To appear in I. Weiner, & E. Craighead (Eds.) *The Corsini encyclopedia of psychology*, 4th Edition. Hoboken, NJ. Wiley & Sons.

Joachim I. Krueger
Brown University

Significance Testing

Over the past 100 years, the practice of significance testing (ST) has become nearly universal in psychological research. The increasing acceptance of the a method has been punctuated by impassioned critiques and search for viable alternatives.

The quantitative core of ST is the probability p that, given a specific state of nature, certain observations, or observations more extreme, will be made. The presumed state of nature is typically framed as a null hypothesis, H_0, and the observations are the data, D, obtained empirically. Many test statistics are available for the evaluation of differences between empirical means, medians, proportions, correlation coefficients, regression weights, and many others.

Psychological research is often conducted with small samples in areas offering little prior knowledge about what to expect and few opportunities for precise, theory-grounded, quantitative modeling. Under these circumstances, many investigators seek to design studies with measurements precise enough and samples large enough to yield $p(D|H_0) \leq .05$. This strategy is weak as it settles for the detection of a directional effect (Meehl, 1990).

Under different circumstances, investigators seek to obtain $p(D|H_0) > .05$. One class of such circumstances are assumptions that permit ST in the first place. Some tests require that the samples, the means of which are to be compared, have variances that do not significantly differ from one another. Other tests require that the correlations among measurements within different samples of the same study are not significantly different from one another. Another set of circumstances arises when a well-articulated theoretical model specifies the set of associations among measurement variables. Here, $p(D|H) \leq .05$—where H is a substantive, non-null hypothesis—suggests that the evidence militates against, not for, a theoretical idea.

Research practitioners try to extract more information from ST than the method can provide, and mathematical statisticians have yet to provide an objective alternative that avoids such misconceptions. One common error is to assume that $p(D|H_0)$ is the probability that the null hypothesis is true. Another error is to assume that $1- p(D|H_0)$ is the probability that an empirical finding will be replicated. The probability that an hypothesis is true given the data, $p(H|D)$, and the probability of a successful replication can be derived with Bayesian methods (Krueger, 2001; Trafimow, 2003). Such calculations require estimates of the prior probabilities—i.e., probabilities estimated before new empirical evidence is introduced—of at least two hypotheses. The proper way of estimating priors, and whether to estimate them at all, remains a topic of debate.

When theories do not provide precise hypotheses, ST takes its original Fisherian form. Without an alternative hypothesis, H_1, it is not possible to estimate the prohability that H_0 is falsely rejected (Type I error) or falsely accepted (Type II error). Consequently, it is impossible to estimate statistical power, which is the complement of the probability of a Type II error. The concepts of different types of error are characteristic of the Neyman-Pearson approach to ST, which Fisher opposed.

Although many psychologists are trained to think in terms of decision errors and urged to provide power estimates when applying for research funds, their research practice remains largely Fisherian. They see low values of $p(D|H_0)$ as strong evidence against the null hypothesis and as strong evidence for the replicability of the findings. Consider a two-sample t-test performed on data where $M_1 = 65$, $M_2 = 50$, $s_1 = s_2 = 30$, $n_1 = n_2 = 5$. The test is not significant, $t(8) = 1.58$, $p = .15$ (two-tailed), although the standardized effect size is intermediate, $d = (M_1 - M_2)/s = .5$. The primary reason for a lack of significance is the large standard error of the mean difference = $s/\sqrt{10} = 9.49$. Shrinking the variance of the observations (e.g., by reducing measurement error) or raising the number of observations reduces the standard error. The latter method is often preferred if only because it requires greater persistence rather than more sophisticated measurement techniques.

The strategy of increasing the number of observations raises the question of how to determine the appropriate sample size. Without an alternative hypothesis, the power of the test—and thus the requisite sample size—cannot be estimated before study. Researchers can, however, gather an arbitrary number of observations, compute the basic statistics, treat the observed effect size as the best estimate of the alternative hypothesis, and then perform a power analysis. Midstream data analyses and updates of sample size estimates are heterodox, but the strategy has merit, particularly if it is openly practiced by everyone.

Many statisticians hold that the null hypothesis is false *a priori* and not worth testing. Only effect sizes are of interest. Yet, small effect sizes with large measurement errors are undesirable. The practical advantage of NHST is that it motivates researchers to gather data until $p(D|H_0) \leq .05$. In the numerical example above, this point is reached with $n_1 = n_2 = 9$. Using $p = .05$ as an anchor, the only remaining variables are the effect size, and the sample size. Although they are perfectly negatively related, these two variables convey unique information. The effect size conveys the state of nature, and the sample size conveys the utility the researcher attaches to detecting it. Large investments made to coax small effects into significance can serve as a cue toward the practical significance of the finding. In published empirical work too, the size of the effect is inversely related to the size of the sample used to test it, which suggests that many investigators informally conduct sequential testing. Making sequential testing explicit and standardizing its practice might turn this presumed vice into a virtue. Sequential testing with $p(D|H_0) = .05$ as a common benchmark makes effect sizes directly comparable, and thus facilitates the assessment of their practical significance.

The *Association for Psychological Science* now endorses the weakest form of ST by requiring authors to report the probability p_{rep} that a replication study will yield an effect of the same sign (Killeen, 2005). The effect obtained in the original study is the mean of a distribution whose variance decreases with the size of the sample, $1\text{-}p_{rep} > p$. This inequality becomes smaller with increasing n. As p_{rep} and p use the same information, their log transforms are perfectly correlated. To demand $p_{rep} = .9$ is to demand $p \approx .03$.

Word count: 1,045

References

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*, 345-353.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*, 16-26.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*, 526-535.

Suggested Readings

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149-169). New York: Guilford.

Keywords

Null hypothesis, probability, replication, effect size