Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

By Tom Siegfried

March 27th, 2010; Vol.177 #7 (p. 26)

But in practice, widespread misuse of statistical methods makes science more like a crapshoot.

It's science's dirtiest secret: The "scientific method" of testing hypotheses by statistical analysis stands on a flimsy foundation.

Statistical tests are supposed to guide scientists in judging whether an experimental result reflects some real effect or is merely a random fluke, but the standard methods mix mutually inconsistent philosophies and offer no meaningful basis for making such decisions.

Even when performed correctly, statistical tests are widely misunderstood and frequently misinterpreted.

As a result, countless conclusions in the scientific literature are erroneous, and tests of medical dangers or treatments are often contradictory and confusing.

Replicating a result helps establish its validity more securely, but the common tactic of combining numerous studies into one analysis, while sound in principle, is seldom conducted properly in practice.

In fact, if you believe what you read in the scientific literature, you shouldn't believe what you read in the scientific literature.

"There is increasing concern," declared epidemiologist John Ioannidis in a highly cited 2005 paper in PLoS Medicine, "that in modern research, false findings may be the majority or even the vast majority of published research claims."

Ioannidis claimed to prove that more than half of published findings are false, but his analysis came under fire for statistical shortcomings of its own. "It may be true, but he didn't prove it," says biostatistician Steven Goodman of the Johns Hopkins University School of Public Health. On the other hand, says Goodman, the basic message stands. "There are more false claims made in the medical literature than anybody appreciates," he says. "There's no question about that."

the standard statistical system for drawing conclusions is, in essence, illogical. "A lot of scientists don't understand statistics," says Goodman. "And they don't understand statistics because the statistics don't make sense."

How could so many studies be wrong? Because their conclusions relied on “statistical significance,” a concept at the heart of the mathematical analysis of modern scientific experiments.

Fisher first assumed that fertilizer caused no difference — the “no effect” or “null” hypothesis. He then calculated a number called the P value, the probability that an observed yield in a fertilized field would occur if fertilizer had no real effect. If P is less than .05 — meaning the chance of a fluke is less than 5 percent — the result should be declared “statistically significant,” Fisher arbitrarily declared, and the no effect hypothesis should be rejected, supposedly confirming that fertilizer works.

Fisher’s P value eventually became the ultimate arbiter of credibility for science results of all sorts — whether testing the health effects of pollutants, the curative powers of new drugs or the effect of genes on behavior. In various forms, testing for statistical significance pervades most of scientific and medical research to this day.

But in fact, there’s no logical basis for using a P value from a single study to draw any conclusion. If the chance of a fluke is less than 5 percent, two possible conclusions remain: There is a real effect, or the result is an improbable fluke. Fisher’s method offers no way to know which is which. On the other hand, if a study finds no statistically significant effect, that doesn’t prove anything, either. Perhaps the effect doesn’t exist, or maybe the statistical test wasn’t powerful enough to detect a small but real effect.

“That test itself is neither necessary nor sufficient for proving a scientific result,” asserts Stephen Ziliak, an economic historian at Roosevelt University in Chicago.

Soon after Fisher established his system of statistical significance, it was attacked by other mathematicians, notably Egon Pearson and Jerzy Neyman. Rather than testing a null hypothesis, they argued, it made more sense to test competing hypotheses against one another. That approach also produces a P value, which is used to gauge the likelihood of a “false positive” — concluding an effect is real when it actually isn’t. What eventually emerged was a hybrid mix of the mutually inconsistent Fisher and Neyman-Pearson approaches, which has rendered interpretations of standard statistics muddled at best and simply erroneous at worst. As a result, most scientists are confused about the meaning of a P value or how to interpret it. “It’s almost never, ever, ever stated correctly, what it means,” says Goodman.

Correctly phrased, experimental data yielding a P value of .05 means that there is only a 5 percent chance of obtaining the observed (or more extreme) result if no real effect exists (that is, if the no-

difference hypothesis is correct). But many explanations mangle the subtleties in that definition. A recent popular book on issues involving science, for example, states a commonly held misperception about the meaning of statistical significance at the .05 level: "This means that it is 95 percent certain that the observed difference between groups, or sets of samples, is real and could not have arisen by chance."

That interpretation commits an egregious logical error (technical term: "transposed conditional"): confusing the odds of getting a result (if a hypothesis is true) with the odds favoring the hypothesis if you observe that result.

Another common error equates statistical significance to "significance" in the ordinary use of the word. Because of the way statistical formulas work, a study with a very large sample can detect "statistical significance" for a small effect that is meaningless in practical terms. A new drug may be statistically better than an old drug, but for every thousand people you treat you might get just one or two additional cures — not clinically significant.

Similarly, when studies claim that a chemical causes a "significantly increased risk of cancer," they often mean that it is just statistically significant, possibly posing only a tiny absolute increase in risk.

"I found that eight or nine of every 10 articles published in the leading journals make the fatal substitution" of equating statistical significance to importance, he said in an interview. Ziliak's data are documented in the 2008 book The Cult of Statistical Significance, coauthored with Deirdre McCloskey of the University of Illinois at Chicago.

Even when "significance" is properly defined and P values are carefully calculated, statistical inference is plagued by many other problems. Chief among them is the "multiplicity" issue — the testing of many hypotheses simultaneously. When several drugs are tested at once, or a single drug is tested on several groups, chances of getting a statistically significant but false result rise rapidly. Experiments on altered gene activity in diseases may test 20,000 genes at once, for instance. Using a P value of .05, such studies could find 1,000 genes that appear to differ even if none are actually involved in the disease. Setting a higher threshold of statistical significance will eliminate some of those flukes, but only at the cost of eliminating truly changed genes from the list. In metabolic diseases such as diabetes, for example, many genes truly differ in activity, but the changes are so small that statistical tests will dismiss most as mere fluctuations. Of hundreds of genes that misbehave, standard stats might identify only one or two. Altering the threshold to nab 80 percent of the true culprits might produce a list of 13,000 genes — of which over 12,000 are actually innocent.

Recognizing these problems, some researchers now calculate a "false discovery rate" to warn of flukes disguised as real effects.

Many researchers now also commonly report results with confidence intervals, similar to the margins of error reported in opinion polls. Such intervals, usually given as a range that should include the actual value with 95 percent confidence, do convey a better sense of how precise a finding is. But the 95 percent confidence calculation is based on the same math as the .05 P value and so still shares some of its problems.

Statistical problems also afflict the "gold standard" for medical research, the randomized, controlled clinical trials that test drugs for their ability to cure or their power to harm. Such trials assign patients at random to receive either the substance being tested or a placebo, typically a sugar pill; random selection supposedly guarantees that patients' personal characteristics won't bias the choice of who gets the actual treatment. But in practice, selection biases may still occur, Vance Berger and Sherri Weinstein noted in 2004 in ControlledClinical Trials. "Some of the benefits ascribed to randomization, for example that it eliminates all selection bias, can better be described as fantasy than reality," they wrote.

Randomization also should ensure that unknown differences among individuals are mixed in roughly the same proportions in the groups being tested. But statistics do not guarantee an equal distribution any more than they prohibit 10 heads in a row when flipping a penny. With thousands of clinical trials in progress, some will not be well randomized. And DNA differs at more than a million spots in the human genetic catalog, so even in a single trial differences may not be evenly mixed. In a sufficiently large trial, unrandomized factors may balance out, if some have positive effects and some are negative.

Still, trial results are reported as averages that may obscure individual differences, masking beneficial or harm-ful effects and possibly leading to approval of drugs that are deadly for some and denial of effective treatment to others.

"Determining the best treatment for a particular patient is fundamentally different from determining which treatment is best on average," physicians David Kent and Rodney Hayward wrote in American Scientist in 2007. "Reporting a single number gives the misleading impression that the treatment-effect is a property of the drug rather than of the interaction between the drug and the complex risk-benefit profile of a particular group of patients."

Another concern is the common strategy of combining results from many trials into a single "meta-analysis," a study of studies. In a single trial with relatively few participants, statistical tests may not detect small but real and possibly important effects. In principle, combining smaller studies to create a larger sample would allow the tests to detect such small effects. But statistical techniques for doing so are valid only if certain criteria are met. For one thing, all the studies conducted on the drug must be included — published and unpublished. And all the studies should have been performed in a similar way, using the same protocols, definitions, types of patients and doses. When combining studies with differences, it is necessary first to show that those differences would not affect the analysis, Goodman notes, but that seldom happens. "That's not a formal part of most meta-analyses," he says.

Meta-analyses have produced many controversial conclusions. Common claims that antidepressants work no better than placebos, for example, are based on meta-analyses that do not conform to the criteria that would confer validity.

Similar problems afflicted a 2007 meta-analysis, published in the New England Journal of Medicine, that attributed increased heart attack risk to the diabetes drug Avandia. Raw data from the combined trials showed that only 55 people in 10,000 had heart attacks when using Avandia, compared with 59 people per 10,000 in comparison groups. But after a series of statistical manipulations, Avandia appeared to confer an increased risk.

More recently, epidemiologist Charles Hennekens and biostatistician David DeMets have pointed out that combining small studies in a meta-analysis is not a good substitute for a single trial sufficiently large to test a given question. “Meta-analyses can reduce the role of chance in the interpretation but may introduce bias and confounding,” Hennekens and DeMets write in the Dec. 2 Journal of the American Medical Association. “Such results should be considered more as hypothesis formulating than as hypothesis testing.”

These concerns do not make clinical trials worthless, nor do they render science impotent. Some studies show dramatic effects that don’t require sophisticated statistics to interpret. If the P value is 0.0001 — a hundredth of a percent chance of a fluke — that is strong evidence, Goodman points out. Besides, most well-accepted science is based not on any single study, but on studies that have been confirmed by repetition. Any one result may be likely to be wrong, but confidence rises quickly if that result is independently replicated.

Most critics of standard statistics advocate the Bayesian approach to statistical reasoning, a methodology that derives from a theorem credited to Bayes, an 18th century English clergyman.

His approach uses similar math, but requires the added twist of a “prior probability” — in essence, an informed guess about the expected probability of something in advance of the study.

Often this prior probability is more than a mere guess — it could be based, for instance, on previous studies.

Bayesian math seems baffling at first, even to many scientists, but it basically just reflects the need to include previous knowledge when drawing conclusions from new observations. To infer the odds that a barking dog is hungry, for instance, it is not enough to know how often the dog barks when well-fed. You also need to know how often it eats — in order to calculate the prior probability of being hungry. Bayesian math combines a prior probability with observed data to produce an estimate of the likelihood of the hunger hypothesis.

“A scientific hypothesis cannot be properly assessed solely by reference to the observational data,” but only by viewing the data in light of prior belief in the hypothesis,

“Bayes’ theorem is ... a logically consistent, mathematically valid, and intuitive way to draw inferences about the hypothesis.”

In medical diagnoses, for instance, the likelihood that a test for a disease is correct depends on the prevalence of the disease in the population, a factor that Bayesian math would take into account.

But Bayesian methods introduce a confusion into the actual meaning of the mathematical concept of “probability” in the real world. Standard or “frequentist” statistics treat probabilities as objective realities; Bayesians treat probabilities as “degrees of belief” based in part on a personal assessment or subjective decision about what to include in the calculation. That’s a tough placebo to swallow for scientists wedded to the “objective” ideal of standard statistics. “Subjective prior beliefs are anathema to the frequentist, who relies instead on a series of ad hoc algorithms that maintain the facade of scientific objectivity,”

Conflict between frequentists and Bayesians has been ongoing for two centuries. So science’s marriage to mathematics seems to entail some irreconcilable differences. Whether the future holds a fruitful reconciliation or an ugly separation may depend on forging a shared understanding of probability.

“What does probability mean in real life?” the statistician David Salsburg asked in his 2001 book The Lady Tasting Tea. “This problem is still unsolved, and ... if it remains un-solved, the whole of the statistical approach to science may come crashing down from the weight of its own inconsistencies.”