Nonsignificance on Trial

Michael Smithson

There is a long-running love-hate relationship between the legal and statistical professions, and two vivid examples of this have surfaced in recent news stories, one situated in a court of appeal in London and the other in the U.S. Supreme Court. Briefly, the London judge ruled that Bayes' theorem must not be used in evidence unless the underlying statistics are "firm;" while the U.S. Supreme Court unanimously ruled that a drug company's non-disclosure of adverse side-effects cannot be justified by an appeal to the statistical non-significance of those effects. Each case, in its own way, shows why it is high time to find a way to establish an effective rapprochement between these two professions.

The Supreme Court decision has been applauded by statisticians, whereas the London decision has appalled statisticians of similar stripe. Both decisions require some unpacking to understand why statisticians would cheer one and boo the other, and why these are important decisions not only for both the statistical and legal professions but for other domains and disciplines whose practices hinge on legal and statistical codes and frameworks.

This post focuses on the Supreme Court decision. The culprit was a homoeopathic zinc-based medicine, Zicam, manufactured by Matrixx Initivatives, Inc. and advertised as a remedy for the common cold. Matrixx ignored reports from users and doctors since 1999 that Zicam caused some users to experience burning sensations or even to lose the sense of smell. When this story was aired by a doctor on Good Morning America in 2004, Matrixx stock price plummeted.

The company's defense was that these side-effects were "not statistically significant." In the ensuing fallout, Matrixx was faced with more than 200 lawsuits by Zicam users, but the case in point here is Siracusano vs Matrixx, in which Mr. Siracusano was suing on behalf of investors on grounds that they had been misled. After a few iterations through the American court system, the question that the Supreme Court ruled on was whether a claim of securities fraud is valid against a company that neglected to warn consumers about effects that had been found to be statistically non-significant. As insider-knowledgeable Stephen Ziliak's insightful essay points out, the decision will affect drug supply regulation, securities regulation, liability and the nature of adverse side-effects disclosed by drug companies. Ziliak was one of the "friends of the court" providing expert advice on the case.

A key point in this dispute is whether statistical nonsignificance can be used to infer that a potential side-effect is, for practical purposes, no more likely to occur when using the medicine than when not. Among statisticians it is a commonplace that such inferences are illogical (and illegitimate). There are several reasons for this, but I'll review just two here.

These reasons have to do with common misinterpretations of the measure of statistical significance. Suppose Matrixx had conducted a properly randomized double-blind experiment comparing Zicam-using subjects with those using an indistinguishable placebo, and observed the difference in side-effect rates between the two groups of subjects. One has to bear in mind that random assignment of subjects to one group or the other doesn't guarantee equivalence between the groups. So, it's possible that even if there really is no difference between Zicam and the placebo regarding the side-effect, a difference between the groups might occur by "luck of the draw."

The indicator of statistical significance in this context would be the probability of observing a difference at least as large as the one found in the study if the hypothesis of no difference were true. If this probability is found to be very low (typically .05 or less) then the experimenters will reject the no-difference hypothesis on the grounds that the data they've observed would be very unlikely to occur if that hypothesis were true. They will then declare that there is a statistically significant difference between the Zicam and placebo groups. If this probability is not sufficiently low (i.e., greater than .05) the experimenters will decide not to reject the no-difference hypothesis and announce that the difference they found was statistically non-significant.

So the first reason for concern is that Matrixx acted as if statistical nonsignificance entitles one to believe in the hypothesis of no-difference. However, failing to reject the hypothesis of no difference doesn't entitle one to believe in it. It's still possible that a difference might exist and the experiment failed to find it because it didn't have enough subjects or because the experimenters were "unlucky." Matrixx has plenty of company in committing this error; I know plenty of seasoned researchers who do the same, and I've already canvassed the well-known bias in fields such as psychology not to publish experiments that failed to find significant effects.

The second problem arises from a common intuition that the probability of observing a difference at least as large as the one found in the study if the hypothesis of no difference were true tells us something about the inverse—the probability that the no-difference hypothesis is true if we find a difference at least as large as the one observed in our study, or, worse still, the probability that the no-difference hypothesis is true. However, the first probability on its own tells us nothing about the other two.

For a quick intuitive, if fanciful, example let's imagine randomly sampling one person from the world's population and our hypothesis is that s/he will be Australian. On randomly selecting our person, all that we know about her initially is that she speaks English.

There are about 750 million first-or second-language English speakers world-wide, and about 23 million Australians. Of the 23 million Australians, about 21 million of them fit the first- or second-language English description. Given that our person speaks English, how likely is it that we've found an Australian? The probability that we've found an Australian given that we've picked an English-speaker is 21/750 = .03. So there goes our hypothesis. However, had we picked an Australian (i.e., given that our hypothesis were true), the probability that s/he speaks English is 21/23 = .91.

See also Ziliak and McCloskey's 2008 book, which mounts a swinging demolition of the unquestioned application of statistical significance in a variety of domains.

Aside from the judgment about statistical nonsignificance, the most important stipulation of the Supreme Court's decision is that "something more" is required before a drug company can justifiably decide to not disclose a drug's potential side-effects. What should this "something more" be? This sounds as if it would need judgments about the "importance" of the side-effects, which could open multiple cans of worms (e.g., Which criteria for importance? According to what or whose standards?). Alternatively, why not simply require drug companies to report all occurrences of adverse side-effects and include the best current estimates of their rates among the population of users?

A slightly larger-picture view of the Matrixx defense resonates with something that I've observed in even the best and brightest of my students and colleagues (oh, and me too). And that is the hope that somehow probability or statistical theories will get us off the hook when it comes to making judgments and decisions in the face of uncertainty. It can't and won't, especially when it comes to matters of medical, clinical, personal, political, economic, moral, aesthetic, and all the other important kinds of importance.