether it makes sense or not.”

Commonly, statistical significance is judged by computing a P value, the probability that the observed results (or results more extreme) would be obtained if no difference truly existed between the factors tested (such as a drug versus a placebo for treating a disease). But there are other approaches. Often researchers will compute confidence intervals — ranges much like the margin of error in public opinion polls. In some cases more sophisticated statistical testing may be applied. One school of statistical thought prefers the Bayesian approach, the standard method’s longtime rival.
Gigerenzer and Marewski emphasize that all these approaches have their pluses and minuses. Scientists should consider them as implements in a statistical toolbox — and should carefully think through which tool to use for a particular purpose. “Good science requires both statistical tools and informed judgment about what model to construct, what hypotheses to test, and what tools to use,” Gigerenzer and Marewski write.

Sadly, this sensible advice hasn’t penetrated a prejudice, born decades ago, that scientists should draw inferences from data via a universal mechanical method. The trouble is that no such single method exists, Gigerenzer and Marewski note. Therefore surrogates for this hypothetical ideal method have been adopted, primarily the calculation of P values. “This form of surrogate science fosters delusions and borderline cheating and has done much harm, creating … a flood of irreproducible results,” Gigerenzer and Marewski declare.

They trace the roots of the current conundrum to the “probabilistic revolution,” the subversion of faith in cause-and-effect Newtonian clockwork by the roulette wheels of statistical mechanics and quantum theory.

Ironically, the specter of statistical reasoning first surfaced in the social sciences. Among its early advocates was the 19th century Belgian astronomer-mathematician Adolphe Quetelet, who applied statistics to social issues such as crimes and birth and death rates. He noted regularities from year to year in murders and showed how analyses of populations could quantify the “average man” (l’homme moyen). Quetelet’s work inspired James Clerk Maxwell to apply statistical methods in analyzing the behavior of gas molecules. Soon many subrealms of physics adopted similar statistical methods.

For a while the statistical mindset invaded physics more thoroughly than in the social, medical and psychological sciences. “With few exceptions, social theorists hesitated to think of probability as more than an error term in the equation observation = true value + error,” Gigerenzer and Marewski write.

But then, beginning about 1940, statistics conquered the social sciences in the form of what Gigerenzer calls the “inference revolution.” Instead of using statistics to quantify measurement errors (a common use in other fields), social scientists began using statistics to draw inferences about a population from tests on a sample. “This was a stunning new emphasis,” Gigerenzer and Marewski write, “given that in most experiments, psychologists virtually never drew a random sample from a population or defined a population in the first place.” Most natural scientists regarded replication of the experiment as the key to establishing a result, while in the social sciences achieving statistical significance (once) became the prevailing motivation.

What emerged from the inference revolution was the “null ritual” — the practice of assuming a null hypothesis (no difference, or no correlation) and then rejecting it if the P value for observed data came out to less than 5 percent (P < .05). But that ritual stood on a shaky foundation. Textbooks invented the ritual by mashing up two mutually inconsistent theories of statistical inference, one devised by Ronald Fisher and the other by Jerzy Neyman and Egon Pearson.

“The inference revolution was not led by the leading scientists. It was spearheaded by humble nonstatisticians who composed statistical textbooks for education, psychology, and other fields and by the editors of journals who found in ‘significance’ a simple, ‘objective’ criterion for deciding whether or not to accept a manuscript,” Gigerenzer and Marewski write. “Inferential
statistics have become surrogates for real replication.” And therefore many study “findings” fail when replication is attempted, if it is attempted at all.

Why don’t scientists do something about these problems? Contrary motivations! In one of the few popular books that grasp these statistical issues insightfully, physicist-turned-statistician Alex Reinhart points out that there are few rewards for scientists who resist the current statistical system.

“Unfortunate incentive structures … pressure scientists to rapidly publish small studies with slapdash statistical methods,” Reinhart writes in *Statistics Done Wrong*. “Promotions, tenure, raises, and job offers are all dependent on having a long list of publications in prestigious journals, so there is a strong incentive to publish promising results as soon as possible.”

And publishing papers requires playing the games refereed by journal editors.

“Journal editors attempt to judge which papers will have the greatest impact and interest and consequently those with the most surprising, controversial, or novel results,” Reinhart points out. “This is a recipe for truth inflation.”

Scientific publishing is therefore riddled with wrongness. It’s almost a miracle that so much truth actually does, eventually, leak out of this process. Fortunately there is enough of the good side of science to nurture those leaks enough so that by the standards of most of the world’s institutions, science still fares pretty damn well. But by its own standards, science fails miserably.

*Follow me on Twitter: @tom_siegfried*

Source URL: https://www.sciencenews.org/blog/context/science-heroic-tragic-statistical-flaw?mode=blog&context=117&tgt=nr