3 INTERLEAF

Why statistical significance is not the same as biological importance

In the early 20th century, the word "significant" had one dominant meaning. If you said that something was "significant" you meant that it *signified* or showed something—that it had or conveyed a meaning. When R. A. Fisher said that a result was significant, he meant that the data showed some difference from the null hypothesis. In other words, we were able to learn something from those data. This sense of the word "significant" has persisted in the scientific literature. When discussing data, a "statistically significant" result means that a null hypothesis has been rejected.

Full of sound and fury, Signifying nothing.

—Shakespeare, *Macbeth*

But languages, like species, evolve. Over the 20th century, the word "significant" came to mean *important*. Today, outside of statistics, when we say that something is significant, we usually mean that it has value and import—that it ought to be paid attention to. This creates ambiguity when the term is applied now to scientific findings. When a newspaper article describes a new scientific result as "a significant new finding" or "a significant advance in our knowledge," for example, is that a statistical statement or a value judgment? Sometimes the meaning is blurred.

A statistically significant result is not the same as a biologically important result. An important result in biology refers to one whose effect is large enough to matter in some way. The mean lengths of the third molars differ between species of early mammals, but this doesn't by itself mean that the difference is vital. To address the importance of the difference, we need to know its magnitude and how it might matter.

The problem is that extremely small, biologically uninteresting effects can be statistically significant, as long as the sample size is sufficiently large. For example, automobile accidents increase during full moons, and this result is statistically significant (Templer et al. 1982). Such results attracted media attention because they bring to mind stories of werewolves and vampire legends. But the size of the effect is about 1%, far too small to make it worth changing our driving habits or public policy. In fact, almost any null hypothesis can be disproved with a large enough sample. Are there really populations out there *exactly* equal to each other in every way? We should care about the differences only if they are large enough to matter.

We already have some sense that statistically significant does not mean important when we read of scientific studies that showed such earth-shattering facts as "teenagers like to listen to music sometimes," "driving fast makes the ride feel more bumpy," and "spiders scare some people." Each of these results was backed by hard data and statistics, but each surprised no one.

On the other hand, a result can be important even if it is not statistically significant. Sometimes new data suggest a pattern that, if true, is very important, provoking further study. For example, the first studies testing whether administering streptokinase prevents strokes did not reject the null hypothesis that this drug had no effect on mortality rate. But they were small studies that showed a suggestive pattern. As a result, further larger studies were conducted, and streptokinase was eventually shown to be effective in reducing mortality rates from strokes. This is why we don't "accept the null hypothesis"—a small study on a small effect will have a low probability of rejecting a false null hypothesis.

At the other extreme, sometimes it is important to show the lack of an effect. A large-scale study of the efficacy of hormone replacement therapy (HRT) showed no statistically significant evidence for a benefit of HRT to postmenopausal women. Moreover, confidence intervals showed that any plausible effect was small. HRT had been in wide use with substantial money being invested and with some known deleterious side effects. Knowing that it had little benefit saved a great deal of resources and prevented many of the side effects. This result was not "statistically significant," but it was very important medically.

When presenting data, we should always report the estimated magnitude of the effect with a confidence interval, not just the *P*-value. The confidence interval gives us a plausible range for the size of the effect, and if this interval includes values with greatly varying interpretations, we know that we have to revisit the question with further data. We should look at a graphical presentation of the data to gauge the magnitude of the effect. The importance of a result depends on the value of the question and the size of the effect. Statistical significance tells us merely how confidently we can reject a null hypothesis, but not how big or how important the effect is.