Politzer-Ahles, S., & Chen, S. In press at *The SAGE Encyclopedia of Human Communication Sciences and Disorders*.

Significance

Statistical significance is a concept used in null hypothesis statistical testing to indicate how much a given sample (e.g., data from an experiment) is inconsistent with a given hypothesis. For example, suppose a researcher has measured the vocabulary size in 100 children with Specific Language Impairment (SLI) and 100 typically developing children, and she finds that it is 25 words lower on average in children with SLI, and the pooled variance of the difference between groups is 40 words; she now wants to decide whether this difference is meaningful. Researchers often use statistical significance to make this decision, but actually statistical significance is only a tool for making indirect inferences about whether a sample reflects a population, and it does not indicate that a finding is "real", large, or important.

In inferential statistics, the properties of a **sample** (e.g., the results of the experiment described above) are known, but the properties of the **population** (e.g., the vocabulary size for all children with SLI in the world and all typically developing children in the world, at the time the experiment is conducted), which the researcher is often interested in, are unknown. In most research and clinical applications, measuring the entire population is infeasible; thus, practitioners collect data from small samples and try to infer the properties of the population. If one knew the vocabulary size for every child in the world, one could know the likelihood of observing any given difference in one experiment. But it is usually not possible to know what kind of population the experiment data came from; for example, even if there is no aggregate difference between the vocabulary sizes of typically developing and SLI children in the population, the researcher could still observe a big difference in her experiment if she happened to randomly sample 100 typically developing children with unusually high vocabulary sizes and 100 SLI children with unusually low ones. Therefore, a common statistical approach is to first

assume that there is no difference between these children in the world population. One can then calculate how probable it is that an experiment could have found a difference as big, or bigger, than the observed difference if there really were no difference in the population. (Technically, this "difference" is operationalized as a test statistic calculated from the observed sample, and often takes into account both the observed mean difference [25 words, in this example] and pooled variance [40 words].) This probability is known as a *p*-value; a small *p*-value means that if there were no difference in the population, the probability that one would have observed a difference (technically a test statistic) this big or bigger in an experiment is small. R. A. Fisher introduced this *p*-value as a tool for showing the strength of evidence against the hypothesis that the true effect in the population is zero (in this example, this hypothesis is the assumption that there is no difference in vocabulary size between the population of SLI children and typically developing children). In communication science, *p*-values below 5% are typically considered **statistically significant**, and treated as being relatively strong evidence against the hypothesis of no difference.

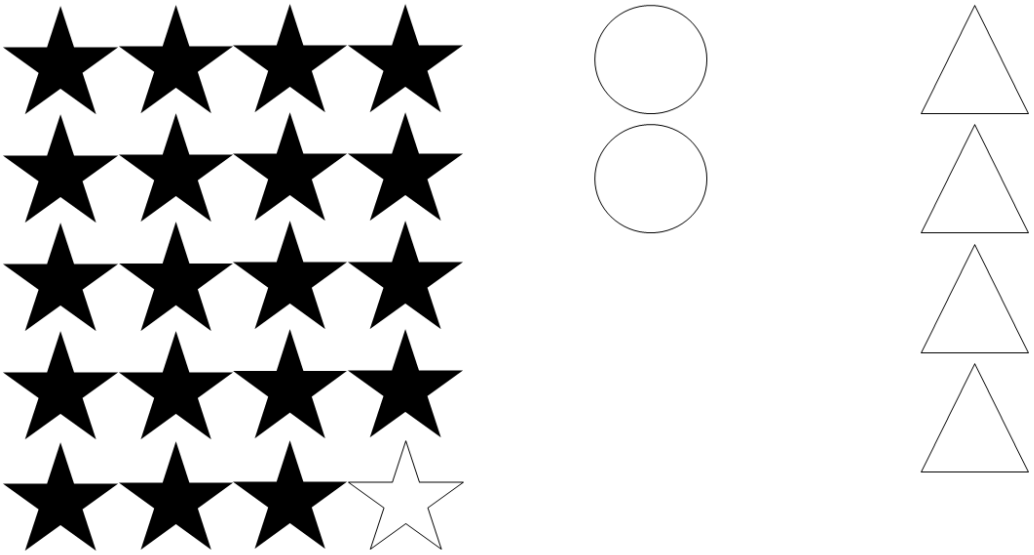

Figure 1. Illustration of a case where a population of stars is 5% likely to yield a white sample, but a white sample is 14% likely to have been drawn from a population of stars.

Note that a *p*-value indicates the probability of an observed result (technically, an observed test statistic) given a hypothesis, *not* the probability of a hypothesis given the observed result. The probability of A given B is not always equal to the probability of B given A. Consider Figure 1: if a shape is a star there's a 5% chance that it's white, but it would be false to say "if a shape is white there's a 5% chance it's a star" (in fact there's an approximately 14% chance). In the SLI example, a *p*-value of 5% would only mean that if there were no difference in the population then there's a 5% chance of observing a difference (test statistic) this big (or bigger) in the experiment; it would *not* mean there is a 5% chance that there is no difference in the population.

Some additional limitations of statistical significance are as follows:

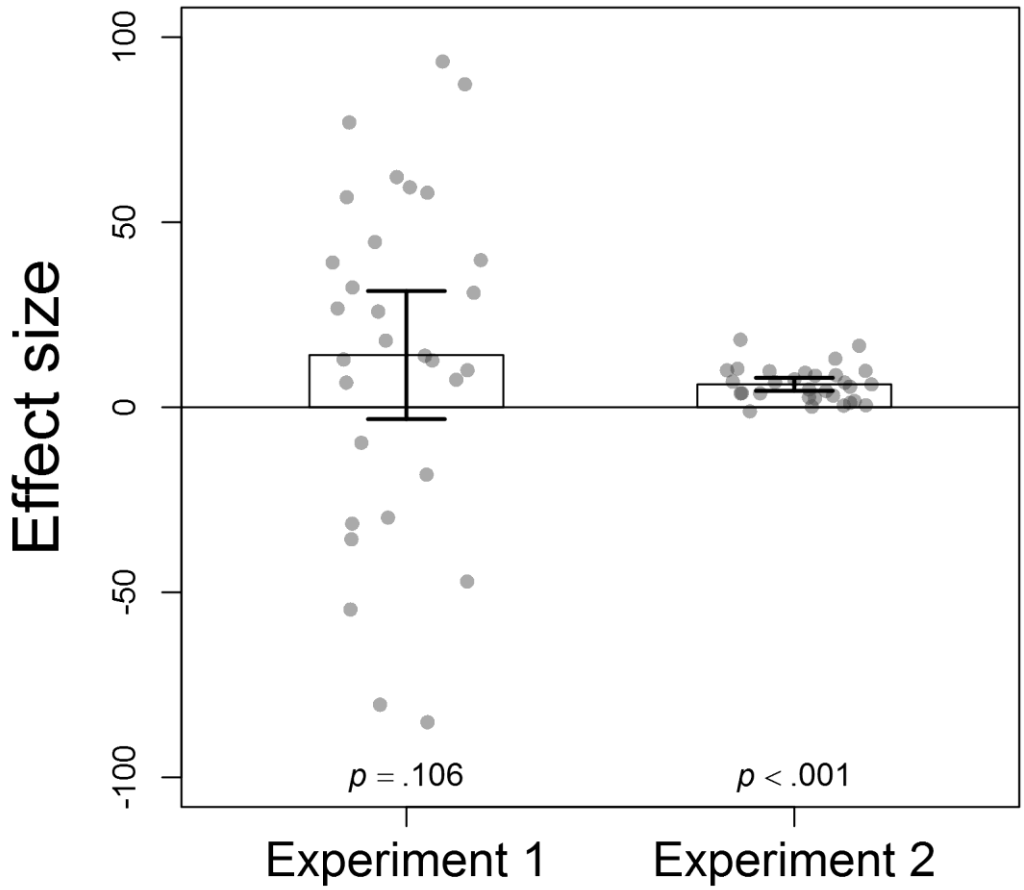

Figure 2. Data from two fake N=30 experiments. Bars show the means, dots show each participant's result, and red error bars show 95% confidence intervals based on the standard error of the mean and the t *statistic. Experiment 1 has a mean that is high but non-significant: the variation around the mean is large, many participants have negative effect sizes, the confidence interval includes zero. Experiment 2 has a mean that is small but significant: the variation is small, almost all participants have effect sizes in the same direction, and the confidence interval does not include zero.*

Significance does not indicate effect size. Test statistics are often based on a *ratio* of the effect size and the sample variance. Thus, effects can be large but non-significant, or small but significant (Figure 2).

Significance does not indicate importance or reality. Significance does not tell whether that result is replicable, of practical importance, or reflective of "real" phenomena. For example, an intervention might increase children's average vocabulary size by the small, but significant, amount of 0.01 words. If the intervention is costly or harmful, then a practitioner may well conclude that the effect is significant but not of practical value.

5% significance is not a bright line. In communication sciences, significance levels below 5% are often treated as statistically significant, and significance levels above 5% as not significant. However, significance is only one tool for identifying attention-worthy phenomena. It must be considered alongside other forms of evidence, including previous research and *a priori* predictions; scientific conclusions should not be blindly made based on *p*-values alone.

Lack of significance does not mean no effect. A non-significant *p*-value means there is *in*sufficient evidence to *reject* the hypothesis that there is no effect, not that there is sufficient evidence to accept it. An effect of the same size may show significant or non-significant *p*-values, due to different sample size or sample variance; experiments observing large but non-significant effect sizes may be consistent with (and indeterminate between) both large effects and negligible or nonexistent effects.

Stephen Politzer-Ahles, The Hong Kong Polytechnic University, Hong Kong

Si Chen, The Hong Kong Polytechnic University, Hong Kong

Further reading

Goodman, S. (2008). A dirty dozen: twelve *p*-value misconceptions. *Seminars in Hematology, 45*, 135-140.

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review, 13*, 1033-1037.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician, 70*, 129-133.

Cross references:

experimental research, quantitative research, research, statistics: descriptive, statistics: predictive, theory of signal detection, validity.