A Brief Explanation of Statistical Significance and *P* Values

Research data can be interpreted in terms of their statistical significance and their practical significance. To understand the statistical significance of results presented in research papers, you must first understand the concept of statistical inference. As you may recall from your statistics class, this concept entails making generalizations about a large population by using data from a relatively small segment of it, which is called a *sample*.

For most studies in the physiological sciences, it's impossible to collect data from the entire population of interest. Consider, for example, an experiment on the effects of a new drug for lowering blood pressure. The researchers would not be able to administer the drug to every single person with hypertension (high blood pressure) around the world. So a sample of hypertensive patients would be selected for the study, perhaps from the researchers' local community. After administering the drug to an experimental group and a placebo treatment to a control group, the researchers would calculate the mean changes in blood pressure for the groups over the study period. The difference in blood pressure change between the two groups cannot, on its own, validly reflect the statistical significance of the data. That's because the value observed from the sample might differ, perhaps even greatly, from the change observed in *another* sample from the larger population. Let's say that the mean value for systolic blood pressure decreased by 6.2 mm Hg and 1.2 mm Hg in the experiment and control groups, respectively. The difference in change between the two groups is 5.0 mm Hg. In repeated experiments on different samples of the population of hypertensive individuals in the world, we would not expect to observe a difference in change of exactly 5.0 mm Hg for systolic blood pressure. Instead, we would expect sampling variability. That is, the values might range quite a bit around the mean difference of 5.0 mm Hg observed in the original sample. To avoid using intuition in interpreting their data, researchers must use statistical tests and probability theory to determine whether observed differences are due to *systematic effects* of treatments or simply to *chance factors*.

Statistical analyses are carried out to test the tenability of the null hypothesis, which is an assumption that in the population of interest no differences exist between groups and that treatments of interest have no effects. To maintain objectivity, researchers establish the null hypothesis prior to conducting their studies. But researchers may also establish a research hypothesis, which is an informed statement (not a wild guess) about their true expectations and beliefs. Statistical tests help scientists determine whether the null hypothesis or the research hypothesis is more tenable. The decision is based on the significance of study results. In statistics, the term "significance" refers to the mathematical probability of obtaining the observed results if the same experiment were conducted many times on different samples of subjects in the population of interest. In conducting tests of statistical significance (such as t-tests and ANOVA), researchers answer this central question: If the null hypothesis were true in the population (that is, if there really is no difference between groups and no treatment effect), what is the probability of obtaining the results that we observed in our experiment?

The key outcome of an inferential statistical test is a *P* value, which is the probability of obtaining a result—such as a difference in blood pressure change between a treatment and control group—as extreme or more extreme than the one observed if no true effect existed in the population.

Consider, again, a study on the effects of a new drug for treating hypertension. Over the study period, the experimental and control groups experienced a decrease in systolic blood pressure of 6.2 mm Hg and 1.2 mm Hg, respectively. Is the 5.0 mm Hg difference truly due to the effects of the drug, or is the difference simply due to chance factors? The statistical test would yield a *P* value, which is the probability of observing a 5.0 mm Hg or greater difference in systolic blood pressure in repeated experiments if the null hypothesis were true. Let's say that the statistical test revealed a *P* value of .99. Here's the interpretation: If it's really true that the drug has no effect on systolic blood pressure, we would observe a 5.0 mm Hg difference between experimental and control subjects in 99 out of 100 repeated experiments. If we obtained such a result so frequently, we would be confident that the null hypothesis is tenable and that the drug doesn't really reduce blood pressure.

What if the test revealed a *P* value of .01? Here's the interpretation: If it's really true that the drug has no effect on systolic blood pressure, we would observe a 5.0 mm Hg difference between experimental and control subjects in 1 out of 100 repeated experiments. If we obtained such a result so rarely under the assumption that the null hypothesis is true, we would have to doubt that the null hypothesis is tenable. We would thus accept the research hypothesis that the drug does reduce blood pressure. We would say that the 5.0 mm Hg difference is statistically *significant*. As you may have learned in statistics class, a more accurate interpretation is that the 5.0 mm Hg difference is statistically *reliable*.

As the *P* value gets lower (i.e., closer to 0% and farther away from 100%), researchers are more inclined to accept the research hypothesis and to reject the null hypothesis. As you know, before beginning studies scientists conventionally establish cutoff points, which are called *alpha values*, to indicate what *P* values they will accept as significant. In many physiological studies, alpha values are set at .05 or .01. As you know, if the *P* values that are calculated in statistical tests are less than alpha—for example, if $P < .05$—the researchers would conclude that their study results are statistically significant.

A relatively simple way to interpret *P* values is to think of them as representing how likely a result would occur by chance. For a calculated *P* value of .001, we can say that the observed outcome would be expected to occur by chance only 1 in 1,000 times in repeated tests on different samples of the population.

Here's a very important point to consider when you're reading scientists' interpretations of their data in research papers: **Statistical significance by no means indicates practical significance, or the importance of the data in an applied setting.** To reach strong interpretations about the practical significance of a study's data, you must deeply understand the motivating research questions and the science that defines the field. In our example scenario above, let's say that the study revealed that the experimental group experienced a statistically significant reduction in blood pressure, lowering their mean values for systolic pressure by 5.0 mm Hg. Practically speaking, we have to ask, "*So what?*" Is a blood pressure reduction of this magnitude clinically significant? Could this reduction noticeably enhance the heart's function to a degree that improves the general health of individuals with hypertension? Could a reduction of this magnitude lower an individual's risks of the negative outcomes that are associated with high blood pressure, such as heart attack and stroke? These questions must be answered, through applying knowledge about cardiovascular physiology, to interpret the *practical significance* of the results.