Deakin University
Faculty of Health and Behavioural Sciences
School of Psychology

What is the meaning of $p<0.05$?

Jessica Jacques

Submitted as an assessment for Research Methods C (HPS 742)

Due Date: Tuesday 26th April 2005

Word Count: 1, 497

Hypothesis testing based on statistical significance has dominated behavioural and social science graduate programs for over 40 years and as a current psychology student I can promise you it still does (Huberty, 1996). A closer review of the history and current status of our beloved significance tests and their computed *p* value, revealed to me that one can probably say that few methodological issues in social science research have generated as much controversy. In fact as Anderson, Burnham, & Thompson (2000) note, across the years and throughout disciplines, the frequency of published criticisms has grown substantially. However before the feeling of the rug being pulled from underneath overwhelms me, one must ask the question what in fact is the meaning of the *p* value such as 0.05? Put more precisely what is being tested and where does it fit into data analysis and research if at all? This essay will unravel what exactly is tested by statistical significance tests, the role of replicability to the progression of scientific knowledge, significance testing based on the falsificationist approach to science and the topic of effect sizes.

Currently, most researchers will employ hypothesis testing involving a computation of the *p* value (Thompson 1996). However the problem arises too frequently that often researchers (and publishers of these papers) do not understand what their computed p-values actually represent (Carver 1993). Therefore we must turn to the question what are p-values and what do they measure? Currently traditional inferential statistics taught include the *t* ratio, the *F* ratio, the chi-square analysis, analysis of variance (ANOVA) and additional methods that test statistical significance. These procedures result in the same

decisions as with the use of the *p* value however with such modernized software packages exact values of either form can be easily obtained. In all procedures the probability is determined at a specified level called alpha (usually at the .05 level) of a particular result, presuming the null hypothesis is true of the population, given random sampling and assignment and with a sample of size n (Shaver 1993). So what exactly does this long-winded sentence mean? To put more precisely this essay will extend on and highlight the imperative elements.

From a falsificationist approach

In a normative account of science, theory is taken to be the starting point for the scientific process. Falsificationism tests whether these theories are scientific or not by whether they allow falsifiable hypotheses/predictions to be made and tested. This falsificationist approach to science as postulated by Popper is based on a form of reasoning, namely Aristotles modus tollens, which is to deny the antecedent by denying the consequent (Chalmers, 1999). When applied to the reasoning of statistical significance tests it follows that: If the null hypothesis is true of the population, then statistical significance would probably not occur, statistical significance has occurred, therefore the null hypothesis is probably not true (Cohen, 1994). This reasoning can be found to appear implicitly in the bulk of research literature from which I have no doubt is the result of the explicit teachings of statistic textbooks. The problem is that this formulation is inherently invalid as it is probabilistic not absolute thus leading to a result that is not sensible (Cohen, 1994). To achieve the Popperian principle we need to represent our theories as null hypotheses and attempt to falsify them and this is

what Meehl (1967) calls strong testing. The problem with the testing procedure as outlined in this essay is that we confirm our theories by rejecting the null hypothesis.

The null hypothesis

At this point it is important to clarify the meaning of the null hypothesis. In general terms it provides propositions of the population expressed as a specified value of a population parameter. Whilst the value can take any form (such as a mean difference, a proportion, a correlation etc) in the bulk of literature it is taken to mean zero or no difference (Cohen, 1986). Although it would be exceedingly rare that there would be no difference in the population (Shaver, 1993). It is important to highlight *p* values that yields statistical significance do not indicate the probability that the null hypothesis is true or false given a particular data (Shaver**,** 1993**)**. This probability is only made available through Bayesian statistics, which combines the likelihood, which is based solely on the observed data from the sample, with information known before the experiment. In the future Bayesian estimation may be a positive step towards better data analysis however and it is not until our theories evolve that we will really be able to make use of them (Cohen, 1986). Confidence intervals, which also reveal information about the outcome of hypothesis tests provide a plausible range for which you can be confident the population parameter will lie (Everitt, 2001). Although recommendations have been made over and over again that they should be considered as a means for which statistical results are presented, it still remains as something researchers 'ought' to do (Gardner and Altman, 1986).

Next, statistical significance tests result in an evaluation of the probability expressed in terms of it being more or less than a pre-specified alpha level (at the expense of assessing the results in other ways) that is, significant or non-significant (Thompson, 1993). I shall briefly turn my attention to the problem with this kind of criteria. As statistical significance tests are predominantly but not exclusively determined by the sample size (Shaver, 1993), a study that involves a sample size of 4,000 with a reported correlation of .03 will result in a statistically significant result but in theory this is NOT a significant predictor. Conversely a study with a sample size of 25 and a large reported effect size of .48 at an alpha level of 0.05 would not be statistically significant. This example shows that both trivial and important results may be statistically significant. As it is, for the results of the study to be statistically significant all the researcher needs is enough participants.

Third, tests of statistical significance must meet randomness assumptions, including random selection and assignment. If these assumptions are not met the study will not result in a meaningful or valid probability statement (Shaver 1993). The fourth element that I shall address as mentioned previously is the assumption that the null hypothesis is true in the population. This means that the commonly used statistical significance tests as mentioned earlier and *p* values make inferences from the population to the sample NOT vice-versa (Schmidt, 1996). However this is not what a 4th year psychology student has been led to believe nor does a researcher want statistical significance tests to do. As scientists we want to

draw samples and deduce inferences about the population from which they came. Thus providing knowledge about result replicability (Thompson, 1996).

Importance of replicability

For science to progress it must build knowledge regarding stable relationships and the only way this can occur is by providing information about result replicability (Vacha-Haase, 2001). However if statistical significance tests do not actually test the population further evidence of the studies replicability is crucial for the research to be scientifically enterprising. Whilst the best results for providing this kind of information involve yielding a similar replication of results, the next best action may be the reporting of effect sizes (Vacha-Haase, 2001) and it is this topic for which I shall now turn my attention.

Effect Sizes

As we have illustrated thus far calculated p values and their equivalent are a function of several effects, particularly sample size. Statistical tests do not then tell us anything about the magnitude or importance of a result (Shaver, 1996). Effect sizes, independent of sample size and measurement scale provide a numeric value for small, medium and large effect sizes (.2, .5, and .8, respectively) calculated by obtaining the difference between the value specified in the null hypothesis and value specified in the research hypothesis (Hinkle, Wiersma & Jurs, 2003). It is important to note that effect sizes whilst provide important information for data analysis, they do not offer a solution for which we can use instead of the problematic .05 statistical significance for they too should be interpreted with caution (Shaver, 1996).

So should the *p* value be abandoned completely? Whilst a majority of the literature on the topic would echo a resounding yes, I believe for psychology this may not be entirely sensible. I believe that such tests utilised in a correct and exploratory way give us hints of the existence of possible relationships and an evaluation of the data to estimate one or more parameters (Schafer, 1996) although some of the assumptions underpinning the tests may be invalid and illogical. If we are to know the full extent of what we are actually testing it is often possible to assess if the result is clear and accurate. Finally if nothing else *p* values are needed for a students understanding due to its centrality in the majority of psychological literature (Everitt, 2001)

My conclusion is that at this point there is no magical solution to the problem of statistical significance tests dominating the research field. However I believe that essays like these that demand consideration of it's inherent flaws and the more logical alternatives such as confidence intervals and effect sizes will serve as a starting point for a reform of data analysis methods in psychology.

References

Anderson, D. R., Burnham, K.P., & Thompson, W. L. (2000). Null hypotheisis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Biskin, B. H. (1998). Comment on significance testing. Measurement and Evaluation in Counseling and Development, 31, 58-62.

Carver, R. P. (1993). The case against statistical significance testing revisited. *Journal of Experimental Education, 61*, 287-292

Chalmers, A. F. (1999). *What is this thing called science?* (3rd ed.). University of Queensland press

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

Everitt, B. S. (2001). *Statistics for psychology: An intermediate course*. London

Gardner, M. J., and Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British medical journal*, *292*, 746-747

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed). Boston: New York.

Huberty, C. J. (1993). Historical Origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of science,* 34, 103-115.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education, 61,* 383-387.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 1082-1086

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61,* 219-224