A Review of Confidence Intervals

Anne-Marie Kimbell Mauk

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 28, 2000.

ABSTRACT

The present paper summarizes the recommendation that statistical significance testing be replaced or at least accompanied by the reporting of effect sizes and confidence intervals and discusses, in particular, confidence intervals. The recent report of the APA Task Force on Statistical Inference suggested that confidence intervals should always be reported.

A REVIEW OF CONFIDENCE INTERVALS

In 1996 the Task Force on Statistical Inference (TFSI) was convened by the Board of Scientific Affairs of the American Psychological Association to evaluate the applications of statistics used in psychological journals (Azar, 1997; Shea, 1996). The task force, which was instigated in part as a result of many years of discussion and disagreement over the use of statistical significance testing, recommended, among other things, revising the statistical sections of the American Psychological Association Publication Manual (APA, 1994). Prior to any revision of this manual, however, the Task Force printed a report in American Psychologist to encourage discussion regarding the subject. This was done in August, 1999, and included proposed guidelines, comments, explanations, and elaborations regarding the use of statistical methods and suggestions for the revision of the APA publication manual and developing related material (Wilkinson & The APA Task Force on Statistical Inference, 1999).

One of the proposed guidelines pertaining to analyzing results is to “always present effect sizes for primary outcomes” (Wilkinson et al., 1999, p. 599), which “enables readers to evaluate the stability of results across

samples, designs, and analyses" (p. 599). Related to this is the recommendation to also provide interval estimates for any effect size involving principal outcomes and "for correlations and other coefficients of association or variation whenever possible" (p. 599). Examining confidence intervals from related studies helps determine stability across studies (Schmidt, 1996), and "helps in constructing plausible regions for population parameters (Wilkinson et al., p. 599).

Many articles and books have been written detailing the flaws in and misuses of statistical significance testing (Chatfield, 1991; Cohen, 1994; Falk, 1998; McGrath, 1998; Oakes, 1986; Roozeboom, 1960; Schmidt, 1996; Steiger & Fouladi, 1997; Thompson, 1993, 1996, 1998), some calling for the actual banning of statistical significance testing (Carver, 1978; Cohen, 1994; Meehl, 1967; Schmidt, 1996). Those arguments will not be detailed again here; instead the present paper examines in some detail the Task Force recommendation that statistical significance testing be replaced or at least accompanied by the reporting of effect sizes and confidence intervals, and explains confidence intervals.

Reasons for Using Confidence Intervals

An article by Cohen (1994) written prior to the TFSI being convened (and partly responsible for the creation of the task force) pointed out that researchers tend pay too much attention to statistical significance testing and not enough to their conclusions about the actual meaning of their results, and should, to change this, “routinely report effect sizes in the form of confidence limits…. which contain all the information to be found in significance tests and more” (p. 1002).

Oakes (1986) found confidence intervals “infinitely preferable to tests of significance” (p. 66):

> Although the underlying logic is essentially similar they are not couched in the pseudo scientific hypotheses testing language of significance tests. They do not carry with them decision-making implications, but, by giving a plausible range for the unknown parameter, they provide a basis for a rational decision should one be necessary. Should sample size be inadequate this is signaled by the sheer width of the interval. (pp. 66-67)

Oakes (1996) also argued that “the researcher armed

with a confidence interval, but deprived of the respectability of statistical significance must work harder to convince himself and others of the importance of his findings. This can only be good" (p. 67).

In a 1994 article entitled "Misuse of Statistical Tests in Three Decades of Psychotherapy Research," Dar, Serlin, and Omer wrote that confidence intervals should be used when judging obtained effects. They noted "In drawing boundaries around obtained effects, confidence intervals provide essential information when estimating effect sizes in the population" (p. 80).

Schmidt (1996) detailed several reasons for using confidence intervals, the first of which is that "point estimates and confidence intervals provide a much more correct picture" (p. 121) than null hypothesis statistical significance testing. Another is that confidence intervals "hold the overall error rate to the desired level" (p. 121). Schmidt also reminded us that "prior to the appearance of Fisher's 1932 and 1935 texts, data analysis in individual studies was typically conducted using point estimates and confidence intervals" (p. 121).

Confidence intervals provide a graphical method for observing results of a study. The APA Task Force also

expressed a clear preference for graphical presentations of results, especially as regards confidence intervals:

> Figures attract the reader's eye and help convey global results. Because individuals have different preferences for processing complex information, it often helps to provide both tables and figures…In all figures, include graphical representations of interval estimates whenever possible. (p. 601, emphasis added)

Steiger and Fouladi (1997) wrote "In general, a confidence interval conveys more information, in a more naturally usable form, than a significance test. This is seen most clearly when confidence intervals from several studies are graphed alongside one another" (p. 227). Vertical or horizontal line segments can be placed through the graphed value of the statistic to show the confidence interval (Huck & Cormier, 1996).

Computing Classical Confidence Intervals

A point estimate (e.g., mean, *r*, *R*) is a number computed from a sample to represent a population parameter. Since there is some sampling error associated with this estimate, the true population parameter could be larger or smaller than the sample statistic. By identifying a range of possible values for the population parameter, the

researcher can control the probability that samples from the population will yield statistics approximating the values within a computed range of values. This range is called a confidence interval. For example, a 95% confidence interval can be computed using alpha = .05 such that 95% of the samples from the population would capture the population parameter (Cohen & Cohen, 1983).

Confidence intervals can be computed for any statistic (e.g., the sample mean, median, *r*, *R*). The critical component in computing a confidence interval is estimating the standard deviation of the sampling distribution (see Breunig, 1995; Rennie, 1997), which is called the "standard error." The standard error, and thus the boundaries for confidence intervals, can be estimated in either of two ways. First, the boundaries of a confidence interval can be computed based on *theoretical assumptions* about the shape of the sampling distribution (cf. Thompson, 1999). Second, the boundaries can be computed by *empirically estimating* the standard error, using a technique such as the "bootstrap" (cf. Lunneborg, 2000; Thompson, 1999).

The present paper focuses on the use of theoretically-based estimates of standard errors, for the sake of simplicity (albeit at possible loss of accuracy, because assumptions regarding sampling distribution shape may not

be perfectly met very often). Also for simplicity sake, the illustrations here involve only the statistic the mean (first one mean and then for the comparison of two means), even though it is emphasized once again that confidence intervals can be computed for any statistics. Of course, the computational formulas for computing intervals differ for different statistics.

Confidence Intervals for the Mean

The confidence interval has as its foundation the Central Limit Theorem, so when the sample size is large enough, generally over 30 ($n \geq 30$), the sampling distribution of the sampling mean is approximately normal. Ninety-five percent of a normal distribution falls within two standard deviations (1.96 exactly) of the mean; 99% of a normal distribution falls within three standard deviations of the mean.

Suppose a researcher wanted to determine the number of hospital treatment days necessary for adults undergoing withdrawal from alcohol. The sample population consists of 100 persons, who required an average of 13.42 days. This mean (X) is the point estimate. The confidence interval is computed around this estimate by using the formula:

$$c.i. = X \pm z(\sigma/ N\text{-}1),$$

where c.i. = confidence interval; X = the sample mean; *z* = the *z* value as determined by the alpha level; and s/ N-1 = the standard deviation of the sampling distribution, or the standard error, assuming that the sampling distribution is normally distributed.

Because the population standard error is unknown, the sampling distribution standard error formula can be used for a sample size of over 30. This substitution should not be made when the sample size is less than 30; *t*-statistics are used instead.

For a 95% alpha level, the corresponding *z*-value will be ± 1.96. The standard deviation of the sample of 100 numbers of days required for detox is 4.48. Therefore, the computation would be:

c.i.= 13.42 ± 1.96 (4.48/ 99).

c.i.= 13.42 ± 1.96 (.45)

c.i.= 13.42 ± .88

or c.i.= (12.54, 14.30)

Confidence intervals are generally expressed either by enclosing the two values in parentheses, separated by a comma, or by providing the point estimate plus or minus the margin of error. For this example, the estimate is that the average length of hospital days for detox for adult

alcoholics is somewhere between 12.54 and 14.30. Since 95% of all possible sample means are within 1.96 *z*'s (or .88 days) of the mean of the sample, the interval will probably contain the population mean. Only if the sample mean is one of the few that is more than 1.96 *z*'s from the mean of the sampling distribution will this interval fail to include the population mean.

Interval Width

The width of a confidence interval is related to (a) the statistical significance level set by the researcher, (b) the standard error, and (c) the sample size. The confidence interval will be wider the higher the percentage of accuracy desired. The confidence interval will be wider the larger the standard error. The confidence interval will be narrower the larger the sample size. The researcher has to make a determination about what risk to take in regards to being wrong--of not including the population value in the estimate--based on the nature of the research. A 99% confidence level using the above data would be constructed as follows:

$$\text{c.i.} = 13.42 \pm 2.58\ (4.48/\ 99)$$

$$\text{c.i.} = 13.42 \pm 2.58\ (.45)$$

$$\text{c.i.} = 13.42 \pm 1.16$$

or c.i.= 12.26, 14.58

The researcher can then be 99% confident that the average days of detox required for the population falls between 12.26 and 14.58. The goodness of this interval estimation procedure can be evaluated by "examining the fraction of times in repeated sampling that the intervals contain the parameter being estimated" (Mendenhall & Ott, 1980, p. 147). This fraction is called the confidence coefficient. This can be illustrated by drawing a number of different samples from the population and computing interval estimates using the formula described previously. Although the intervals will be different, most of them will contain μ (the population mean). If repeated over and over, approximately 95% of the intervals would contain μ (Mendenhall & Ott, 1980).

Huck and Cormier (1996) point out that the correct way to interpret confidence intervals is to imagine constructing many same-size samples from the same population, constructing confidence intervals separately around each sample's statistic, and then observing that "some of these intervals would 'capture' the parameter" (p. 140) and some of them will not. They note that "It would turn out that 95 percent of these 95 percent confidence intervals contain the parameter" (p. 140).

Confidence Intervals for Comparison of Two Means

The above are confidence intervals constructed around a single mean. Confidence intervals can also be constructed for the difference between two means, a single contrast on means, a single variance, the ratio of two variances, a single correlation, and a single proportion.

To compare two means and construct confidence intervals, the formula is as follows:

$$c.i. = (X_2 - X_1) \pm z\sigma_{x2-x1},$$

where $X_2 - X_1$ = difference between the sample means, and σ_{x2-x1} = standard error of the sampling distribution of the estimated difference between the sample means.

The *z*-score will depend, as usual, on the confidence coefficient determined by the researcher. The standard error is estimated by the following: “When two estimates are formed from independent samples, the sampling distribution of their difference has variance equal to the sum of the variances of the sampling distributions of the separate estimates” (Agresti & Finlay, 1997, p. 213).

Therefore:

$$\sigma_{x1-x2} =$$

For a large sample the formula will be:

$$(X_2 - X_1) \pm z$$

again using the sample standard deviation instead of an unknown population standard deviation. For a small sample, the *t* distribution is substituted for the normal distribution, providing the formula:

$$(X_2 - X_1) \pm t$$

Formulas for computing confidence intervals for other statistics, some of them quite complex, can be located in statistical textbooks. The statistical computer program SPSS provides confidence interval computations quickly and easily, which allows the researcher to graphically observe and display results of their various statistical calculations if so desired.

Summary

As more and more journals require the reporting of effect sizes and confidence intervals, as is occurring slowly, the argument for providing confidence intervals becomes stronger. Oakes (1986) in defense of confidence intervals rather than significance testing, stated “Above all, interval estimates *are* estimates of effect size. It is incomparably more useful to have a plausible range for the value of a parameter than to know, what [sic] whatever degree of certitude, what single value is untenable” (p. 67).

Cohen (1994) who, along with Schmidt (1996), has called for replacing statistical significance testing with point estimates and confidence intervals, wrote “as researchers, we have a considerable array of statistical techniques that can help us find our way to theories of some depth, but they must be used sensibly and be heavily informed by informed judgment” (p. 1002).

REFERENCES

Agresti, A., & Finlay, B. (1997). Statistical methods for the statistical sciences (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Azar, B. (1997). APA task force urges a harder look at data. The APA Monitor, 28(3), 26.

Breunig, N. A. (1995, November). Understanding the sampling distribution and its use in testing statistical significance. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS (ERIC Document Reproduction Service No. ED 393 939)

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

Chatfield, C. (1991). Avoiding statistical pitfalls. Statistical Science, 6, 240-268.

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

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

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Dar, R.; Serlin, R. C.; & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy

research. Journal of Consulting and Clinical Psychology, 62(1), 75-82.

Falk, R. (1998). In criticism of the null hypothesis statistical test. American Psychologist, 53, 798-799.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). Applied statistics for the behavioral sciences (4th ed.). Boston: Houghton Mifflin.

Huck, S. W., & Cormier, W. H. (1996). Reading Statistics and Research (2nd ed.). New York: HarperCollins.

Lunneborg, C. E. (2000). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Duxbury.

McGrath, R. E. (1998). Significance testing: Is there something better? American Psychologist, 53, 796-797.

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

Mendenhall, W., & Ott, L. (1980). Understanding statistics (3rd ed.). North Scituate, MA: Duxbury.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Rennie, K. M. (1997, January). Understanding the sampling distribution: Why we divide by n-1 to estimate the population variance. Paper presented at the annual meeting

of the Southwest Educational Research Association, Austin. (ERIC Document Reproduction Service No. ED 406 442)

Roozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61, 350-360.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42 (49), A12, A16.

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 significance tests? (pp. 221-257). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 334-349.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. Invited address presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

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