DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 2

11-1-2004

Confidence Intervals On Subsets May Be Misleading

Juliet Popper Shaffer
University of California, Berkeley, shaffer@stat.berkeley.edu

Recommended Citation

Shaffer, Juliet Popper (2004) "Confidence Intervals On Subsets May Be Misleading," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 2.
DOI: 10.22237/jmasm/1099267320
Available at: http://digitalcommons.wayne.edu/jmasm/vol3/iss2/2

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 261-270

INVITED ARTICLES

Confidence Intervals On Subsets May Be Misleading

Juliet Popper Shaffer
University of California, Berkeley

A combination of hypothesis testing and confidence interval construction is often used in social and behavioral science studies. Sometimes confidence intervals are computed or reported only if a null hypothesis is rejected, perhaps to see whether the range of values is of practical importance. Sometimes they are constructed or reported only if a null hypothesis is accepted, in order to assess the range of plausible nonnull values due to inadequate power to detect them. Even if always computed, they are interpreted differently, depending on whether the null value is or is not included. Furthermore, many studies in which the null hypothesis is not rejected are never published (the "file drawer" problem). This article discusses the coverage probability of nominal $1-\alpha$ confidence intervals when examining intervals that do or do not cover some specified null value, usually zero. A briefer treatment considers interval coverage when undesirable results are suppressed. The coverage probability of such conditional confidence intervals may be very far from the nominal value. The magnitude of the effect of selection on interval coverage probability and possible resultant biases in inference are illustrated, and discussed in relation to effect sizes of importance in social and behavioral science research and to estimation of effect sizes.

Keywords: Hypothesis tests, selected confidence intervals, censored studies

Introduction

There has been an enormous amount of literature, much of it in the social sciences, recommending that confidence intervals always be constructed, either in addition to or instead of p-values or other information related to testing hypotheses. The purpose of this article is to point out a problem in interpreting confidence intervals when they are pertinent to a hypothesis of interest.

The correct interpretation of $1-\alpha$ confidence intervals is that these randomly-chosen intervals have probability $1-\alpha$ of covering the true values of the parameter being estimated. Given a set of intervals, on the average $1-\alpha$ proportion should cover the true values. However, it is often true that special attention is paid to intervals depending on their coverage. Often there is special interest in a particular value of the parameter involved, either zero (often in comparing two groups) or some specified nonzero value. This article will consider the situation in which zero is of special interest; the results generalize to any other value with only obvious changes.

In such cases of selective interest, special attention may be paid to intervals that

Juliet Popper Shaffer is Senior Lecturer Emeritus of Statistics. Her research is primarily in the area of linear models--regression and analysis of variance, multivariate analysis, and simultaneous inference. She also has a background in psychology and in educational measurement and methodology. Email: shaffer@stat.berkeley.edu.

don't include zero, in order to estimate the size of plausible parameter values. There may be special interest in intervals that are far from zero. Or on the contrary, special attention may be paid to intervals that do include zero, to see whether there might be differences of substantial interest that could be verified by more powerful studies. Usually the direction of departure from the null hypothesis is of special interest, and intervals in one or the other direction may be especially scrutinized.

Furthermore, it is well known that studies with insignificant results often are not reported, and therefore not known, as is sometimes true of studies with results in a direction opposite to that of the desires or expectations of the sponsoring organization. Then only some selected intervals are available to be considered.

As soon as there is special consideration of a subset of intervals based on the values they include, the probability that they cover the true parameter value, in other words their conditional coverage probability, may be considerably different from the nominal $1-\alpha$ probability that applies to the whole set.

Such conditional considerations apply to all situations in which confidence intervals are obtained. This article will give detailed quantitative results for the comparison of the means of two distributions, assuming independent, normally-distributed observations with equal variance and equal sample sizes. All quantitative results reported here for known variance apply also to the case of matched pairs of observations with variances of the matched differences known, given the appropriate one-sample test in that case, provided the tabled effect sizes are divided and tabled sample sizes multiplied by the square root of 2.

Section 1 will give a general overview of conditional probability coverage both when the intervals do and when they do not cover the value zero, with most attention on the former. The coverage depends on the noncentrality parameter, a function of the sample size and the effect size. Sections 2 and 3 will examine the coverage for effect sizes and sample sizes that are frequently encountered in social and behavioral science research: Section 2 primarily when zero is not covered, and Section 3 when intervals in one direction are not calculated or reported. Section 4 will discuss effect size estimation issues as they relate to conditional coverage. Section 5 discusses and summarizes the issues raised.

Comparing the means of two distributions: Conditional on coverage or noncoverage of a specified value

Consider two groups of independent, normally-distributed observations with equal, known variance, and of equal sample size. With unknown variance, the standard test of equality of the means is the two-sample t test. With known variance, the known value σ is used in place of the estimate s; the test statistic then has a normal distribution, and the resulting test will be referred to as the two-sample z test. Since the t distribution tends to the normal distribution as the number of degrees of freedom tends to ∞, the properties of the z test hold approximately for the t test when the variance is estimated with large degrees of freedom.

Suppose a $1-\alpha$ confidence interval is constructed for a difference between the means of the two groups, where $\alpha = .05$ is assumed throughout the paper. Consider separately the probability of covering the correct value for confidence intervals that do not include the value zero, and the same probability for confidence intervals that do include zero. Figure 1 gives the conditional coverage of those intervals, as a function of the noncentrality parameter, which is the standard effect size measure $(\mu_1-\mu_2)/\sigma$ (Cohen, 1962, 1988), multiplied by the square root of $n/2$, where n is the sample size of each group. Given the known sample size n of each group, the noncentrality parameter, and therefore the conditional coverage, is a function of the unknown true effect size.

What Figure 1 illustrates is the well-known fact that intervals that do not cover zero also have very small conditional probabilities (given that fact) of covering values close to zero (see, e.g., Olshen, 1973). Correspondingly, intervals that do cover zero are also more likely than the nominal confidence coefficient to cover values close to zero. These properties are true for intervals of fixed length as in this case, when the standard deviation is known. For the same true effect size, the coverage probabilities depart even further from the nominal values when the standard deviation is unknown and must be estimated, so that the size of the conditional intervals varies with the estimated standard deviation. In that case, for a given effect size, intervals that don't cover zero are likely to be shorter than intervals that do, so both location and interval length affect the conditional coverage. Figure 2 gives the correlation between the interval length and the probability that the interval includes the correct value, for t intervals with varying degrees of freedom.

Relation of conditional true value coverage and non-coverage to effect sizes and sample sizes frequently encountered in social-behavioral science research.

The noncentrality parameter that determines the coverage probability is a function of the known sample size n of each group and of the effect size. Thus, consideration of effect sizes is crucial in examining conditional confidence interval coverage. Of course, there is no direct way of making use of the quantitative information in a particular case, since the true effect size is unknown. However, many studies in the social sciences, as noted in Cohen's (1962) pioneering paper, support the assumption that effect sizes in these fields are often between .1 and .5. Cohen suggested the now-standard terminology of small effects = .2, medium effects = .5, and large effects = .8. He stated "Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined…" (Cohen, 1988, p. 13).

Examples of estimated effect sizes in the literature show many around .2 or less. For instance, Fukkuk and Glopper (1998), in a meta-analysis of studies of learning of word-meaning from context, found out of 22 effect size estimates that nine were smaller than .20, ten were between .21 and .40, and only three were greater than .40. Grissmer (1999), in a meta-analytic study of the effects of class size reduction on achievement, found effect sizes between .15-.25 for grades K-3, and .11-.20 for grades 4-7. Although researchers carrying out meta-analytic studies try to find as many studies as possible, it seems clear that it is easier to locate studies with significant effects, and thus probably larger real or apparent effect sizes, than those with insignificant effects, which may never have been reported. Furthermore, in the former study (Fukkuk and Glopper), it was noted that the data for some studies, even though the studies were found, could not be obtained. Thus, the obtained values reported above are likely to represent an upwardly-biased sample. It follows that even when all reported confidence intervals are considered equally, the available studies are likely to include an overabundance of intervals that do not include zero.

In summary, a small effect size of .2 or smaller is likely to be a feature of many studies of this kind, and furthermore, the reported values may be upwardly biased. Since the conditional coverage probability of confidence intervals is a function of the effect size, an examination of effect sizes in the range assumed to be common in social and behavioral science research, and their relation to conditional coverage, is called for.

Table 1 gives the coverage probability for the two-sample z Test, equal sample sizes, with sample sizes ranging from 5 per group to 50 per group, assuming effect sizes of .1 to .5, and assuming the null hypothesis is rejected, so that the intervals do not include zero. The values in parentheses are the probabilities of rejecting the null hypothesis for the associated sample size-effect size combination. All values hold approximately when variances are estimated with large degrees of freedom.

If the variance must be estimated from the information in the two sets of observations, the confidence coverage results are still further from the nominal values. When there are 5 observations per group, so that t is based on 8df , the first row entries in Table 1 would be

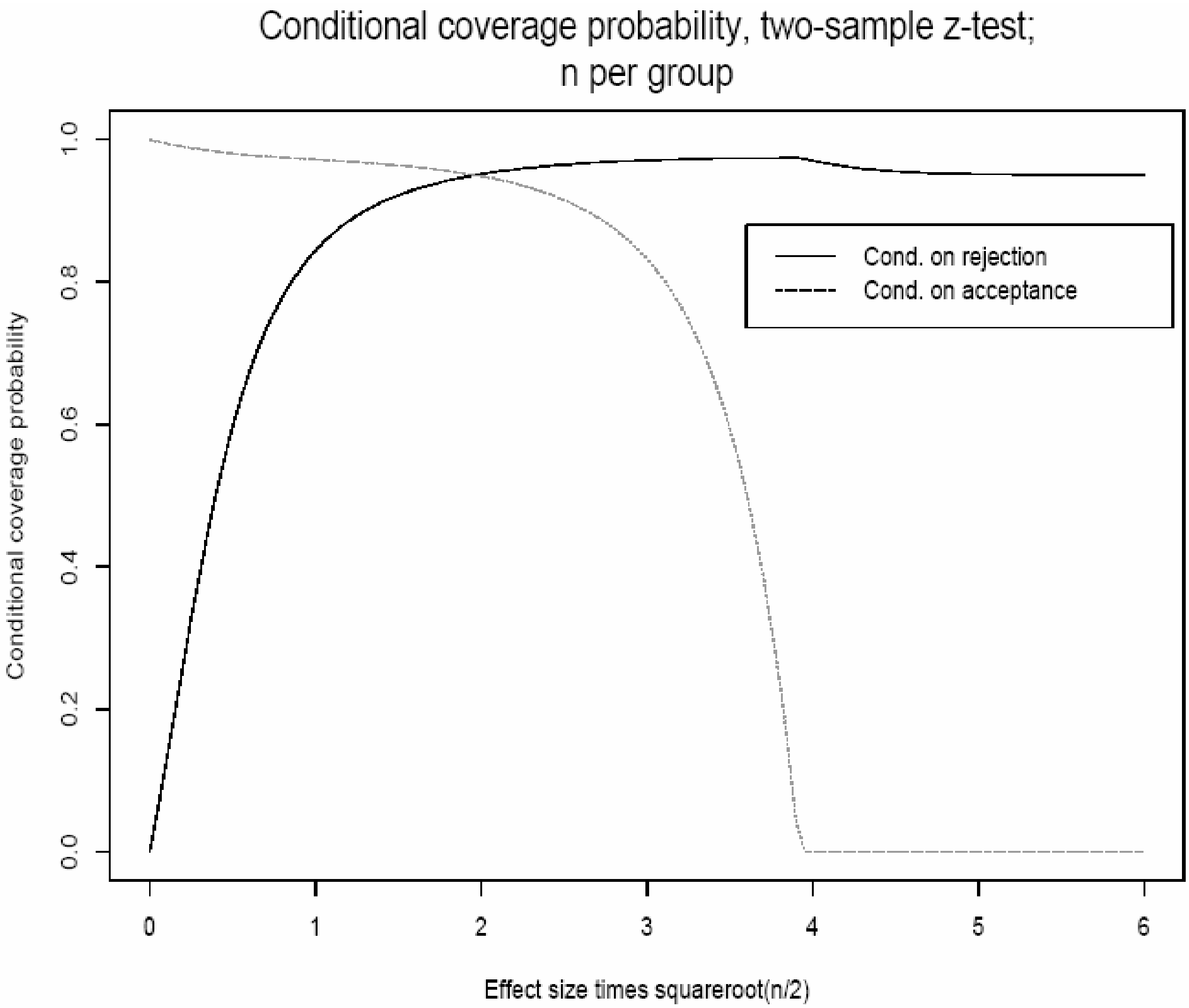

Figure 1: Conditional coverage probability, two-sample z test, n per group

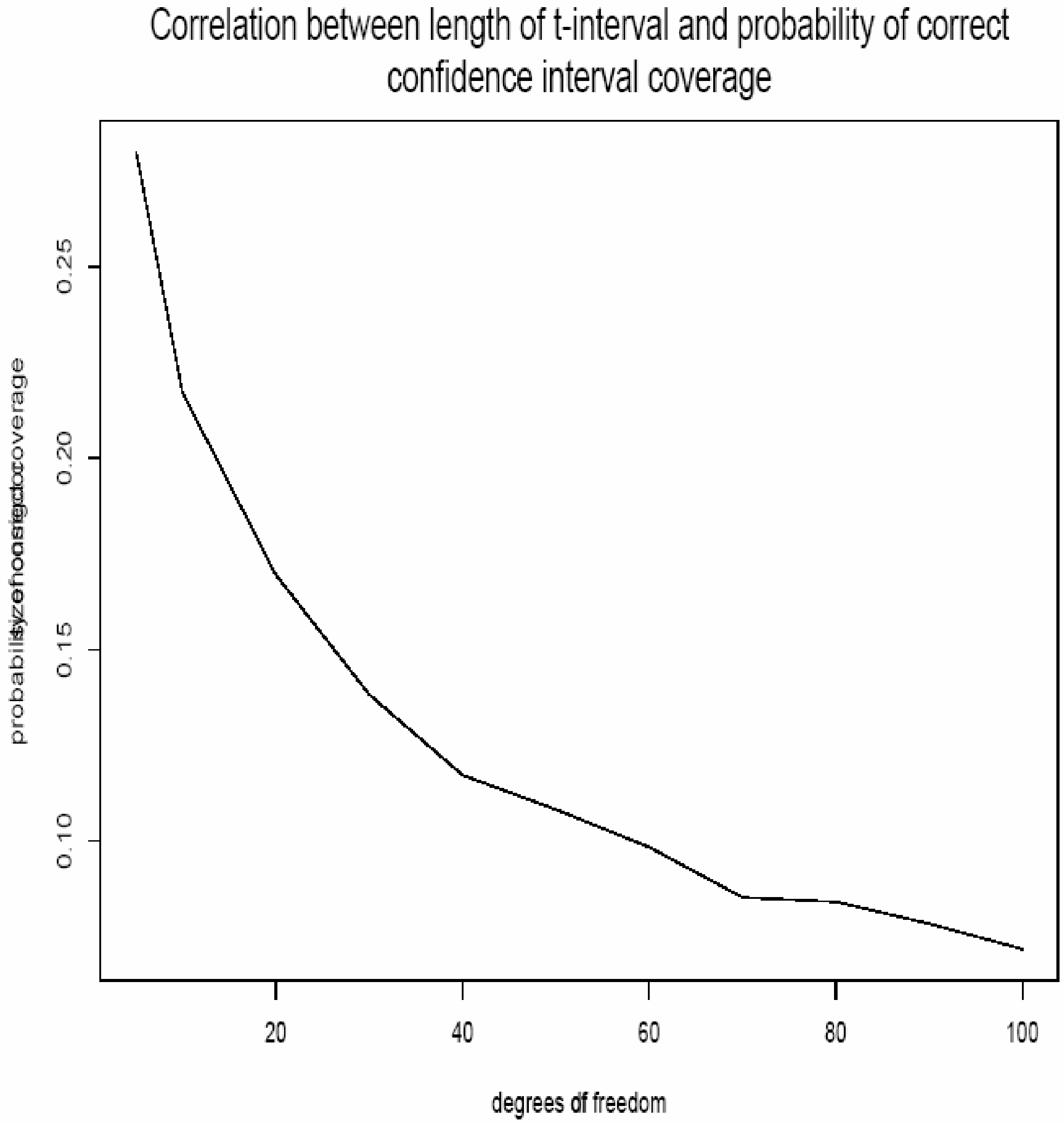

Figure 2: Correlation between length of t-interval and probability of correct confidence interval coverage

Table 1: True conditional probability that the nominal .95 confidence interval based on the z test covers the correct value, given rejection of the null hypothesis (values in parentheses are probabilities of rejection).

	Effect size				
Sample size	.1	.2	.3	. 4	.5
5	.20(.05)	.41(.06)	.57 (.07)	.69(.10)	.77 (.12)
10	.29(.05)	.55(.07)	.72(.10)	.81(.15)	.87(.20)
20	.41(.06)	.69(.09)	.83(.16)	.90(.24)	.93(.35)
30	.49(.07)	.77(.11)	.88(.21)	.93(.34)	.95(.48)
40	.55(.07)	.81(.14)	.91(.26)	.94(.43)	.96(.60)
50	.60(.08)	.84(.17)	.92(.32)	.95(.51)	.96(.70)

replaced by .18(.05), .36(.06), .52(.07) .64(.09), and .73(.11), respectively, and when there are 10 observations per group (18df), the second row entries would be replaced by .28(.06), .53(.07), .69(.10), .80(.14), and .86(.18), respectively. For larger df , the differences are very small, so the results for known variance are approximately correct.

For an effect size of .1, the probability of rejecting the null hypothesis, i.e. the probability that the interval does not include zero, is quite small, even for samples of size 50 in each group. However, if these cases are the ones that get attention, perhaps the only ones that get published, the extreme departure from the nominal coverage probability of the associated confidence intervals means that incorrect quantitative inferences are highly likely. Even for effect sizes larger than .1, the under-coverage of the intervals can be non-negligible, and the probability that the intervals don’t contain zero becomes much larger. As noted above, effect sizes within the range .1 to .3 are very common in social-behavioral science research.

Values that are covered when the true value is not covered

When intervals that do not include zero also do not include the true values, they will include either only values in the wrong direction from the true effect, smaller than the true effect in the correct direction, or, more likely with small effect sizes, values in the correct direction but farther away from zero than the true values. When the true effect is barely different from zero, clearly the probability of a range of values more extreme in the right direction and a range in the wrong direction will each be approximately .50. When the true effect is extremely large, the probability of ranges of values more extreme in the right direction and less extreme in the right direction will each be approximately .50. For the effect sizes and sample sizes in Table 1, the probabilities of intervals covering only smaller values in the correct direction are all equal to zero. Table 2 gives the conditional probabilities that the results do not cover the true values; the entries in parentheses are the expected proportion of these non-covering intervals that are in the right direction but more extreme. Subtracting these proportions from one gives the conditional probabilities of confidence intervals with ranges entirely in the wrong direction.

Note that for the smaller effect sizes and/or sample sizes in this table, the probability that the intervals do not cover the true values can be quite substantial, as can the probabilities that they cover values in the correct direction but larger. In some cases, the probability of intervals entirely in the wrong direction is non-negligible. Thus, the calculated intervals may lead to either incorrect directional inferences or unwarranted optimism about the true sizes of the effects under study.

It has been noted that when studies with insignificant effects are not reported, many studies in the literature claim real differences when in fact the null hypotheses are true. However, it is shown here that even when the null hypotheses are false, the confidence intervals are likely to indicate that the effect sizes are larger than they really are. This is true if special attention is paid to confidence intervals that do not include zero, even when there is no withholding of studies showing insignificant effects.

Suppose, however, that confidence intervals including zero are specially noted, in order to estimate the range of plausible nonzero values. When the true value is small, these intervals are likely to have probability higher than the nominal probability of covering true values, and thus also to give falsely optimistic impressions of possible null hypothesis departures.

If the variance must be estimated from the two samples themselves, the first row would be replaced by the values for t with 8 *df* : .82(.59), .67(.33), .48 (.74), .36 (.81), and .27 (.87), respectively, while the entries in the second row, replaced by the values for t with 18 *df* , would be .72(.63), .47(.75),.31(.84),.20(.91), and .14 (.95), respectively. For larger degrees of freedom, the values are very close to those for known variance. As for known variance, the probability of coverage in the correct direction but smaller than the true value is zero for the sample sizes and effect sizes in the table.

Conditioning when significant results in one direction only are noted

According to an Associated Press article in the September 9, 2004 *San Francisco Chronicle*, and also reported in other places, editors of 11 medical journals are adopting a policy requiring the results of all clinical studies to be made public, noting that "drug company-sponsored studies with negative results rarely are submitted to medical journals" (Tanner, 2004). In this case, "negative" means results contrary to the desires of the company. This can be interpreted in two ways, noted by (a) and (b) below.

(a) The results may be reported only if significant and in the direction desired by the company. If the results are significant, but the true value is in the direction that is not reported, then reported confidence intervals will have probability zero of including the correct value, and from Table 2 it is possible to calculate the probability of results in the false direction being reported (multiply the probability of rejection by the conditional probability of intervals in the incorrect direction, given rejection). If the true value is in the direction that is reported, the values in Table 1 are the probabilities that the reported intervals cover the true values.

Table 2: Conditional probability of noncoverage (of true values) of the nominal .95 confidence interval, and (in parentheses) the proportion of noncovering intervals containing larger values in the correct direction.

	Effect size				
Sample size	.1	.2	.3	.4	.5
5	.80(.59)	.59(.69)	.43(.77)	.31(.84)	.23(.89)
10	.71(.63)	.45(.76)	.28(.85)	.19(.92)	.13(.96)
20	.59(.69)	.31(.84)	.17(.93)	.10(.98)	.07(.99)
30	.51(.73)	.23(.89)	.12(.97)	.07(.99)	.05(1.00)
40	.45(.76)	.19(.92)	.09(.98)	.06(1.00)	.04(1.00)
50	.40(.78)	.16(.94)	.08(.99)	.05(1.00)	.04(1.00)

Table 3: True conditional probability that the nominal .95 confidence interval covers the correct value, as a function of effect size and sample size per group, given that the the results are not significant in the true direction, for a two sample z test (values in parentheses are probabilities that the interval is reported).

	Effect size				
Sample size	.1	.2	.3	.4	.5
5	.94(.96)	.92(.95)	.91(.93)	.88(.91)	.85 (.88)
10	.93(.96)	.91(.93)	.88(.90)	.83(.86)	.78(.80)
20	.92(.95)	.88(.91)	.82(.84)	.73(.76)	.62(.65)
30	.92(.94)	.86(.88)	.76(.79)	.63(.66)	.48(.51)
40	.91(.93)	.83(.86)	.71(.73)	.54(.57)	.37(.39)
50	.90(.93)	.81(.83)	.65(.68)	.46(.48)	.27(.29)

(b) Suppose results are reported if either nonsigificant or significant in the desired direction, i.e. suppressed only when the results are significant in the less-favored direction, as might be the case if some studies suggested undesirable side effects of a medication. If the favored direction happens to be the true one, the confidence interval coverage will be equal to the nominal coverage, .95 in the example, regardless of the true effect size. Table 3 gives the probabilities that the confidence intervals cover the true values, variance known, when the true values are in the less-favored direction: This is the probability that the null hypothesis is accepted and contains the true values. The probability that the interval is reported is given in parentheses.

Coverage probabilities and effect size estimation

Given the type of conditioning, conditional confidence interval coverage depends on the noncentrality parameter, which is a function of the sample size (known) and the effect size (unknown). Thus, if the effect size were known, the conditional coverage probability would be known, and vice versa. It would appear, then, that estimating the effect size would be helpful in estimating the confidence interval coverage. The relation between effect size estimation and confidence interval coverage, however, is complex.

If the variance were known, estimation of effect size would be equivalent to estimation of the mean difference. With unknown variance, however, estimation of the effect size, which requires an estimate of the unknown standard deviation in the denominator, is considerably more difficult and less robust than estimation of the mean difference. In either case, estimation of effect size is unlikely to be helpful in estimating confidence coverage of the true mean difference. Although the confidence interval coverage when the variance is estimated with small degrees of freedom is not drastically different from the coverage with known variance, estimation of the effect size is very much poorer in the former case.

Note that problems with effect size estimation exist even if there is equal information on and attention to any outcome, while in that case confidence interval coverage is equal to the nominal level, given the assumptions of the model. Calculation of the confidence interval is straightforward, while there are a number of different estimates of effect size even in this simplest case (Hedges & Olkin, 1985).

Hedges (1984) studied the theoretical properties of effect size estimation when only significant effect sizes are observed; see also Hedges and Olkin (1985). The standard estimate

$$g = (\bar{X}_E - \bar{X}_C)/s \text{ as}$$

an estimate of

$$\partial = (\mu_E - \mu_C)/\sigma ,$$

where E is the experimental group mean and C is the control group mean, is biased towards more extreme absolute values even with no censoring, and is also biased when such censoring occurs. The exception is for $\partial = 0$, in which case neither is biased. Note that in this case, with censoring, the confidence interval coverage is zero. The variance of g when $\partial = 0$ is much larger under censoring than without censoring, and is bimodal, so highly nonnormal, for small sample sizes and/or effect sizes. Thus, under the conditions for which confidence interval coverage is far from optimal coverage, estimation of effect sizes is no help in trying to estimate the non-coverage probability.

Even under known censoring conditions, effect size estimation for single studies is of little value when the noncentrality parameter is small. The value of effect size estimation comes through meta-analysis, when a series of estimates of the same effect size are available. One of the problems, even in that case, is that there is almost certainly some censoring, but the type and extent are usually unknown.

Conclusion

It is often claimed that confidence interval or confidence set procedures give more useful information than hypothesis-testing procedures, since they indicate not only whether the hypothesis of a specified value for a parameter would be accepted or rejected, but indicate plausible values of that parameter. This article points out some difficulties in interpreting confidence intervals when there is a parameter value of special interest, as is true when hypotheses are tested. Confidence intervals conditional on covering or especially on not covering that particular value may have coverage probabilities considerably different from their nominal probabilities, under conditions frequently encountered in research in the social and behavioral sciences.

Although the conditional considerations are obviously important when confidence intervals are computed or known only for selected cases, they are also important when confidence intervals are calculated in all cases, if subsets of intervals are examined for different purposes. As noted, intervals that do not include zero are often examined to see whether the range of plausible values is of practical importance, while intervals that do include zero may be examined to see whether studies with greater power would be worth carrying out. Under the conditions reported in this paper, the true coverage probability of each of these subsets of intervals may be very different from their nominal coverage probabilities.

This article has dealt only with inference concerning a single test or interval. Additional problems arise in multiple testing or estimation. A recent paper on this subject (Benjamini & Yekutieli, 2004, with discussion) notes that the conditional problem is insoluble when there are no plausible assumptions about the possible effect sizes; thus, the conditional coverage properties noted in this paper are relevant for the ranges of sample sizes and effect sizes covered. The authors suggest an alternative interpretation involving unconditional aspects that allows some bounds on the probability of non-covering intervals.

References

Benjamini, Y., & Yekutieli, D. (2004, in press). False discovery rate adjusted multiple confidence intervals for selected parameters (with discussion). *Journal of the American Statistical Association*.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition. NY: Academic Press.

Fukkuk, R. G. & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, *68*, 450-469.

Grissmer, D. (1999). Class size effects: Assessing the evidence. its policy implications, and future research agenda. *Educational Evaluation and Policy Analysis*, *21*, 231-248.

Hedges, L. V., (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61-85.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. NY: Academic Press.

Olshen, R. A. (1973). The conditional level of the F test. *Journal of the American Statistical Association*, *68*, 692-698.

Tanner, L. (2004, September 9). Medical editors stress better access to data. *San Francisco Chronicle*, C4.