DOCUMENT RESUME

ED 325 524 TM 015 782

AUTHOR Palomares, Ronald S.
TITLE Alternatives to Statistical Significance Testing.
PUB DATE 8 Nov 90
NOTE 20p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (19th, New Orleans, LA, November 14-16, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; Estimation (Mathematics); *Evaluation Methods; *Research Design; Research Problems; *Sample Size; *Statistical Significance
IDENTIFIERS Bootstrap Methods; Cross Validation

ABSTRACT

Researchers increasingly recognize that significance tests are limited in their ability to inform scientific practice. Common errors in interpreting significance tests and three strategies for augmenting the interpretation of significance test results are illustrated. The first strategy for augmenting the interpretation of significance tests involves evaluating significance test results in a sample size context. A second strategy involves interpretation of effect size estimates; several estimates and corrections are discussed. A third strategy emphasizes interpretation based on estimated likelihood that results will replicate. The bootstrap method of B. Efron and others and cross-validation strategies are illustrated. A 28-item list of references and four data tables are included. (Author/SLD)

ALTERNATIVES TO STATISTICAL SIGNIFICANCE TESTING

Ronald S. Palomares

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, November, 1990.

ABSTRACT

Researchers increasingly recognize that significance tests are limited in their ability to inform scientific practice. Common errors in interpreting significance tests, and three strategies for augmenting the interpretation of significance test results, are illustrated. The first strategy for augmenting the interpreCation of significance tests involves evaluating significance test results in a sample size context. A second strategy involves interpretation of effect size estimates; several estimates and corrections are discussed. A third strategy emphasizes interpretation based on estimated likelihood that results will replicate. The methods of Efron, and cross-validation strategies are illustrated.

Statistically significant results are obtained, but thoughtful interpretation of the results suggests that observed effects are not noteworthy. This situation occurs with increasing frequency as researchers become more aware that significance tests are limited in their potential to inform valid interpretations in scientific inquiry (Carver, 1978). Selecting the best indices to use when evaluating empirical results has become a subject of much debate (Huberty, 1987; Thompson, 1989a, 1989b; Rosnow & Rosenthal, 1989; and Welge-Crow, LeCluyse & Thompson, 1990).

Shaver (1979, pp. 5-6) has argued that

> The emphasis on statistics and the "test of significance" procedure has resulted in a methodological orientation toward establishing generalizability that has been deleterious in it's effects on the scientific accumulation of knowledge in education.

Similarly, Carver (1978, p. 378) states that, "Statistical significance testing has involved more fantasy than fact."

The traditional emphasis on significance testing and the direct interpretation of inferential test results has led to the need for improved editorial policies and scholarly practices (Thompson, 1987). It has been increasingly recognized that significance testing only directly aids the interpretation of the results in special cases, e.g., significant results from small sample sets, and not in all cases (Thompson, 1987). Thus, significance testing is only useful when employed by thoughtful researchers, and especially by researchers who are aware of

strategies that make the interpretation of the test results most meaningful. The purpose of the present paper is to discuss three strategies useful in augmenting the interpretation of significance test results. First, however, a review of the influence of sample size on the outcomes of significance testing warrants some consideration.

The Influence of Sample Size on Significance Test Results

The size of the sample used in a significance test has direct bearing on the results of the test. Selecting an appropriate sample size is one of the most difficult problems found in designing research (McNamara, 1990a). When working with large samples, virtually all null hypotheses will be rejected, since the "null hypothesis of no difference is almost never exactly true in the population" (Thompson, 1987, p. 14). Stated a different way, Hays (1981, p. 293) argues that "virtually any study can be made to show significant results if one uses enough subjects." Welge-Crow, LeCluyse and Thompson (1990) demonstrate Hays' argument using a concrete heuristic example that will serve here to emphasize the influence that sample size has on significance test results.

Presume that a researcher decided to compare the mean IQ scores of 12,000 students located in one zip code of the Houston Independent School District with the mean IQ of the remaining 188,000 students residing in other zip codes. If the mean IQ of the 12,000 students in the zip code of interest is only 100.15 (SD=15), and the mean of the remaining 188,000 students is 99.85 (SD=15), the two means differ to a statistically significant

(Zcalc = 2.12 > Zcrit = 1.96, p<.05) degree. The less thoughtful interpretation of these statistically significant findings would be that the 12,000 students differ appreciably in their intellects from their 188,000 peers. Perhaps it will even be suggested that a school for the gifted be built in this zip code. Yet, these differences can hardly be considered noteworthy.

A thoughtful researcher would note, in such a situation, that the standardized difference in these two mean (.3/15 = 0.02) is trivially small. The researcher would also be aware that all measurements are limited and imperfect, and that the difference of the means (0.3 being one-third of one IQ point) is substantially less than the standard error of measurement (SEM) of an IQ measure with a reliability coefficient of 0.92 (SEM = 15 (1-.92)^.5 = 15 (.08)^.5 = 15 (.2828) = 4.243). The researcher who applies the significance test and only interprets the inferential results (as against the researcher that truly understands the interpretation and realizes that there is not a real difference between the samples) is less likely to offer helpful recommendations for policy or worthy contributions to theory elaboration.

In choosing sample size, Hinkle, Wiersma, and Jurs (1988, p. 293) state three relevant assumptions:

1. It is sensible to view research findings based on large samples as more reliable than findings based upon smaller samples, all other things being equal.
2. However, inferential statistical methods will not result in rejecting the null hypothesis if,

in the design of the study, an inappropriately small sample was selected.

3. In a well-planned research study, in which the variance of the criterion variable is likely to be quite large and the treatment effects rather small, large samples are appropriate and justifiable.

Thus, sample size is of the utmost importance for the interpretation of significance testing results. Morrison and Henkel (1970) and Carver (1978) explain the limits of significance testing as an aid to scientific practice. Only once a sample is determined to be large enough to detect certain effect sizes can interpretation of significance test results be made more directly. In short, sample size must be considered when making interpretations of the empirical results. Otherwise, testing significance becomes only a test of whether a large sample is in hand, which the researcher presumably already knows, before the significance test is even conducted.

1. Evaluating Significance Test Results in a Sample Size Context

The first strategy for augmenting interpretation of significance tests is the strategy of evaluating the expected or the obtained effect size in relation to changes in sample size (Thompson, 1989a). Prospectively, before data are collected, the researcher must determine what sample size is needed to obtain a statistically significant result for a given effect size. Retrospectively, once data have been collected, the researcher must determine how variations in sample size might have altered

the significance decision, assuming that the effect size is generalizable and thus is taken as fixed.

Welge-Crow, LeCluyse and Thompson (1990) illustrate this approach with the following application. Say a researcher detected a large effect size of 33.6%. Table 1 presents significance tests associated with this effect size taken as a fixed value, but assuming different sample sizes had been used. The table can be viewed as presenting results for either a multiple regression analysis involving two predictor variables (in which case the "r sq" effect size would be called the squared multiple correlation coefficient, R^2) or an analysis of variance involving an omnibus test of differences in three means in a one-way design (in which case the "r sq" effect size would be called the correlation ratio or eta^2).

INSERT TABLE 1 ABOUT HERE.

The table presents results for fixed effect sizes but increasing sample sizes (n=4, 13, 23, or 33). For the effect size (33.6%) reported in the table, the result become statistically significant when there are somewhere between 13 and 23 subjects in the analysis (Welge-Crow, LeCluyse & Thompson, 1990, p. 5).

Craig, Eison and Metze (1976) have found serious distortions in interpretations of research studies when researchers failed to understand the effect that sample size has upon significance tests. When there is a failure to understand how the sample size affects results, researchers may ignore large effect sizes involving nonsignificant results attributable to small sample

sizes, while at the same time over-interpreting significant results when the effect size is actually small (Welge-Crow, LeCluyse & Thompson, 1990).

2. Interpretation of Effect Size Estimates

Effect size can be characterized as the "degree to which the phenomenon exists" (Cohen, 1977, p. 9). There are numerous methods that a researcher may choose with which to estimate the effect size for the data (e.g., Hays, 1981; Tatsuoka, 1973). The effect size is used by the researcher to "garner some insight regarding result importance" (Welge-Crow, LeCluyse & Thompson, 1990).

McNamara (1990b) demonstrates the usefulness of the effect size in allowing the researcher to infer if a meaningful true difference occurs within the targeted population. In the example that McNamara (1990b) uses, he compares the difference between two means from a survey administered to a sample of teachers and a sample of administrators. The mean difference for a particular item was 0.34, which when divided by the common standard deviation of 0.66, yielded an effect size of 0.52. With an effect size of 0.52, the researcher can then conclude that "on average the questionnaire item score for administrators was a 0.52 standard deviation higher than the same questionnaire item score for teachers" (McNamara, 1990b, p. 29). Thus, the effect size is over the one-half common standard deviation effect size that Borg (1987) argues represents a meaningful difference between two means.

However, Welge-Crow, LeCluyse, and Thompson (1990) explain

that sample effect sizes "overestimate" the effect size actually found in the full population, as well as the effect size that is likely to be found in future studies with different samples. This inflation occurs because all classical parametric (e.g., t-tests, ANOVA) methods are correlational methods (Knapp, 1978; Thompson, 1988) that capitalize upon the sampling error as a part of the least squares analysis (Welge-Crow, LeCluyse & Thompson, 1990). However, there are correction formulas available (Maxwell, Camp & Arvey, 1981; Rosnow & Rosenthal, 1988) that can be applied to correct the estimated population effect sizes based on sample results, or in the estimation of effect sizes likely to be found in future samples (Welge-Crow, LeCluyse & Thompson, 1990). These corrections tend to be larger when the sample sizes are small or if the original effect size is small (Thompson, 1990).

3. Evaluation of Result Replicability

Replication of the results of a study, one the eight elements of the scientific method (Babbie, 1990), is a third strategy that can be used to facilitate accurate interpretation of results. Increasing the estimated likelihood that the results will replicate is one of the goals of research and plays a crucial part in scholarly inquiry. As Welge-Crow, LeCluyse and Thompson (1990) state, contrary to many misconceptions, "significance tests do not evaluate the probability that results will replicate" (p. 7).

One of the easiest and most commonly used methods with which to scrutinize the replicability of results is that of

partitioning the sample and replicating the study on the other portion. Then comparison of results will either support or will bring into question the replicability of the study's results and conclusions. Various sample partioning methods have been devised, including conventional cross-validation strategies and the "bootstrap" methods developed by Efron and his colleagues (Diaconis & Efron, 1983; Efron, 1979).

One of the cross validation strategies that is the easiest to apply is a three step process. Initially the researcher randomly splits the sample into two separate subgroups. Next, the researcher conducts the same analyses separately on both subgroups. Finally, the researcher empirically compares the results, attempting to demonstrate the replication of results for both subgroups, thus increasing confidence in the replicability of the study's results.

Welge-Crow, LeCluyse and Thompson (1990) describe the interpretation of results that the researcher should expect from this procedure. The invariance coefficients obtained for the two samples should approach 1.0 for the results to be indicative of replicability across samples. If the results do indicate replicability, "... the researcher can interpret the set of results involving all the subjects with more confidence" (p. 8). Welge-Crow, LeCluyse and Thompson (1990) explain that the interpretation of results should always be based upon the total sample, not the subgroup splits, because the "full sample should theoretically provide the most generalizable results; sample splitting is only performed to evaluate the replicability of the results" (p. 8). It is also stressed that the results of any

replicability study must use empirical methods to evaluate replicability, not subjective comparison of solutions (Welge-LeCluyse & Thompson, 1990). Results that appear on the surface to be different (e.g., yield markedly different beta weights for variables) may be remarkably similar in the population effects that are estimated.

The "bootstrap" method devised by Efron and his colleagues (Diaconis & Efron, 1983; Efron, 1979) is considered to be one of the most powerful strategies for evaluating the replicability of results. The process underlying the "bootstrap" method involvess copying the original data set numerous times on top of itself and thus creating a "mega" data set. From this "mega" data set, hundreds or even thousands of samples are randomly selected and undergo the specific analyses required by the particular study. These results are all computed separately for each sample. Once all the samples have been analyzed, they are then averaged together. The power from this type of method is realized through the analytic consideration of "so many configurations of subjects and informs the researcher regarding the extent to which results generalize across different configurations of subjects" (Welge-Crow, LeCluyse & Thompson, 1990, p. 9).

Thompson and Melancon (1990) provide examples of "bootstrap" methods, and further explanation of the methods. Welge-Crow, LeCluyse and Thompson (1990) include in their paper also an example of "bootstrap" estimation. Table 2 presents the small data set used in this example. Table 3 presents the descriptive statistics for the data in hand (n=12) for this example--these

are the results conventionally calculated by researchers. Table 4 presents the "bootstrap" estimates of the population correlation coefficients based on the data original data. Lunneborg's (1987) software, which automates the "bootstrap" method on a microcomputer, was used to derive the tabled estimates based upon 500 resamplings with replacement from the small data set.

INSERT TABLES 2, 3 AND 4 ABOUT HERE.

For the Table 4 example illustrated by Welge-Crow, LeCluyse and Thompson (1990), it can been seen that the results of the "bootstrap" method indicate that the first correlation coefficient (0.052) is very close to the mean found in 500 bootstrap resamplings (0.0514). Such a result would suggest to the researcher that some confidence can be vested in these results in hand, since the sample result so closely approximates the result over several hundred configurations of the subjects. The relatively large standard deviation (0.3541) for the eighth coefficient, however, suggests that this estimate (r=.008) is least stable over different groups of the subjects.

Summary

Increasingly, researchers note that they obtain statistically significant results, but careful scrutiny of the data demonstrates that such differences are not necessarily meaningful. As Holmes (1990, p. 72) observes:

> The trouble with reporting statistically significant results is twofold. First, all too often the word "statistically" gets lost or left off. Thus, the researcher reports a "significant

> difference was obtained." ...This leads to the second problem. When the word "significant" is used in this way, most people naturally equate it with the words "important," "meaningful," or "practical." Just the phrase, "A significant difference was found..." carries a certain amount of authority.

Three strategies for augmenting the interpretation of significance test results were illustrated. The first strategy deals with the size of the sample and the effects that too large or too small a sample size may have upon significance test results. A second strategy involves using the effect size to determine the degree to which the identified phenomenon is found to exist in the data. Finally, evaluating the replicability of the results, using either cross-validation or the "bootstrap" methods developed by Efron and his colleagues, was discussed.

Researchers spend valuable time and money conducting a research project. Numerous hours spent on the interpretation of the results can all be for nothing if the researcher fails in the fundamental determination of whether there is actually a meaningful effect size found within the data. Conducting a t-test, ANOVA, or other statistical significance test will inform the researcher if there is "statistical significance", but the researcher must delve further into the data to determine if the results are "meaningful". This paper described three such methods that can help the researcher make these determinations.

REFERENCES

Babbie, E.R. (1990). Survey research methods (2nd ed.). Belmont, CA: Wadsworth Publishing.

Borg, W.R., (1987). Applying educational research: A practical guide for teachers (2nd ed.). New York: Longman.

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

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Craig, J.R., Eison, C.L., & Meltze L.P. (1976). Significance tests and their interpretaiton: AN example utilizing published research and omega-squared. Bulletin of the Psychonomic Society, 7, 280-282.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-140.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

Hays W.L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Hinkle, D.E., Wiersma, W. & Jurs, S.G. (1988). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin Company.

Holmes, C.B. (1990). The honest truth about lying with statistics. Springfiled, IL: C.C. Thomas.

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

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Lunneborg, C.E., (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

McNamara, J.F. (1990a). The sample size issue in educational Administration. Research in Educational Administration and Supervision, 10(2), 46-49.

McNamara, J.F. (1990b). Statistical power in educational research. National Forum of Applied Educational Research Journal, 3(2), 23-36.

Maxwell, S.E., Camp, C.J., & Arvey, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

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

Rosnow, R.L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. Journal of Counseling Psychology, 35, 203-208.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Shaver, J.P. (1979). The productivity of educational research and the applied-basic research distinction. Educational Researcher, 8(1), 3-9.

Tatsuoka, M.M. (1973). An examination of the statistical properties of a multivariate measure of strength of relationships. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Thompson, B. (1987, April). The use (and misuse) of statistical

significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1988, April). Canonical correlation analysis: An explanation with comments on correct practice. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50, 15-31.

Thompson, B., & Melancon, J.G. (1990, November). Bootstrap versus statistical effect size corrections: A comparison with data from the Finding Embedded Figures Test. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans.

Welge-Crow, P., LeCluyse, K., & Thompson, B. (1990, June). Looking beyond statistical significance: Result importance and result generalizability. Paper presented at the annual meeting

of the American Psychological Society, Dallas.

Table 1
Statistical Significance at Various Sample Sizes
for a Fixed Effect Size (Large Effect Size)

	SOS	r sq	df	MS	Fcalc	Fcrit	Dec.
SOSexp	337.2	0.336	2	168.600	0.253	200.00	Not Rej
SOSunexp	665.1		1	665.100			
SOStot	1002.3		3	334.100			
SOSexp	337.2	0.336	2	168.600	2.535	4.10	Not Rej
SOSunexp	665.1		10	66.510			
SOStot	1002.3		12	83.525			
SOSexp	337.2	0.336	2	168.600	5.070	3.49	Rej
SOSunexp	665.1		20	33.255			
SOStot	1002.3		22	45.559			
SOSexp	337.2	0.336	2	168.600	7.605	3.32	Rej
SOSunexp	665.1		30	22.170			
SOStot	1002.3		32	31.322			

Note. As sample size increases, tabled "critical F" values get smaller. Additionally, as sample size increases, error df gets larger, mean square error gets smaller, and thus "calculated F" also gets larger. Entries in bold remain fixed for the purposes of these analyses. From Thompson (1989b), with permission.

Table 2
Data Set for Hueristic Example

n	MILESEC	SYSTOLAV	POND	TOTCHOL	HDLCHOL
1	890(+0.18)	94.0(-1.04)	11.5(-0.91)	180(+0.64)	80.1(+1.17)
2	1097(+1.16)	108.7(+1.42	12.0(-0.69)	142(-1.56)	51.1(-1.02)
3	1300(+2.12)	97.7(-0.42)	13.1(-0.21)	165(-0.23)	63.3(-0.10)
4	948(+0.45)	90.3(-1.66)	12.6(-0.43)	199(+1.74)	75.7(+0.84)
5	940(+0.41)	100.7(+0.08)	19.3(+2.49)	187(+1.04)	61.0(-0.27)
6	760(-0.44)	104.3(+0.69)	14.7(+0.48)	148(-1.22)	76.0(+0.86)
7	740(-0.53)	95.3(-0.82)	14.2(+0.26)	164(-0.29)	78.5(+1.05)
8	571(-1.33)	97.7(-0.42)	13.6(+0.00)	174(+0.29)	54.3(-0.78)
9	748(-0.50)	102.7(+0.42)	10.9(-1.17)	190(+1.22)	62.2(-0.18)
10	640(-1.01)	96.0(-0.70)	11.4(-0.95)	161(-0.46)	67.4(+0.21)
11	642(-1.00)	107.0(+1.14)	14.6(+0.44)	159(-0.58)	34.8(-2.25)
12	957(+0.49)	108.0(+1.30)	15.2(+0.70)	159(-0.58)	70.8(+0.47)

Note. From Thompson (1990), with permission.

Table 3
Descriptive Statistics and Correlation Coefficients

	MILESEC	SYSTOLAV	POND	TOTCHOL	HDLCHOL	PREDC1	CRITC1
Mean	852.8	100.2	13.6	169.0	64.6	0.0	0.0
SD	211.0	6.0	2.3	17.3	13.2	1.0	1.0
MILESEC		.052	.046	-.047	.140	.063	.048
SYSTOLAV			.244	-.624	-.559	-.981	-.752
POND				.008	-.121	-.084	-.064
TOTCHOL					.243	.637	.830
HDLCHOL						.569	.742
PREDC1							.767

Note. From Thompson (1990), with permission.

Table 4
Bootstrap Estimates of r's for Table 2 Data
Based on 500 Samples with Replacement

Coef.	Table 3 Estimate	Bootstrap Mean	Bootstrap Median	Bootstrap SD
1	0.052	0.0514	0.0417	0.2819
2	0.046	0.0421	0.0603	0.2287
3	-0.047	-0.0233	-0.0489	0.2690
4	0.140	0.1135	0.1551	0.3092
5	0.244	0.2598	0.2428	0.2343
6	-0.624	-0.5878	-0.6196	0.2135
7	-0.559	-0.5430	-0.5737	0.2166
8	0.008	-0.0649	-0.0519	0.3541
9	-0.121	-0.0971	-0.1198	0.2224
10	0.243	0.2189	0.2486	0.2560

Note. From Thompson (1990), with permission.