THE WALL STREET JOURNAL.
WSJ.com

APRIL 1, 2011, 10:04 PM ET

A Statistical Test Gets Its Closeup

My print column this week examines the concept of statistical significance — a concept that the Supreme Court recently weighed in on, but that remains elusive even to some scientists who use it to determine whether their experimental findings are worth reporting.

"The concept is so difficult to understand that misunderstandings are forgivable," said Donald Berry, a biostatistician at the University of Texas M.D. Anderson Cancer Center.

I asked several statisticians to offer definitions of statistical significance. Shane Reese had the briefest one, tailored for a clinical trial for a drug: "It is unlikely that chance alone could have produced the improvement shown in our clinical trial. Because it seems unlikely that chance produced the improvements, we logically conclude that the improvement is due to the drug." Reese and other statisticians noted that this definition is backwards: It is based on assuming there is no link, then finding the probability that chance alone could have produced the experimental results seen.

Reese and Brad Carlin, who also offered a definition, suggest that Bayesian statistics are a better alternative, because they tackle the probability that the hypothesis is true head-on, and incorporate prior knowledge about the variables involved.

There are other problems with statistical significance. It can be ill-suited to cases where it is unclear if all data is being collected, such as with the reporting of adverse events experienced by users of a drug that is past the clinical-trial stage — or never had to go through clinical trials — and is now on the market. In such a situation, "you have to make a lot of assumptions in order to do any statistical test, and all of those are questionable," said Susan Ellenberg, a biostatistician at the University of Pennsylvania's medical school.

"Every statistical test relies on half a dozen assumptions," echoed Aris Spanos, an economist at Virginia Tech. "Before you use that test, you have to check your assumptions."

Spanos wishes the Supreme Court had gone further in its recent ruling, in which it determined that a lack of statistical significance didn't always provide drug companies with enough cover to avoid disclosing reports of adverse events from users of their drugs. Spanos would have liked to see more guidance for how to proceed without relying strictly on statistical significance. "It was a move in the right direction but then you open the system to different kinds of abuses," Spanos said.

Joel Cohen, a partner at the New York law firm Davis Polk, agreed: "It would have been helpful if the court had provided some specific guidance, but they obviously were not comfortable with a black-or-white rule based on statistical significance."

The U.S. Food and Drug Administration also doesn't use such a black-and-white rule. In January the FDA warned women who have gotten breast implants or might get breast implants, because of an elevated risk of the rare cancer anaplastic large-cell lymphoma. The FDA did so even though the link wasn't statistically significant — in part because the agency reasoned that perhaps not all such incidents were reported. "It underscores the importance of not solely relying on a statistical test to tell you there is a public-health issue," said William Maisel, the chief scientist for the agency's Center for Devices and Radiological Health.

There also are cases where seemingly statistically significant results aren't, statisticians say. For example, a very large sample size reduces the effects of statistical noise, so it can yield very high levels of significance for fairly minor relationships — or, roughly speaking, a large degree of confidence in the existence of a very small effect.

Checking for lots of potential effects can also lead to results that appear to be statistically significant, but aren't. "In the early days of clinical trials, it wasn't unusual for people to keep looking at data as they go along," said Ellenberg. "It was a fishing expedition, completely subverting the whole notion of chance findings."

Also, a statistically significant effect may not matter much in practice. "Statistics and value judgment belong to different domains," Siu L. Chow, professor emeritus in psychology at the University of Regina in Saskatchewan, wrote in a written response to questions. "It follows that statistical decision and assessment of substantive impact have their own respective metrics. Hence, it is incorrect to use statistical significance or any other statistical indices (e.g., effect size) to index real-life importance."

Stephen Ziliak, an economist at Roosevelt University in Chicago, and co-author of the 2008 book "The Cult of Statistical Significance" with Deirdre McCloskey, an economist at the University of Illinois at Chicago, said he would like to see large effect sizes reported even when they are not statistically significant. Researchers "probably ought to go ahead and report what happened anyway," Ziliak said. "There's probably a lot of stuff out there that didn't see the light of day."

Ziliak and McCloskey filed a brief against statistical significance in the Supreme Court case, in a considerably calmer tone than their book, which makes a sometimes fierce case against statistical significance, one that not all researchers thought was warranted.

Stephen Stigler, a statistician at the University of Chicago, agrees with the general premise that "you can have a real effect which is nonetheless trivial in the practical sense." He doesn't think this is widely misunderstood, though: "I don't think in science we generally sanction the unequivocal acceptance of significance tests."

Ziliak disagrees, saying of his book: "It is passionate in the sense that we do reveal anger. We had collectively been working on this issue in a calm fashion for 45 years. We deserved to open up the conversation a little more widely in this way."