Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 4, Number 5, November, 1994 ISSN=1531-7714

The Concept of Statistical Significance Testing.

Bruce Thompson,
Texas A & M

Too few researchers understand what statistical significance testing does and doesn't do, and consequently their results are misinterpreted. Even more commonly, researchers understand elements of statistical significance testing, but the concept is not integrated into their research. For example, the influence of sample size on statistical significance may be acknowledged by a researcher, but this insight is not conveyed when interpreting results in a study with several thousand subjects.

This article will help you better understand the concept of significance testing. The meaning of probabilities, the concept of statistical significance, arguments against significance testing, misinterpretation, and alternatives are discussed.

WHAT ARE THOSE PROBABILITIES IN STATISTICAL SIGNIFICANCE TESTING?

Researchers may invoke statistical significance testing whenever they have a random sample from a population, or a sample that they believe approximates a random, representative sample. Statistical significance testing requires subjective judgment in setting a predetermined acceptable probability (ranging between 0 and 1.0) of making an inferential error caused by the sampling error--getting samples with varying amounts of "flukiness"--inherent in sampling. Sampling error can only be eliminated by gathering data from the entire population.

One probability (p), the probability of deciding to reject a null hypothesis (e.g., a hypothesis specifying that $Mean_1 = Mean_2 = Mean_3$, or $R^2 = 0$) when the null hypothesis is actually true in the population, is called "alpha," and also $p_{(CRITICAL)}$. When we pick an alpha level, we set an upper limit on the probability of making this erroneous decision, called a Type I error. Therefore, alpha is typically set small, so that the probability of this error will be low. Thus, $p_{(CRITICAL)}$ is selected based on subjective judgment regarding what the consequences of Type I error would be in a given research situation, and given personal values regarding these consequences.

A second probability, $p_{(CALCULATED)}$ (which, like all p's, ranges between .0 and 1.0), is calculated. Probabilities can only be calculated in the context of assumptions sufficient to constrain the computations such that a given problem has only one answer.

What's the probability of getting mean IQ scores of 99 and 101 in two sample groups? It depends, first, on the actual statistical parameters (e.g., means) in the populations from which the samples were drawn. These two sample statistics ($Mean_1 = 99$ and $Mean_2 = 101$) would be most probable (yielding the highest $p_{(CALCULATED)}$ if the population means were respectively 99 and 101. These two sample statistics would be less likely (yielding a smaller $p_{(CALCULATED)}$ if the population means were both 100. Since the actual population parameters are not known, we must assume what the parameters are, and in statistical significance testing we assume the parameters to be correctly specified by the null hypothesis, i.e., we assume the null hypothesis to be exactly true for these calculations.

A second factor that influences the calculation of p involves the sample sizes. Samples (and thus the statistics calculated for them) will potentially be less representative of populations ("flukier") as sample sizes are smaller. For example, drawing two samples of sizes 5 and 5 may yield "flukier" statistics (means, r's, etc.) than two samples of sizes 50 and 50. Thus, the $p_{(CALCULATED)}$ computations also must (and do) take sample size influences into account. If the two samples both of size 5 had means of 100 and 90, and the two samples both of size 50 also had means of 100 and 90, the test of the null that the means are equal would yield a smaller $p_{(CALCULATED)}$ for the larger samples, because assuming the null is exactly true, unequal sample statistics are increasingly less likely as sample sizes increase. Summarizing, the $p_{(CALCULATED)}$ probability addresses the question:

Assuming the sample data came from a population in which the null hypothesis is (exactly) true, what is the probability of obtaining the sample statistics one got for one's sample data with the given sample size(s)?

Even without calculating this p, we can make logical judgments about $p_{(CALCULATED)}$. In which one of each of the following pairs of studies will the $p_{(CALCULATED)}$ be smaller?

- In two studies, each involving three groups of 30 subjects: in one study the means were 100, 100, and 90; in the second study the means were 100, 100, and 100.
- In two studies, each comparing the standard deviations (SD) of scores on the dependent variable of two groups of subjects, in both studies $SD_1 = 4$ and $SD_2 = 3$, but in study one the sample sizes were 100 and 100, while in study two the samples sizes were 50 and 50.
- In two studies involving a multiple regression prediction of Y using predictors X_1, X_2, and X_3, and both with samples sizes of 75, in study one $R^2 = .49$ and in study two $R^2 = .25$.

WHAT DOES STATISTICAL SIGNIFICANCE REALLY TELL US?

Statistical significance addresses the question:

> "*Assuming the sample data came from a population in which the null hypothesis is (exactly) true, and given our sample statistics and sample size(s), is the calculated probability of our sample results less than the acceptable limit ($p_{(CRITICAL)}$) imposed regarding a Type I error?*"

When $p_{(CALCULATED)}$ is less than $p_{(CRITICAL)}$, we use a decision rule that says we will "reject" the null hypothesis. The decision to reject the null hypothesis is called a "statistically significant" result. All the decision means is that we believe our sample results are relatively unlikely, given our assumptions, including our assumption that the null hypothesis is exactly true.

However, though it is easy to derive $p_{(CRITICAL)}$, calculating $p_{(CALCULATED)}$ can be tedious. Traditionally, test statistics (e.g., F, t, X squared) have been used as equivalent (but more convenient) reexpressions of p's, because Test $Statistics_{(CALCULATED)}$ are easier to derive. The $TS_{(CRITICAL)}$ exactly equivalent to a given $p_{(CRITICAL)}$ can be derived from widely available tables; the tabled value is found given alpha and the sample size(s). Different $TS_{(CALCULATED)}$ are computed depending on the hypothesis being tested. The only difference in invoking test statistics in our decision rule is that we reject the null (called "statistically significant") when $TS_{(CALCULATED)}$ is greater than $TS_{(CRITICAL)}$. However, comparing p's and TS's for a given data set will always yield the same decision.

Remember, knowing sample results are relatively unlikely, assuming the null is true, may not be helpful. An improbable result is not necessarily an important result, as Shaver (1985, p. 58) illustrates in his hypothetical dialogue between two teachers:

> Chris: ...*I set the level of significance at .05, as my thesis advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that surely must be important.*
>
> Jean: *Wait a minute. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?*

WHY NOT USE STATISTICAL SIGNIFICANCE TESTING?

Statistical significance testing may require an investment of effort that lacks a commensurate benefit. Science is the business of isolating relationships that (re)occur under stated conditions, so that knowledge is created and can be cumulated. But statistical significance does not adequately address whether the results in a given study will replicate (Carver, 1978). As scientists, we must ask (a) what the magnitudes of sample effects are and (b) whether these results will generalize; statistical significance testing does not respond to either question (Thompson, in press). Thus, statistical significance may distract attention from more important considerations.

MISINTERPRETING STATISTICAL SIGNIFICANCE TESTING

Many of the problems in contemporary uses of statistical significance testing originate in the language researchers use.

Several names can refer to a single concept (e.g., "$SOS_{(BETWEEN)}$"= "$SOS_{(EXPLAINED)}$" = "$SOS_{(MODEL)}$" = "$SOS_{(REGRESSION)}$"), and different meanings are given to terms in different contexts (e.g., "univariate" means having only one dependent variable but potentially many predictor variables, but may also refer to a statistic that can be computed with only a single variable).

Overcoming three habits of language will help avoid unconscious misinterpretations:

- **Say "statistically significant" rather than "significant."** Referring to the concept as a phrase will help break the erroneous association between rejecting a null hypothesis and obtaining an important result.
- **Don't say things like "my results approached statistical significance."** This language makes little sense in the context of the statistical significance testing logic. My favorite response to this is offered by a fellow editor who responds, "How did you know your results were not trying to avoid being statistically significant?".
- **Don't say things like "the statistical significance testing evaluated whether the results were 'due to chance'."** This language gives the impression that replicability is evaluated by statistical significance testing.

WHAT ANALYSES ARE PREFERRED TO STATISTICAL SIGNIFICANCE TESTING?

Two analyses should be emphasized over statistical significance testing (*Journal of Experimental Education,* 1993). First, effect sizes should be calculated and interpreted in all analyses. These can be r squared-type effect sizes (e.g., R squared, eta squared, omega squared) that evaluate the proportion of variance explained in the analysis, or standardized differences in statistics (e.g., standardized differences in means), or both. Second, the replicability of results must be empirically investigated, either through actual replication of the study, or by using methods such as cross-validation, the jackknife, or the bootstrap (see Thompson, in press).

RECOMMENDED READING

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304-1312.

Journal of Experimental Education. (1993). Special Issue--"The role of statistical significance testing in contemporary analytic practice: Alternatives with comments from journal editors". Washington, DC: Heldref Publications. (Available from ERIC/AE).

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan*, 67(1), 57-60.

Thompson, B. (in press). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*.

Descriptors: Data Analysis; Data Interpretation; Decision Making; *Effect Size; Hypothesis Testing; Probability; Research Methodology; Research Problems; *Sampling; *Statistical Analysis; *Statistical Significance; Test Interpretation; *Test Use; *Testing

Citation: Thompson, Bruce (1994). The concept of statistical significance testing. *Practical Assessment, Research & Evaluation*, 4(5). Available online: http://PAREonline.net/getvn.asp?v=4&n=5.