Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 33

11-1-2002

Some Reflections On Significance Testing

Thomas R. Knapp

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 240-242

1538 – 9472/02/$30.00

Some Reflections On Significance Testing

Thomas R. Knapp
Kailua-Kona, Hawaii

This essay presents a variation on a theme from my article "The use of tests of statistical significance", which appeared in the Spring, 1999, issue of *Mid-Western Educational Researcher*.

Key words: significance tests; confidence intervals

Introduction

In addition to $.25 Senior Coffee at McDonald's, one of the few advantages of being old at the beginning of the 21st century is that you have actually lived through certain events (World War II comes immediately to mind), rather than reading about them in history books.

An interesting statistical event that I have lived through is the controversy regarding the use of tests of significance. As David Salsburg (2001) points out in his book, *The lady tasting tea*, that controversy started in the 1930s as part of the ongoing feud between R.A. Fisher and Jerzy Neyman. It was resurrected about 35 years later with the publication of the book, *The significance test controversy*, edited by Morrison and Henkel (1970); and was revisited recently in a subsequent book entitled *What if there were no significance tests?*, edited by Harlow, Mulaik, and Steiger (1997), by a task force of the American Psychological Association (see Wilkinson, 1999), and elsewhere (e.g., Nickerson, 2000).

Thomas R. Knapp, Ed. D. (Harvard, 1959) is Professor Emeritus of Education and Nursing, University of Rochester and The Ohio State University. Email him at tknapp5@juno.com.

Much nonsense has been written in attempts to resolve this controversy. In what follows I would like to suggest a middle-of-the-road solution. I leave it to you, dear reader (as the late Ann Landers used to say), to decide whether or not my suggestion is more nonsense.

Significance testing vs. hypothesis testing

Some writers (see Huberty, 1987; Huberty & Pike, 1999) distinguish between significance testing (a la Fisher) and hypothesis testing (a la Neyman & Pearson). Although the distinction is sometimes important and sometimes not, for the purposes of this paper I will not make the distinction. Here, a significance test is something one uses to test statistical hypotheses. I will also not get into null vs. nil hypotheses or one-tailed tests vs. two-tailed tests. If you are interested in such things, I recommend that you read Cohen (1965), Cohen (1994), or almost any of the late Jacob Cohen's other work.

Significance tests vs. confidence intervals

Since most of the controversy revolves around this matter, I will concentrate on it, along with the associated matter of "effect sizes" and what to do about them. It has often been claimed

that confidence intervals subsume significance tests: If the hypothesized value of a parameter is outside of the interval, reject it; if it is inside the interval you can't reject it. (See, for example, Steiger & Fouladi's contention that "the significance test rejects at the α significance level if and only if the 1-α confidence interval for the mean difference excludes the value zero—1997, p. 226.) Unfortunately, it's not that simple, as Dixon and Massey (1983) and others have pointed out, especially when the parameter of interest is a population proportion or percentage, as the following example will illustrate.

An example

Suppose you were interested in the proportion of nurses who smoke cigarettes. (As a former holder of joint appointments in education and nursing in two different universities, I've always wondered why ANY nurses smoke!) Suppose further that you have rather limited resources and you must restrict your efforts to a relatively small population (all nurses in Rochester, New York, say) and a relatively small sample size (16, say) from same. You are familiar with some of the literature on cigarette smoking and some of the literature regarding the significance testing controversy, so you believe that you have two choices: (1) test the hypothesis that P, the population proportion, is equal to some number, say .25 (that's roughly the national average); or (2) put a confidence interval around p, the sample proportion. Let's assume that you decide on the latter choice, you draw your random sample of 16 nurses, and you find that one of the nurses smokes cigarettes.

Here is a summary of your results:

Sample n = 16 Sample p = .0625

Estimated standard error =

$$\sqrt{p(1-p)/n} = \sqrt{(.0625)(.9375)/16} = .0642$$

95% confidence interval = .0625 ± 1.96 (.0642) = .0625 ± .1258, i.e., from 0 (since you can't have a negative proportion) to .1883.

But something isn't quite right here. First of all, the normal approximation to the binomial doesn't work so well for sample sizes of 16. Secondly, the p for this particular sample is used to estimate the population P in the calculation of the standard error, so that's a problem, since the P for this population of nurses is unknown Finally, and perhaps most importantly, that standard error is almost certain to be an under-estimate of the "true" standard error. (It would be even worse if you just happened to draw a sample that consisted of no smokers, in which case the estimated standard error would be equal to zero!) As Wilcox (1996) and others have pointed out, you need special techniques to handle the small n, small p case.

So what? The "so what?" is that for examples like this the interval estimation approach DOES NOT subsume the hypothesis testing approach. The otherwise hypothesis-tested value of .25 is not inside the interval around your effect size of .0625 ("no effect" would be a proportion of 0), but that's not the right interval. It's too narrow. The standard error that would be used in significance testing would be a function of the .25, not the .0625.

Conclusion

Tom Knapp's bottom line

If you have hypotheses to test (a null hypothesis you may or may not believe a priori and/or two hypotheses pitted against one another), use a significance test to test them. If you don't, confidence intervals are fine.

I think that makes sense. Do you?

References

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology*, New York: McGraw-Hill.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, 49, (12), 997-1003. Reprinted in L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?*. Mahwah, NJ: Erlbaum.

Dixon , W. .J., & Massey, F. J. (1983). *Introduction to statistical analysis* (4th ed.). New York: McGraw-Hill.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.) (1997). *What if there were no significance tests?*. Mahwah, NJ: Erlbaum.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, 16 (8), 4-9.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. *Advances in Social Science Methodology*, 5, 1-22.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy*. Chicago: Aldine.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5 (2), 241-301.

Salsburg, D. (2001). *The lady tasting tea*. New York: Freeman

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no sgnificance tests?* Mahwah, NJ: Erlbaum.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego: Academic Press.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54 (8), 594-604.