DOCUMENT RESUME

ED 449 214 TM 032 359

AUTHOR	Pizzitola, Kelly M.
TITLE	A Primer on Confidence Intervals.
PUB DATE	2001-02-01
NOTE	18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Research Methodology; *Statistical Significance
IDENTIFIERS	*Confidence Intervals (Statistics)

ABSTRACT

The recent American Psychological Association Task Force on Statistical Inference report suggested that confidence intervals should always be reported. Other researchers also agree that confidence intervals should be used as a supplement to statistical significance testing. The confidence interval is defined as a range of values constructed around a point estimate that makes it possible to state the probability that the interval contains the population parameter between its upper and lower confidence limits. This paper examines specific flaws in statistical significance testing, as reported by these researchers, and suggests the use of confidence intervals based on the recommendations of the Task Force. The construction of confidence intervals using range null hypothesis, as opposed to null hypothesis, is also discussed. (Contains 26 references.) (SLD)

Running head: A PRIMER ON CONFIDENCE INTERVALS

A Primer on Confidence Intervals

Kelly M. Pizzitola

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1, 2001.

Abstract

The recent APA Task Force on Statistical Inference report suggested that confidence intervals should always be reported. Other researchers also agree that confidence intervals should be used as a supplement to statistical significance testing. The present paper will examine specific flaws in statistical significance testing, as reported by these researchers, and will suggest the utilization of confidence intervals based on the recommendations of the Task Force. The construction of confidence intervals using range null hypothesis, as opposed to point null hypothesis, is also discussed in this paper.

A Primer on Confidence Intervals

For many years researchers have debated the credibility and usefulness of null hypothesis statistical testing (NHST). In fact, the American Psychological Association (APA) created a committee called the Task Force on Statistical Inference (TFSI) to, as one of it's primary goals, investigate some of the controversial issues surrounding statistical significance testing and possible alternatives (Wilkinson & APA Task Force on Statistical Inference, 1999). Although TFSI did not recommend a ban on the use of significance testing, as some had hoped, the Task Force did recommend ways to supplement NHST, including the use of effect sizes and confidence intervals. One of the proposed guidelines is to "always present effect sizes for primary outcomes...." because "it enables readers to evaluate the stability of results across samples, designs, and analyses" (Wilkinson et al., 1999, p. 599). Also, among other recommendations, the Task Force suggested that:

> [Confidence] [i]nterval estimates should be given for any effect sizes involving principal outcomes.... Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies (Schmidt, 1996). Collecting intervals across studies also helps in constructing plausible regions for population parameters. (p.599)

The Task Force added that the use of many good computer programs makes these supplements easily manageable.

Many researchers agree that there has been an overreliance on null hypothesis significance testing and that supplementary procedures, such as the construction of confidence intervals, could be profitable (Cohen, 1994; Falk, 1998; Howard, Maxwell & Fleming, 2000; Loftus & Masson, 1994; Thompson, 1998). Falk (1998), in response to Hagan's 1997 article "In praise of the null hypothesis statistical test", stated that "science has also done well without using NHST. The trouble with NHST is that it assumes the appearance of inferential validity and it may easily lead us astray". I will examine specific flaws in statistical significance testing, as reported by these researchers, and will suggest the utilization of confidence intervals based on the recommendations of the Task Force and other researchers.

Problems with Statistical Tests

As stated above, there have been numerous critiques written on the overreliance of null hypothesis significance testing in psychology. Cohen (1994) adamantly stated:

> What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that $\underline{H}_o$ is true?" But as most of us know, what it tells us is "Given that $\underline{H}_o$ is true, what is the probability of these (or more extreme) data? These are not the same…" (p. 997)

Many, including Cohen, argue that one problem with NHST is that the null hypothesis is often mistaken for what he calls the "nill hypothesis", defined as effect size equaling zero. Testing "nil" nulls of zero difference allows the $\underline{H}_o$ to be unrightfully rejected in almost all cases (Cohen, 1994). Cohen (1990) added, "so if the null hypothesis is always false, what's the big deal about rejecting it?" (p. 1308). In the case of a non-nill hypothesis, the $\underline{H}_o$ is also never literally true when dealing with a sufficiently large sample size (Howard et al., 2000; Thompson, 1996). Thompson (1992) portrayed this point in his statement:

> Statistical significance testing can involve a tautological logić in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. (p. 436)

Another problem with statistical significance testing involves the misinterpretation of $\underline{p} < .05$ and the importance of research results. Researchers often interpret statistical significance as if it meant "plain-English" significance (Cohen, 1994). It is assumed that the lower the $\underline{p}$ value the more important or significant the results become. Cohen (1995) explained:

> Incidentally, I do not question the validity of NHST, but rather its widespread misinterpretation. If I reject the null hypothesis at the 5% level, then I can correctly assert that if it were true, I would have obtained results like those in hand less than 5% of the time. I cannot correctly assert

> that the probability that the null hypothesis is true is less than 5%. (p. 1103)

Cohen (1994) also stated, "Even a correct interpretation of p values does not achieve very much, and has not for a long time" (p. 1001). Other researchers, such as Svyantek and Ekeberg (1995), also commented that the misrepresentation of p values has become a "major stumbling block" in their research when interpreting results for individuals being evaluated in a field intervention.

Other studies have investigated how researchers and graduate students interpret various levels of significance and the importance placed on specific p values (Nelson, Rosenthal & Rosnow, 1986; Rosenthal & Gaito, 1963). For example, Rosenthal and Gaito (1963) conducted a study with 9 graduate students and 10 faculty members at the University of North Dakota using a questionnaire requesting the participants to rate their degree of confidence in a variety of p levels. These researchers found that greater confidence was placed in lower p levels and in p levels based on a larger *n*. Also, graduate students tended to show greater confidence in any given p level than did faculty members. In addition, Rosenthal and Gaito detected "cliff characteristics", sudden drops in confidence in p levels, just beyond the .05 level.

Nelson et al. (1986) conducted a follow-up study where similar results were found when various p levels were rated for confidence. The sampled researchers tended to place more confidence in lower p levels and in samples with a larger *n*. Similar "cliff characteristics" were again detected with .05 and .10

levels. These researchers included effect size as a variable and found that respondents placed more confidence with the increasing magnitude of effect size (Nelson et al., 1986). These two studies reflect the overreliance that researchers have on $\underline{p}$ levels less than .05.

Constructing Confidence Intervals

The definition of a confidence interval as defined by Bohrnstedt and Knoke (1982) is "a range of values constructed around a point estimate which makes it possible to state the probability that the interval contains the population parameter between its upper and lower confidence limits" (p. 144). With boundaries defined by two standard errors above and below the mean, it is expected that approximately 95% of the intervals constructed in repeated sampling of the same size will contain the population mean (Bohrnstedt & Knoke, 1982). When constructing confidence intervals for a particular population parameter, the use of a point null hypothesis is the popular practice (Serlin, 1993). The problem with point null hypothesis is that it is always false. Serlin proposed that a more appropriate null hypothesis to test would be a range null hypothesis, allowing the magnitude of the population effect to be determined.

Serlin's range null hypothesis proposal entailed "performing, in principle, an infinite number of tests, given the observed data, of the hypotheses $\underline{H}_o$: $\mu = \mu_o$, where μ_o is varied over all possible values" (p. 354) to construct a central confidence interval. He also suggested that for the behavioral science experiment, an infinite number of tests of the hypotheses $\underline{H}_o$: $|\mu - \mu_o| \leq \Delta_o$, where Δ_o is varied over all possible values should be used. Those values of μ_o and Δ_o, respectively, that lead to non-rejection constitute the confidence interval. The confidence interval derived in the second fashion "has the form $|\mu - \mu_o| > \Delta_L$, where Δ_L is the lower limit to the confidence interval, and hence it allows the direct determination of whether the true parameter satisfies the good-enough requirements" (p. 355).

The "good-enough requirements" referred to in the previous paragraph is a concept developed by Serlin and Lapsley (1985). These qualifying factors are standards that the researcher sets in advance to indicate what kinds of outcomes are "good enough." For example, consider a case with the point-null hypothesis demonstrated by Serlin and Lapsley. The null hypothesis states that a particular variable, σ, will equal 0, but it is known that "because no theory is absolutely true, the value of σ can never be exactly equal to the theoretical value, 0" (p. 79). In this case a good-enough belt of widthΔ must also be used, so that $0 \pm \Delta$ is predicted. A statistical test is then performed on the experiment to determine if σ is in the range of $0 \pm \Delta$. If the value lies in the good-enough belt then the null hypothesis cannot be rejected (Serlin & Lapsley, 1985).

Reasons for Using Confidence Intervals

Confidence intervals can profitably supplement common hypothesis-testing procedures in many ways and also provide a graphical representation of the results (Fidler & Thompson, in press). Thompson (1999) stated that there is a current bias in the literature towards the publication of Type I errors and the non-publication of replication studies detecting previous Type I errors as failures to replicate. He suggested that the priority for reporting Type I errors has put to the side the statistically non-significant results, therefore not allowing science to naturally "self-correct" itself through replication. The use of confidence intervals may aid in reducing these publication biases and help to achieve consistency of findings across studies when studies are interpreted in relation to each other (Thompson, 1999).

Schmidt (1996) also agreed that point estimates complemented by confidence intervals provide a better picture of data than NHST alone. He stated that one reason that confidence intervals are needed is because they "hold the overall error rate to the desired level" (p. 121). Schmidt also pointed out 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).

Loftus and Masson (1994) suggested two simple reasons to support confidence intervals. First, confidence intervals ask the question, "What are the population means?" instead of, "Is it not true that some set of population means are all equal to one another?" (p. 478). Second, if it is unlikely that the null hypothesis is true in any experiment, then it "makes little sense to test the validity of such a null hypothesis" (p. 478). Loftus and Masson also write that constructing a set of sample means along with their respective confidence intervals provides "the best estimate of the underlying pattern of population means, and the degree to which the observed pattern of sample means should be taken seriously as a reflection of the underlying pattern of population means..." (p. 478).

The push for reform of NHST is not backed by the behavioral sciences alone. Medical and industrial researchers are also searching for techniques to supplement NHST (Borenstein, 1994; Morris & Lobsenz, 2000). Most medical and psychiatric research is conducted "to estimate the clinical import of a treatment or risk factor, and not merely to determine whether or not the effect is nil" (Borenstein, 1994, p. 236). Borenstein reported that the magnitude of the effect is needed to make an informed clinical decision; therefore the study should be framed as an estimate of the treatment effect, not as a test of the null. Confidence intervals allow the researcher to report not only the magnitude of the effect, but the precision with which the effect is estimated (Borenstein, 1994).

Another study conducted by Morris and Lobsenz (2000) investigated adverse impact statistics and their role in employment discrimination cases. Adverse impact occurs when minority-group members are selected at a substantially lower rate than majority-group members. Currently, two statistical methods are being used to assess adverse impact: The four-fifths rule and the Z-test for differences in selection rates. The current problem with using these two methods is the difficulty in integrating the results because the use of effect sizes differs: The four-fifths rule is based on the ratio of selection rates and the Z-test uses the difference between selection rates (Morris & Lobsenz, 2000).

The Z-test is a statistical significance test that evaluates whether the selection rates in the population are equal (Morris & Lobsenz, 2000). As noted above, there is controversy over the appropriateness of statistical significance tests in the behavioral sciences and Morris and Lobsenz claimed that the misunderstanding of these tests is even greater in the courtroom. They give this example:

> When applied in a discrimination suit, the plaintiff would have to show that the data are clearly inconsistent with the null hypothesis of no adverse impact, whereas the defendant need only show that the data are not clearly inconsistent with the null hypothesis. (p. 91)

Thus the Z-test places greater burden on the plaintiff than on the defendant and this burden is stronger when the test has low power (Morris & Lobsenz, 2000).

In legal cases such as these, the use of effect sizes and confidence intervals may provide a better method of presenting information regarding adverse impact (Morris & Lobsenz, 2000). Morris and Lobsenz agreed that "the effect size provides the best estimate of the parameter of interests... while the confidence interval communicates the precision of the estimate; that is, the degree to which the estimate is expected to vary across samples" (p. 99). The precision of the results, or the width of the confidence interval, can practically be used in a courtroom to determine the necessary weight that should be given to statistical and non-statistical evidence (Morris & Lobsenz, 2000). Adverse impact research also would benefit from "building multiple confidence intervals reflecting different degrees of precision.... This would allow decision makers to evaluate the likelihood of different levels of adverse impact and the degree of confidence which should be placed in the statistical results" (p. 104).

Intervals versus NHST

Some researchers argue that confidence intervals summon the same logic as statistical tests and are an extension of the null hypothesis testing procedure (Frick, 1995; Hagan, 1997). Thompson (1998) agreed that confidence intervals may invoke the same logic but suggests that the utility of the interval depends greatly on interpretation. As noted elsewhere:

> If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing. But if the confidence intervals in a study are interpreted in the context of the intervals in all related previous studies, the true population parameters will eventually be estimated across studies, even if our prior expectations regarding the parameters are wildly wrong. (Thompson, 1998, p. 799)

Many researchers also have been slow to adopt techniques such as confidence interval construction because their understanding of statistics is limited (Howard et al., 2000; Svyantek & Ekeberg, 1995; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). In a study by Zuckerman et al. (1993), 551 active psychological researchers took a survey covering basic issues in statistical analysis, including the distinction between Type I and Type II errors, interpretation of interaction, omnibus versus focused tests and the role of power and effect sizes as criteria for successful replications. Although the study covered a small number of statistical issues and had only five questions, the participants did not perform well. The level of accuracy was .59 compared to a .46 baseline (Zuckerman et al., 1993).

Summary

Recently, a compelling argument has been suggested for the use of confidence intervals in analysis. The increasing number of journals requiring the reporting of effect sizes and confidence intervals, the APA Task Force, and many other researchers have all aided in the fight to stop the sole use of null hypothesis significance testing. As described above, supplements such as effect sizes and confidence intervals provide a better look at the data.

The effect size estimates the parameter of interest while the confidence interval provides the magnitude of the effect and a more precise view of the parameter (Borenstein, 1994; Morris & Lobsenz, 2000). Range null hypothesis and good-enough belts (Serlin & Lapsley, 1985; Serlin, 1993) provide an excellent way to utilize confidence intervals. It is like Cohen (1994) described:

> 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. Even null hypothesis testing complete with power analysis can be useful if we abandon the rejection of point nil hypotheses and use instead "good-enough" range null hypotheses. (p. 1002)

Particularly promising is the construction of confidence intervals about effect size estimates (Thompson, in press). First, the APA Task Force strongly emphasized that effect size reporting is essential (Wilkinson et al., 1999). Second, the Task Force strongly recommended confidence interval reporting and interpretation, because (a) unlike NHST the use of intervals does not require that any particular population parameters be presumed, and (b) using intervals across studies will isolate the true population parameters even when our original expectations are wildly wrong. The logical combination of these recommendations is to report confidence intervals about effect sizes.

For some effect sizes computing intervals about effect sizes is quite difficult. Fortunately, many of these difficulties have now been resolved with the software and the methods provided by Cumming and Finch (in press) and Smithson (in press).

References

Bohrnstedt, G. W., & Knoke, D. (1982). Statistics for social data analysis. Illinois: Peacock.

Borenstein, M. (1994). A note on the use of confidence intervals in psychiatric research. Psychopharmacology Bulletin, 30(2), 235-238.

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. (1995). The earth is round ($p < .05$): Rejoinder. American Psychologist, 50, 1103.

Cumming, G., & Finch, S. (in press). A primer on the calculation and interpretation of both central and noncentral confidence intervals. Educational and Psychological Measurement, 61.

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

Fidler, F., & Thompson, B. (in press). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. Educational and Psyhological Measurement, 61.

Frick, R. W. (1995). A problem with confidence intervals. American Psychologist, 50, 1102-1103.

Hagan, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and bayesian analysis. Psychological Methods, 5, 315-332.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin & Review, 1(4), 476-490.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. Personnel Psychology, 53, 89-112.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. The Journal of Psychology, 55, 33-38.

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.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40, 73-83.

Smithson, M. (in press). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. Educational and Psychological Measurement, 61.

Svyantek, D. J., & Ekeberg, S. E. (1995). The earth is round (so we can probably get there from here). American Psychologist, 50, 1101.

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). If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory & Psychology, 9(2), 165-181.

Thompson, B. (in press). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development.

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 588-611.

Zuckerman, M, Hodgins, H. S., Zuckerman, A, & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4(1), 49-53.

TM032359

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A PRIMER ON CONFIDENCE INTERVALS

Author(s): KELLY M. PIZZITOLA

Corporate Source:

Publication Date: 2/1/01

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ Sample sticker to be affixed to document

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KELLY M. PIZZITOLA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Kelly Pizzitola

Position: RES ASSOCIATE

Printed Name: KELLY M. PIZZITOLA

Organization: TEXAS A&M UNIVERSITY

Address:
TAMU DEPT EDUC PSYC
COLLEGE STATION, TX 77843-4225

Telephone Number: 979/845-1335

Date: 1/24/01

OVER