YOU MIGHT ALREADY KNOW THIS

BENEDICT CAREY

http://www.nytimes.com/2011/01/11/science/11esp.html?pagewanted=all

In recent weeks, editors at a respected psychology journal have been taking heat from fellow scientists for deciding to accept a research report that claims to show the existence of extrasensory perception.

The report, to be published this year in The Journal of Personality and Social Psychology, is not likely to change many minds. And the scientific critiques of the research methods and data analysis of its author, Daryl J. Bem (and the peer reviewers who urged that his paper be accepted), are not winning over many hearts.

Yet the episode has inflamed one of the longest-running debates in science. For decades, some statisticians have argued that the standard technique used to analyze data in much of social science and medicine overstates many study findings often by a lot. As a result, these experts say, the literature is littered with positive findings that do not pan out: effective therapies that are no better than a placebo; slight biases that do not affect behavior; brain-imaging correlations that are meaningless.

By incorporating statistical techniques that are now widely used in other sciences genetics, economic modeling, even wildlife monitoring social scientists can correct for such problems, saving themselves (and, ahem, science reporters) time, effort and embarrassment. I was delighted that this ESP paper was accepted in a mainstream science journal, because it brought this whole subject up again, said James Berger, a statistician at Duke University. I was on a mini-crusade about this 20 years ago and realized that I could devote my entire life to it and never make a dent in the problem.

The statistical approach that has dominated the social sciences for almost a century is called significance testing. The idea is straightforward. A finding from any well-designed study say, a correlation between a personality trait and the risk of depression is considered significant if its probability of occurring by chance is less than 5 percent.

This arbitrary cutoff makes sense when the effect being studied is a large one for example, when measuring the so-called Stroop effect. This effect predicts that naming the color of a word is faster and more accurate when the word and color match (red in red letters) than when they do not (red in blue letters), and is very strong in almost everyone.

But if the true effect of what you are measuring is small, said Andrew Gelman, a professor of statistics and political science at Columbia University, then by necessity anything you discover is going to be an overestimate of that effect.

Date: Published January 10, 2011, in The New York Times.

Consider the following experiment. Suppose there was reason to believe that a coin was slightly weighted toward heads. In a test, the coin comes up heads 527 times out of 1,000. Is this significant evidence that the coin is weighted?

Classical analysis says yes. With a fair coin, the chances of getting 527 or more heads in 1,000 flips is less than 1 in 20, or 5 percent, the conventional cutoff. To put it another way: the experiment finds evidence of a weighted coin with 95 percent confidence.

Yet many statisticians do not buy it. One in 20 is the probability of getting any number of heads above 526 in 1,000 throws. That is, it is the sum of the probability of flipping 527, the probability of flipping 528, 529 and so on.

But the experiment did not find all of the numbers in that range; it found just one 527. It is thus more accurate, these experts say, to calculate the probability of getting that one number 527 if the coin is weighted, and compare it with the probability of getting the same number if the coin is fair. Statisticians can show that this ratio cannot be higher than about 4 to 1, according to Paul Speckman, a statistician, who, with Jeff Rouder, a psychologist, provided the example. Both are at the University of Missouri and said that the simple experiment represented a rough demonstration of how classical analysis differs from an alternative approach, which emphasizes the importance of comparing the odds of a study finding to something that is known.

The point here, said Dr. Rouder, is that 4-to-1 odds just arent that convincing; its not strong evidence.

And yet classical significance testing has been saying for at least 80 years that this is strong evidence, Dr. Speckman said in an e-mail.

The critics have been crying foul for half that time. In the 1960s, a team of statisticians led by Leonard Savage at the University of Michigan showed that the classical approach could overstate the significance of the finding by a factor of 10 or more. By that time, a growing number of statisticians were developing methods based on the ideas of the 18th-century English mathematician Thomas Bayes.

Bayes devised a way to update the probability for a hypothesis as new evidence comes in. So in evaluating the strength of a given finding, Bayesian (pronounced BAYZ-ee-un) analysis incorporates known probabilities, if available, from outside the study.

It might be called the Yeah, right effect. If a study finds that kumquats reduce the risk of heart disease by 90 percent, that a treatment cures alcohol addiction in a week, that sensitive parents are twice as likely to give birth to a girl as to a boy, the Bayesian response matches that of the native skeptic: Yeah, right. The study findings are weighed against what is observable out in the world. In at least one area of medicine diagnostic screening tests researchers already use known probabilities to evaluate new findings. For instance, a new lie-detection test may be 90 percent accurate, correctly flagging 9 out of 10 liars. But if it is given to a population of 100 people already known to include 10 liars, the test is a lot less impressive.

It correctly identifies 9 of the 10 liars and misses one; but it incorrectly identifies 9 of the other 90 as lying. Dividing the so-called true positives (9) by the total number of people the test flagged (18) gives an accuracy rate of 50 percent. The false positives and false negatives depend on the known rates in the population.

In the same way, experts argue, statistical analysis must find ways to expose and counterbalance all the many factors that can lead to falsely positive results among them human nature, in its ambitious hope to discover something, and the effects of industry money, which biases researchers to report positive findings for products.

And, of course, the unwritten rule that failed studies the ones that find no effects are far less likely to be published than positive ones. What are the odds, for instance, that the journal would have published Dr. Bems study if it had come to the ho-hum conclusion that ESP still does not exist?