even 0.00003-percent level if there's a 10-percent or 25-percent (or more) chance that the data is gravely undermined by a laboratory error? Even the most ironclad findings of statistical validity lose their meaning when dwarfed by the specter of error. Or worse yet, fraud.

Nevertheless, even though statisticians warn against the practice, a one-size-fits-all finding of statistical significance is all too often taken as a shortcut to determine whether or not an observation is credible or a finding is "publishable." As a consequence, the peer-reviewed literature is littered with statistically significant findings that are irreproducible and implausible—absurd observations with effect sizes orders of magnitude beyond what's even marginally believable.

The concept of "statistical significance" has become a quantitative crutch for the essentially qualitative process of whether or not to take a study seriously. Science would be much better off without it.

SCIENTIFIC INFERENCE VIA STATISTICAL RITUALS

GERD GIGERENZER

Psychologist; director, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin; author, Risk Savvy

As a young man, Gottfried Wilhelm Leibniz had a beautiful dream: to discover the calculus that could map every single idea in the world into symbols. Such a universal calculus would put an end to all scholarly bickering. Every passionate *Edge* discussion, for one, could be swiftly resolved by dispassionate calculation. Leibniz optimistically estimated that a few skilled people should be able to work the whole thing out in five years. Yet nobody, Leibniz included, has yet found that Holy Grail.

Nonetheless, Leibniz's dream is alive and thriving in the social and neurosciences. Because the object of the dream has not been found, ersatz objects serve in its place. In some fields it's multiple regression, in others Bayesian statistics. But the champ is the null ritual:

1. Set up a null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your own research hypothesis.
2. Use 5 percent as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < .05$, $p < .01$, or $p. < .001$, whichever comes next to the obtained p-value.
3. Always perform this procedure.

Not for a minute should anyone think this procedure has much to do with statistics proper. Sir Ronald Fisher, to whom it has been wrongly attributed, in fact wrote that no researcher should use the same level of significance from experiment to experiment. The eminent statisticians Jerzy Neyman and Egon Pearson would roll over in their graves if they knew about its current use. Bayesians, too, have always detested p-values. Yet open any journal in psychology, business, or neuroscience and you're likely to encounter page after page with p-values. To give just a few illustrations: In 2012, the average number of p-values in the *Academy of Management Journal*, the flagship empirical journal in its field, was 116 per article, ranging between 19 and 536! Typical of management, you might think. But 89 percent of all behavioral, neuropsychological, and medical studies with humans published in 2011 in *Nature* reported p-values only—without considering effect size, confidence interval, power, or model estimation.

A ritual is a collective or solemn ceremony consisting of actions performed in a prescribed order. It typically involves sacred numbers or colors, delusions to avoid thinking about why one is performing the actions, and fear of being punished if one stops doing so. The null ritual contains all these features.

The number "5 percent" is held sacred, allegedly telling us the difference between a real effect and random noise. In fMRI studies, the numbers are replaced by colors, and the brain is said to light up.

The delusions are striking. If psychiatrists had any appreciation of statistics, they would have entered these aberrations into the *Diagnostic and Statistical Manual of Mental Disorders*. Studies in the U.S., U.K., and Germany show that most researchers don't (or don't want to) understand what a p-value means. They con-

fuse it with the probability of a hypothesis—that is, p(Data|Ho) with p(Ho|Data)—or with some other bit of wishful thinking, such as the probability that the data can be replicated. Startling errors are published in top journals. For instance, an elementary point is that in order to investigate whether two means differ, we should test their difference. What shouldn't be done is to test each mean against a common baseline, such as: "Neural activity increased with training ($p < .05$) but not in the control group ($p > .05$)." A 2011 paper in *Nature Neuroscience* presented an analysis of neuroscience articles in *Science, Nature, Nature Neuroscience, Neuron,* and *The Journal of Neuroscience* and showed that although seventy-eight articles did as they should, seventy-nine used the incorrect procedure.

Not performing the ritual can provoke great anxiety, even when it makes absolutely no sense. In one study (the authors' names are irrelevant), Internet participants were asked whether there was a difference between heroism and altruism. The great majority felt so: 2,347 respondents (97.5 percent) said yes; 58 said no. What did the authors do with that information? They computed a chi-square test, calculated that $c2(1) = 2178.60$, $p < .0001$, and came to the astounding conclusion that there were indeed more people saying yes than no.

One manifestation of obsessive-compulsive disorder is the ritual of compulsive hand washing even if there is no reason to do so. Likewise, researchers adhering to the null ritual perform statistical inferences all the time, even in situations where there's no point—that is, when no random sample was taken from a population, or no population was defined in the first place. In those cases, the statistical model of repeated random sampling from a population doesn't even apply and good descriptive statistics is called for. So even if a significant p-value

has been calculated, it's not clear what population is meant. The problem isn't statistics but its mistaken use as an automatic inference machine.

Finally, just as compulsive worrying and hand washing can interfere with the quality of life, the craving for significant p-values can undermine the quality of research. Which it has: Finding significant theories has been largely replaced by finding significant p-values. This surrogate goal encourages such questionable research practices as selectively reporting studies and conditions that "worked," or excluding data after looking at their effect on the results. According to a 2012 survey in *Psychological Science* of some 2,000 psychologists, over 90 percent admitted to having engaged in at least one of these or other questionable research practices. This massive borderline cheating in order to produce significant p-values is likely more harmful to progress than the rare cases of outright fraud. One harmful outcome is a flood of published but irreproducible results. Genetic and medical research using Big Data has encountered similar surprises when trying in vain to replicate published findings.

I don't mean to throw out the baby with the bathwater; statistics offers a highly useful toolbox for researchers. But it's time to get rid of statistical rituals that nurture automatic and mindless inferences. Scientists should study rituals, not perform rituals themselves.

THE POWER OF STATISTICS

EMANUEL DERMAN

Professor of financial engineering, Columbia University; former head, Quantitative Strategies Group, Equities Division, Goldman Sachs; author, Models.Behaving.Badly

I grew up among physicists, whose *modus operandi* is to observe the world, experiment with it, develop hypotheses and theories and models, suggest further experiments, and use statistics to analyze the results, thereby comparing mental imaginings with actual events. Statistics is simply their tool for confirmation or denial.

But nowadays the world, and especially the world of the social sciences, is increasingly in love with statistics and data science as a source of knowledge and truth itself. Some people have even claimed that computer-aided statistical analysis of patterns will replace our traditional methods of discovering the truth—not only in the social sciences and medicine but in the natural sciences, too.

We must be careful not to get too enamored of statistics and data science and thereby abandon the classical methods of discovering the great truths about nature (and man is nature, too). A good example of the classical power is Kepler's 17th-century discovery of his second law of planetary motion, which is in fact less a law than the recognition and description of a pattern. Kepler's second law states that the line between the sun and a moving planet sweeps out equal areas in equal times. This deep symmetry of planetary motion implies that the closer a planet is to the sun, the more rapidly it moves along its orbit.