SIGNIFICANCE TEST, CONFIDENCE INTERVAL, BOTH, OR NEITHER?

Thomas R. Knapp

Introduction

It is reasonably well-known that you can usually get a significance test "for free" by constructing a confidence interval around an obtained statistic and seeing whether or not the corresponding hypothesized parameter is "captured" by the interval. If it isn't inside the 95% confidence interval, for example, reject it at the .05 significance level and conclude that the sample finding is statistically significant. If it is, don't reject it; the sample finding is not statistically significant at that level. So if you want a significance test you can either carry it out directly or get it indirectly via the corresponding confidence interval.

If you want a confidence interval you can carry it out directly (the usual way) or you can get it indirectly by carrying out significance tests for all of the possible "candidates" for the hypothesized parameter (not very practicable, since there is an infinite number of them!).

But should you ever carry out a hybrid combination of hypothesis testing and interval estimation, e.g., by reporting the 95% confidence interval and also reporting the actual p-value that "goes with" the obtained statistic, even if it is greater than or less than .05? Some people do that. Some journals require it.

It is also reasonably well-known that if you don't have a random sample you really shouldn't make any statistical inferences. (Just get the descriptive statistic(s) and make any non-statistical inferences that may be warranted.) Exception: If you have random assignment but not random sampling for an experiment, randomization tests (permutation tests) are fine, but the inference is to all possible randomizations for the given sample, not to the population from which the sample was [non-randomly] drawn.

In what follows I will try to convey to you what some of the practices are in various disciplines, e.g., education, nursing, psychology, medicine, and epidemiology (the disciplines that I know best). I will also give you my personal opinion of such practices and in an appendix to this paper I will provide a brief test of the correctness of various wordings of statistical inferences.

The significance test controversy

Up until about 40 years ago or thereabouts, traditional significance tests were about the only statistical inferential methods that were used. There were occasional arguments among research methodologists concerning the approach of R.A. Fisher vs. that of Jerzy Neyman and Egon Pearson; see, for example, the interesting discussion between Berkson (1942, 1943) and Fisher (1943) regarding the linearity of a set of data, and Salzburg's (2001) fascinating account of Fisher's conflicts with Karl Pearson (Egon's better-known father) and with Neyman. [Huberty (1987) has referred to Fisher's approach as significance testing and Neyman & Pearson's approach as hypothesis testing.] There were also a few researchers (e.g., Meyer, 1964), who argued in favor of the use of Bayesian inference, but most articles published in the professional journals continued to emphasize traditional significance testing.

That all started to change when Morrison and Henkel (1970) compiled a book with the same title as that of this section. The individual chapters were written by various people who were concerned about the overuse and/or misuse of significance tests, especially in sociology, along with a few defenders of the status quo. Things really came to a head in the late 80s with the publication of a chapter by Woolson and Kleinman (1989) regarding practices in medicine and epidemiology, and in the late 90s with the appearance of the book, What if there were no significance tests?, edited by Harlow, Mulaik, and Steiger (1997), with an emphasis on psychology and education. The latter work, like the Morrison and Henkel book, consisted of chapters written by people with different points of view, most of whom argued that significance tests should be replaced by confidence intervals around the corresponding "effect sizes".

[Interesting aside: Berkson's 1942 article (but not Fisher's response or Berkson's rejoinder) was included in Morrison and Henkel's 1970 book and was also cited in Cohen's 1994 article--see below--that was reprinted in the Harlow, et al. 1997 book.]

In the last ten years or so, confidence intervals have begun to replace significance tests, but significance tests still have their defenders. In epidemiology and medicine, and to a lesser extent in nursing, there has been a recent tendency to emphasize interval estimation (usually 95% confidence intervals) while at the same time reporting a variety of p-values that correspond to the area(s) in the tail(s) of the relevant sampling distribution(s).

Confidence intervals: The alleged panacea

One of the arguments against significance tests has been that many users of them botch the wording when they report the results of their studies. For example, many methodologists have rightly objected to statements such as "the probability is less than .05 that the null hypothesis is true". [Cohen (1994) made

the unfortunate mistake of claiming that some people say "the probability is less than .05 that the null hypothesis is false". I've never heard anyone say that.] The null hypothesis is either true or false. There is no probability associated with it, at least in the classical, non-Bayesian context. The probability applies to the likelihood of the sample finding, given that the null hypothesis is true; i.e., it is a conditional probability.

The claim is often made that the wording of confidence intervals is much more straightforward, and researchers are less likely to say the wrong things. Not so, say Cumming (2007), Cumming and Finch (2005), Moye (2006), Sober (n.d.), and others. For every user of significance tests who says "the probability is less than .05 that the null hypothesis is true" you can find some user of confidence intervals who says "the probability is .95 that my interval includes the parameter". Your particular interval doesn't have that .95 probability; the probability, if that word is even relevant for confidence intervals, applies to all such intervals created in the same way.

The one sense in which confidence intervals constitute a panacea is that you don't have to do any hypothesizing beforehand! Researchers often find it difficult to specify the magnitude of a parameter in which they are interested, whether the basis for that specification be theory, previous research, or whatever. With interval estimation all you need to do is specify the confidence you want to have and the margin of error that is tolerable (usually the half-width of the confidence interval), and the requisite sample size for "capturing" the parameter can be determined.

<u>One size confidence interval, different p-values</u>

There recently appeared two articles concerned with smoking cessation efforts, one in the medical literature (Peterson, et al., 2009) regarding teenagers who smoke, and one in the nursing literature (Sarna, et al., 2009) regarding nurses who smoke. Although the former was a randomized clinical trial and the latter was an observational study, both used the same statistical inferential approach of constructing 95% confidential intervals throughout, accompanied by actual p-values.

The principal finding of the Peterson study was "an intervention effect on 6-month prolonged smoking abstinence at 12 months after becoming intervention eligible (21.8% vs 17.7%, difference = 4.0%, 95% CI = – 0.2 to 8.1%, P =.06" (page 1383). [They called that "almost conclusive evidence" (same paragraph, same page).] Two supplementary findings were: "Among female and male smokers, respectively, the corresponding intervention effects were 5% (95% CI = 0.5 to 10%, P = .03) and 2.9% (95% CI = – 4.3 to 9.7%, P = .41)" (also same paragraph, same page).

One of the principal findings of the multiple logistic regression analysis reported in the Sarna study, comparing "any quit attempt" with "no quit attempt" (the dichotomous dependent variable) for smokers of 10-19 cigarettes per day vs. smokers of 20+ cigarettes per day (one of the independent variables) was an odds ratio of 2.43, 95% confidence interval 1.07 to 5.52, P = .03 (Table 4, page 253). Another finding from that analysis for another independent variable, baccalaureate vs. graduate degree, was an odds ratio of 1.54, 95% confidence interval 0.65 to 3.66, P = .33 (same table, same page).

What we have here in both articles is what I referred to earlier in this paper as a hybrid combination of constant confidence interval and varying p-values. I personally don't like it. If the authors are concerned solely with 95% confidence intervals I think they should be concerned solely with .05 p-values. In the Peterson study, for example, the 95% confidence interval for that difference of 4.0% in prolonged smoking abstinence [it should be 4.1%] didn't include an odds ratio of 1.00, so of course p is less than .05. Should the reader of the article care that p is actually .03? I don't think so. [And I don't think they should have used the phrase "almost conclusive evidence"!] The only justification I can see for reporting actual p-values in conjunction with 95% confidence intervals is the incorporation in a meta-analysis with p-values from other studies carried out on the same topic.

No significance tests or confidence intervals

Whether to use significance tests, confidence intervals, or both, pales in comparison to the more serious matter of the appropriateness of any inferential statistics. The standard gospel is easy to espouse: Use traditional inferential statistics if and only if you have a random sample from a well-defined population. So what's the problem?

First of all, there are researchers whom I call the "regarders", who don't have a random sample but who like to think of it as a sample from which a statistical inference can be made to a population of entities "like these". They refuse to quit after reporting the descriptive statistics, apparently because they find it difficult and/or unsatisfying to interpret the data without the benefits of inferential statistics. (Example: Sarna, et al., 2009. But they're not the only ones; it is clearly the modal approach in the research literature in education, nursing, psychology, medicine, and epidemiology.)

Secondly, there are the "populations are samples, too" advocates, who insist on carrying out some sort of statistical inference when they actually have data for an entire population. (Example: The negative correlation between land area and number of inhabitants for the 50 states is statistically significant at the .05 level.) The inference is allegedly from a population at one point in time to that same population at other points in time, even though the time point has not been

selected at random. (See the article by Berk, Western, & Weiss, 1995 about this, along with the various reactions to that article in the same journal.)

Then there are the "random is random" folks who use traditional t-tests or other general linear model techniques to carry out significance tests, rather than randomization (permutation) tests, when they have random assignment but do not have random sampling. Edgington and Onghena (2007) and others (e.g., Ludbrook & Dudley, 2000) have tried to get people to stop doing that, but to little avail. [See also the articles by Levin (1993) and by Shaver (1993), who come down on opposites of the matter.] A traditional t-test can occasionally be used as an *approximation* to a randomization test, if the researcher does not have easy access to the computer software that is necessary for carrying out a randomization test.

Shortly after Morrison and Henkel (1970) compiled their book, the famous statistician John W. Tukey (1977) wrote his treatise on Exploratory data analysis. In that book he claimed that descriptive statistics had been given short shrift and researchers should "massage" their data more carefully before, or instead of, carrying out statistical inferences. He provided several techniques for summarizing sample data, e.g., stem-and-leaf diagrams and q-q plots, that help to bring out certain features in the data that other descriptive statistics do not, and inferential procedures can not. I agree with Tukey's claim about descriptive statistics getting short shrift [but I'm not attracted to some of his statistical graphics]. I have even seen articles that provide an analysis of variance (ANOVA) summary table but not the sample means that produced it!

A final note

In this paper I have alluded to some criticisms that I have made in previous sources (Knapp, 1970; 1998; 1999; 2002). I could go on and on regarding some controversies regarding other practices. For example, why do some people test the statistical significance of baseline differences between experimental and control groups in a randomized experiment? (Don't they trust probability to balance the groups, and don't they understand that the significance test takes care of chance differences?) Or how about one-sided vs. two-sided significance tests and confidence intervals? (See Cohen's delightful 1965 piece about an argument between Doctor One and Doctor Two). But I wanted to keep this short and sweet. I know it's short. I hope you've found it to be sweet.

References

Berk, R.A., Western, B., & Weiss, R.E. (1995). Statistical inference for apparent populations. Sociological Methodology, 25, 421-458.

Berkson, J. (1942). Tests of significance considered as evidence. Journal of the American Statistical Association, 37 (219), 325-335.

Berkson, J. (1943). Experience with tests of significance: A reply to Professor R.A. Fisher. Journal of the American Statistical Association, 38 (222), 242-246.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), Handbook of clinical psychology (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49 (12), 997-1003.

Cumming, G. (2007). Pictures of confidence intervals and thinking about confidence levels. Teaching Statistics, 29 (3), 89-93.

Cumming, G., & Finch, S. (2005). Confidence intervals and how to read pictures of data. American Psychologist, 60 (2), 170-180.

Edgington, E.S., & Onghena, P. (2007). Randomization tests (4th. ed.). New York: Chapman&Hall/CRC.

Fisher, R.A. (1943). Note on Dr. Berkson's criticism of tests of significance. Journal of the American Statistical Association, 38 (221), 103-104.

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.

Knapp, T.R. (1970). A scale for measuring attitudes towards the use of significance tests. [The "original"] Educational Researcher, 21, 6-7.

Knapp, T.R. (1998). Comments on the statistical significance testing articles. Research in the Schools, 5 (2), 39-41.

Knapp, T.R. (1999). The use of tests of statistical significance. Mid-Western Educational Researcher, 12 (2), 2-5.

Knapp, T.R. (2002). Some reflections on significance testing. Journal of Modern Applied Statistical Methods, 1 (2), 240-242.

Levin, J. R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61, 378-382.

Ludbrook, J., & Dudley, H.A.F. (2000). Why permutation tests are superior to t- and F-tests in biomedical research. American Statistician, 54, 85-87.

Meyer, D.L. (1964). A Bayesian school superintendent. American Educational Research Journal, 1 (4), 219-228.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). The significance test controversy. Chicago: Aldine. [Reprinted in 2006.]

Moye, L.A. (2006). Statistical reasoning in medicine: The intuitive p-value primer (2nd. ed.). New York: Springer.

Peterson, A.V., Jr., Kealey, K.A., Mann, S.L., Marek, P.M., Ludman, E.J., Liu, J., & Bricker, J.B. (2009). Group-randomized trial of a proactive, personalized telephone counseling intervention for adolescent smoking cessation. Journal of the National Cancer Institute, 101 (20), 1378-1392.

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

Sarna, L., Bialous, S., Wewers, M.E., Froelicher, E.S., Wells, M.J., Kotlerman, J., & Elashoff, D. (2009). Nurses trying to quit smoking using the internet. Nursing Outlook, 57 (5), 246-256.

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

Sober, E. (n.d.) What does a confidence interval mean? 1-4. [Retrievable from the internet.]

Tukey, J.W. (1977). Exploratory data analysis. New York: Addison Wesley.

Woolson, R.F., & Kleinman, J.C. (1989). Perspectives on statistical significance testing. Annual Review of Public Health, 10, 423-440.

See www.tomswebpage.net/images/sigconf.doc

Appendix: The wording of significance tests and confidence intervals

The situation (adapted from Cumming & Finch, 2005): You are interested in the verbal ability of a population of school children, and you have administered a test of verbal ability to a sample of 36 children drawn randomly from the population. You are willing to assume that the test scores are normally distributed in that population. The sample mean is 62 and the sample standard deviation (with division by n-1) is 30. For those data the estimated standard error of the mean is 5. A two-sided t-test of the hypothesis that the population mean is 70 (the test of verbal ability has been normed on a different population having that mean) produces a two-tailed p-value of .12. Using a critical value of t = 2.03, the two-sided 95% confidence interval extends from 51.85 to 72.15.

On a scale from 1 to 3 (where 1= just plain wrong, 2= wrong but generally conveys the right idea, and 3 = correct), rate each of the following wordings for the significance test and for the confidence interval:

1. The probability is .12 that the sample mean is 62.

2. The probability is less than .12 that the sample mean is 62.

3. The population mean is 70.

4. The probability is .12 that the population mean is 70.

5. The probability is less than .12 that the population mean is 70.

6. The population mean is not 70.

7. The probability is less than .12 that the population mean is not 70.

8. If you were to test another random sample of 36 schoolchildren from that same population, their sample mean would be 62.

9. If you were to test another random sample of 36 schoolchildren from that same population, the probability is .12 that their sample mean would be less than 70.

10. If the population mean is 70, the probability is less than .12 that you would get a sample mean that differs from the population mean by 8 points or more.

11. You are 95% confident that the sample mean is between 51.85 and 72.15.

12. The population mean is greater than or equal to 51.85 and less than or equal to 72.15.

13. The probability is .95 that the population mean is between 51.85 and 72.15.

14. You are 95% confident that the population mean is between 51.85 and 72.15.

15. The probability is .95 that the interval from 51.85 to 72.15 includes the population mean.

16. You are 95% confident that the interval from 51.85 to 72.15 includes the population mean.

17. If you were to test another random sample of 36 schoolchildren from that same population, their sample mean would be between 51.85 and 72.15.

18. If you were to test another random sample of 36 schoolchildren from that same population, the probability is .95 that their sample mean would be between 51.85 and 72.15.

19. The 95% confidence interval includes all of those values of the population mean that would be rejected at the .05 level of significance.

20. 95% of intervals constructed in this same manner would include the population mean.

I won't give you the right answers (partly because there is room for some disagreement), but I'll give you the hint that these items would come fairly close to constituting a perfect Guttman Scale. If you don't know what that is, you can look it up!