JARGON ALERT

Statistical Significance

BY EAMON O'KEEFE

A drug company has developed a new treatment for high cholesterol. It finds that patients who take the new drug experience fewer heart attacks and other negative effects from the condition. But how confident are we in those results? This is the question of statistical significance, and it can be applied to the social sciences to help economists better determine the effects of a certain policy change or business decision.

To determine statistical significance, a researcher begins by creating a null and an alternative hypothesis to test if a relationship exists between two events or characteristics. The null hypothesis typically states that no relationship exists, and the alternative hypothesis asserts that a relationship does exist. For example, an economist might suspect a rise in the minimum wage will affect the employment of less-skilled workers. The null hypothesis would be that, on average, there is no change in the unemployment rate for less-skilled workers after a state raises its minimum wage. The alternative hypothesis would be that there is a change in unemployment after an increase in the minimum wage.

Suppose the economist runs a regression analysis and the coefficient on the minimum wage variable is positive — suggesting a possible correlation between unemployment for less-skilled workers and a state's minimum wage. The next step is to determine our level of confidence in that result. Researchers use what is called a p-value to communicate the probability of finding a relationship when no such relationship exists. If the p-value is below a certain threshold — 5 percent is commonly used — the relationship is deemed statistically significant and the null hypothesis can be rejected.

Of course, correlation is not the same as causation. Just because a change in one variable coincides with a change in the other does not necessarily mean they cause one another. For example, playing tennis might be correlated with wealth, but unless one is a professional tennis player, it won't lead to greater wealth. Without a controlled experiment, it's very difficult to prove causality. Controlled experiments are relatively rare in economics; for example, it's unlikely that legislators would allow an economist to tinker with their state's minimum wage in the name of scientific inquiry. But economists can take advantage of "natural experiments," such as one state raising its minimum wage while a neighboring state leaves its wage unchanged. Or they can use statistical techniques to control for other factors that might affect employment. A considerable amount of research has used such methods to study the minimum wage. Most studies have found disemployment effects, although the magnitude varies considerably. (See "Raise the Wage?" *Econ Focus*, Third Quarter 2014.)

Just as it's important to distinguish between correlation and causation, it's also important to distinguish between statistical significance and economic significance. Statistical significance is about your confidence in the result, but just because a result is statistically significant doesn't mean the result is large or meaningful. For example, say a large increase in the state minimum wage caused a few people in that state to lose their jobs. The statistical relationship might be strong, but the magnitude of job loss could be small enough to be inconsequential to policymakers.

The problem of error is implicit in any discussion of statistical significance. There exists, in a statistical test, the possibility for two types of error: type 1 and type 2. A type 1 error indicates a "false positive" or rejecting the null hypothesis when it is true. A type 2 error is when one accepts the null when it is false. Both can be problematic, but the extent to which the researcher is concerned about the error depends on the question being explored.

It's important to take type 1 and type 2 errors into account when considering the threshold for statistical significance. The smaller the p-value, the higher the bar for significance. So a researcher who is especially concerned about making a type 1 error might look for significance well below 0.05. In a 2012 column, Carl Bialik, the *Wall Street Journal*'s "The Numbers Guy," detailed how this concept was used to validate the existence of the elusive Higgs boson particle — sometimes referred to as the "God particle." Researchers used a statistical significance of "five sigmas" to reject a result with a p-value greater than one in 3.5 million. They wanted to set an extremely high burden of proof for discovering a new particle in the universe.

This discussion of error can be applied to other questions society faces. For example, many might argue that determining guilt in a death penalty case should require a higher burden of proof than in a normal trial. Implicitly, one is determining a p-value in this situation because it is desirable to have a very low probability of type 1 error (convicting someone and sentencing them to death for a crime they didn't commit).

In a sense, then, statistical significance reflects value judgments. Setting a high or low p-value indicates a researcher's belief about what constitutes significance — an additional nuance to be mindful of when interpreting research findings.   **EF**