MISINFORMATION IN MIS RESEARCH: THE PROBLEM OF STATISTICAL POWER

Jack J. Baroudi
and
Wanda J. Orlikowski

June 1986

Center for Research on Information Systems
Information Systems Area
Graduate School of Business Administration
New York University

Working Paper Series

CRIS #125

GBA #86-62

Table of Contents

ABSTRACT

This study reviews 57 MIS articles employing statistical inference testing published in leading MIS journals over the last five years. The statistical power of the articles was evaluated and found on average to fall substantially below the accepted norms. The consequence of low power is that it can lead to misinterpretation of data and results. Collectively these misinterpretations result in a body of MIS research that is built on potentially erroneous conclusions.

1. INTRODUCTION

There is a danger that much of the inference testing performed by management information systems (MIS) researchers is seriously flawed. This is a direct result of the failure of researchers to attend to the power of their statistical tests. The importance of power in statistical inference has been emphasized in the applied and social psychology literatures for many years [Cohen 1962, 1965, 1977; Hays 1981], but this emphasis appears to have been ignored by the MIS literature.

This paper reviews recent publications in the management information systems field and reveals that not only do researchers ignore the determination of power in their statistical testing but also, and more significantly, that they are unaware of the implications of low power on the findings of their tests. This paper attempts to bring these issues to the attention of MIS researchers, and also presents the findings of a survey conducted of recent MIS research to assess its status with regards to statistical power.

2. POWER DEFINED

The power of a statistical test is the probability of rejecting the null hypothesis [Cohen 1977]. Statistical power is important in the case where the null hypothesis is in fact false, that is, when the phenomenon being investigated does exist. In these circumstances if the test reveals nonsignificant results, the usual response is to accept the null hypothesis and to conclude that the effect being examined does not exist. To conclude that an effect "does not exist" unequivocally is never appropriate. Rather, the researcher should qualify his/her conclusions, that the effect had "not been demonstrated" by this study. However a graver danger occurs if in fact the phenomenon does exist, but was not detected due to low statistical power. This would be incorrect, as we would be generating a spuriously negative result, that is, committing a Type II error.

There is a distinct asymmetry in the attention paid to the two types of statistical inference errors (Type I and Type II) in the MIS literature. While the focus on Type I errors is clearly appropriate, we should not ignore Type II errors. Type I error (the probability of mistakenly rejecting a null hypothesis) is carefully guarded against by setting the α to a prudently low level of .05 or .01. However the second type of error, that of mistakenly accepting a false null hypothesis, is often ignored. Yet, it need and indeed should not be. The probability of committing a Type II error (β), can be controlled and planned for. In this way researchers can ensure that their statistical tests have sufficient power to detect an effect, if it exists.

Clearly the relationship between the two risks, Type I and Type II error, needs to be kept at a reasonable level. Cohen [1965, p.98] noting that the consequences of false positive claims are more serious than those of false negative claims, recommends that Type I errors be guarded against four times as stringently as Type II errors. As the convention for α is .05, this would mean setting β to .20. Accepting these conventional values for α and β results in the conventional value for power ($1 - \beta$) to be .80. It is this standard value (Cohen [1965, 1977], Welkowitz et al. [1982]) that we employ as a benchmark against which to judge the acceptability of power levels in statistical tests.

Studies that employ high power levels (.80 or higher) offer advantages in the interpretation of their results. Keppel [1973, p.534] notes that in such cases, "... we do have a relatively sensitive experiment and, consequently, we feel more comfortable in concluding that treatment effects are probably not present in this particular comparison. In short, an experiment with high power provides strong support for our decision not to reject the null hypothesis, while an experiment with low power provides little support for either the null or alternative hypotheses." It should be noted however (although we suspect that this is not very frequent), that sometimes it is not possible, given the constraints of the study, to obtain the desired level of power. Under these circumstances, researchers need to carefully determine whether the costs involved in conducting the research are worth the substantial risk of not demonstrating any effect. Where such studies are undertaken and negative results are reported, they need to be qualified by an indication of the low power of the study.

3. POWER ANALYSIS

Empirical researchers need to attend to both significance testing and power analysis. The following briefly explains the procedures for calculating power.

The power of any statistical test of a null hypothesis is mathematically defined as $1 - \beta$ and refers to the probability of rejecting the null hypothesis. It is a function of the following three parameters:

- The criterion for rejection of the null hypothesis (the α level), and whether the test is directional (one-tailed) or non-directional (two-tailed). As α increases, ceteris paribus, so the power of the test increases, and given the same α level, ceteris paribus, power increases as one moves from a non-directional to a directional test.

- The magnitude of the phenomenon in the population, that is the size of the effect being investigated. This is determined from the specific hypothesis value posited as an alternative to the null hypothesis (in the Neyman-Pearson formulation [Cohen 1965,p.96]). Such an alternative value can be determined by examining prior research on the phenomenon and estimating its size from reasonable expectations, or one can adopt conventional effect size values of large, medium and small effects [Cohen 1977]. The larger the effect size posited, ceteris paribus, the greater the power.

- The sample size, n, such that, ceteris paribus, as n increases so does the power of the statistical test.

These three parameters are related to power in such a way, that when they are all specified, the power of the test is completely determined. The power of a given statistical test can thus be altered by changing any of the values of the three parameters: α, n or effect size.

The effect size is typically given, corresponding to the expected nature of the phenomenon of interest, and cannot be altered. The α level may be increased to improve power, although there is a strong reluctance among researchers to deviate above the accepted norm of .05. This value would thus seem to be non-negotiable in most cases. A directional rather than a non-directional formulation of the test might be used to increase power, but caution is required in the adoption of such an approach. It is only valid where results in the opposite direction of any magnitude are not to be distinguished from null results. Such a position is rare, and it is usually recommended not to adopt directional tests except in very narrow circumstances [Cohen 1965d; Hays 1981]. This leaves the sample size as the easiest value to manipulate. Researchers, however, are often constrained by pragmatics, settling for the size they can get, or what is traditional in a research area, with little concern for the impact of this choice on the power of their test. The appropriate procedure should be to determine the effect size, set α and the desired power level, and only then determine what the sample size needs to be.

It is also possible to perform a post hoc evaluation of the power of a given statistical test, once it is completed, based on the type of test, n, α and the assumed effect size. This provides an indication of how much weight to attach to negative findings. It is this post hoc procedure that we employed to evaluate the power of recent MIS research, which we present below.

4. THE RESEARCH STUDY

This study examined the issues of four major journals publishing MIS research over the past five years (January 1980 - July 1985). The journals selected were: Communications of the ACM, Decision Sciences, Management Science and MIS Quarterly. Only empirical studies were of relevance, and in particular those employing inferential statistics. Sixty-three articles matched these criteria, and Table 1 shows their distribution.

----Insert Table 1 here-----

Both parametric and non-parametric tests were included, the power of the latter tests being determined by using analogous parametric tests where appropriate [Hays 1981; Welkowitz et al. 1982]. For example, the t-test for means approximates for the Mann-Whitney U test, the parametric F test for the Kruskal-Wallis H test, and so on. As noted in Cohen [1962, p.149] the effect of such approximations is to slightly overestimate the power of the test. Overall, however, the effect of this positive bias is trivial.

Of the 63 articles selected, five included statistical tests for which appropriate power calculations were not available and hence these were not included in the sample. These studies utilized tests such as Wilk's lambda and ANAVA (sic) direction statistics. One other article provided insufficient information to conduct a power analysis and was also excluded from the study. The remaining 57 articles generated 149 statistical tests for which power analyses were conducted.

It should be noted that while most of the articles involved a number of tests, not all of these tests were equally relevant to the key hypotheses of the research. Only tests of the major hypotheses were included. Table 2 presents the distribution of the types of tests that constituted our final sample.

-----Insert Table 2 here-----

5. RESULTS

For each of the statistical tests in the sample, the power of the test was determined by using the study's given sample size, setting the α level to the accepted standard of non-directional .05, and positing the conventions of small,

JOURNAL	NUMBER	PERCENT
MIS QUARTERLY	27	42.9%
COMM. of the ACM	18	28.5%
MANAGEMENT SCIENCE	9	14.3%
DECISION SCIENCES	9	14.3%
TOTAL	63	100%

Table 1: Distribution of MIS Studies employing statistical inference testing, for the period: January 1980 - July 1985.

Table 2: Distribution of Statistical analyses employed in MIS studies.

Statistic Analysis	Frequency	Percent
ANOVA	46	30.9%
Correlation	22	14.8%
T-test	21	14.1%
Chi-Square	20	13.4%
Non-parametric	18	12.1%
Regression	13	8.7%
Partial Correlation	7	4.6%
Proportion	1	.7%
Difference between Proportions	1	.7%
TOTAL	149	100%

medium and large effect sizes. Cohen [1977] provides values for small, medium and large effect sizes for a number of typical parametric tests which have become accepted norms. These norms vary for each statistical test, for example the small, medium and large effect sizes posited for a t-test of difference between means are .10, .25, and .40 respectively; while the small, medium and large effect sizes suggested for a Chi-square test are .10, .30 and .50 respectively [Cohen 1977]. For each of the 149 tests culled from the articles, power values at three levels were determined by employing Cohen's [1977] power tables. The mean power of the tests at each effect size level was determined. Table 3 shows the distribution of sample sizes for the MIS studies by type of statistical analysis. Caution is needed in interpreting this data. The high standard deviation levels in many of the entries reveal a large amount of variation in sample sizes. For example, among the non-parametric subsample the average sample size was 93, yet this is misleading as further investigation reveals that of the 18 non-parametric tests we examined, 13 had an average sample size of 14, while two had sample sizes of over 300. Table 4 presents the power distributions for the studies at small, medium and large effect sizes.

-----Insert Tables 3 and 4 Here-----

The average power of the 149 statistical tests examined was 19%, 60%, and 83% if one assumes small, medium, and large effect sizes respectively. On average, if one assumes that the phenomenon being investigated exhibited only small effects then the studies had only a one in five chance of detecting the phenomenon. When one assumes a medium effect size the power increases but still the researcher had less than a two-thirds probability of detecting the relationship. It is not until one assumes a large effect size that the tests have a good chance of discovering the relationship. Eighty percent is the recommended conventional level for power [Cohen 1965, 1977; Welkowitz et al. 1982]. This means that if a relationship exists we should have a four in five chance of detecting it. From Table 4 it can be seen that 99 percent of the studies fall below

Table 3: Distribution of Sample Sizes for 52 MIS studies for the period: January 1980 - July 1985.

Type of Statistical Analysis	Sample Size	
	Mean	Std. Dev*
ANOVA	64	79
Correlation	132	203
T-test	45	41
Chi-Square	119	79
Regression	216	163
Non-Parametric	93	165
Partial Correlation	47	14
Proportion	24	
Difference between Proportions	18	
Grand Mean	84	

*Blank entry indicates those tests for which there was only one in the sample and hence no standard deviation is calculable.

Table 4: Frequency and Cumulative Percentage Distribution of the Power of 52 MIS studies at Small, Medium and Large Population Effects under non-directional .05 level conditions.

POWER	SMALL EFFECT		MEDIUM EFFECT		LARGE EFFECT	
	Frequency	Cumulative Percentage	Frequency	Cumulative Percentage	Frequency	Cumulative Percentage
.91 -	-		40	100%	90	100%
.81 - .90	2	100%	11	73%	8	40%
.71 - .80	-		8	66%	11	34%
.61 - .70	2	99%	18	60%	15	27%
.51 - .60	6	97%	12	48%	11	17%
.41 - .50	5	93%	6	40%	3	9%
.31 - .40	2	90%	21	36%	7	7%
.21 - .30	30	89%	20	22%	1	3%
.11 - .20	42	68%	11	9%	2	2%
.01 - .10	60	40%	2	1%	1	1%
TOTAL	149		149		149	
Average Power	19%		60%		83%	

this level when assuming small effects while 66 and 34 percent of the tests fall below this level if one assumes medium or even large effects respectively.

Given that the power values vary so dramatically depending on which effect size one assumes, it is important to clarify what is meant by each. With regard to differences between means, for example, a small effect is defined as a .20 s.d. between population means, a medium effect is a .50 s.d. and a large effect would be .80 s.d. between population means [Cohen 1977]. From a psychological standpoint a medium difference should be large enough to be noticeable. A small effect, on the other hand, is relatively imperceptible and finally a large effect is so apparent as "... to render a statistical test virtually superfluous" [Cohen 1965, pp. 97]. If a researcher is uncertain of the effect size, the standard has been to assume a medium effect (i.e. s.d. of .50).

In the case of MIS research it is only when one assumes that the effect is so large as to make statistical testing unnecessary, that our studies on average reach adequate power levels, and even then 34 percent fall below the .80 standard. This situation is unfortunate as the power of a study is one of the few things that a researcher, given a posited effect size, can control prior to gathering any data. Given the expense and difficulty of actually collecting data one would want better than a fifty-fifty chance of finding the phenomena under investigation.

Table 5 lists the power of the studies by type of statistical test employed. With the exception of regression, none of the tests reach the 80 percent power level when assuming a medium effect size. ANOVA tests account for almost one third of all MIS statistical analyses, yet their average power level (assuming medium effect size) is only 56 percent.

-----Insert Table 5 Here-----

Table 5: Distribution of Power Values for 52 MIS studies assuming small, medium and large effects under non-directional .05 level conditions.

Type of Statistical Analysis	POWER LEVELS ASSUMING:					
	SMALL EFFECT		MEDIUM EFFECTS		LARGE EFFECTS	
	Mean	Std. Dev*	Mean	Std.Dev*	Mean	Std. Dev*
ANOVA	20	19	56	30	79	26
Correlation	19	17	68	28	89	18
T-test	16	10	53	27	79	22
Chi-Square	16	10	67	32	89	17
Regression	29	14	91	12	99	.3
Non-Parametric	16	165	42	28	69	22
Partial Correlation	23	9	62	1.6	93	8
Proportion	10		36		77	
Difference Between Proportions	9		32		67	

*Blank entry indicates those tests for which there was only one in the sample, and hence no standard deviation is calculable.

6. DISCUSSION

The real danger of neglecting to consider the power of a study is that the authors may erroneously conclude that the treatment introduced or the phenomenon examined has no effect or makes no difference. In fact, what they may be finding is not no effect, but no demonstration of an effect, quite possibly due to the inadequate power of their tests. The other problem is that readers cannot determine if no effects were found because no relationship exists or because the study had poor power. The MIS research literature is rife with these problems. It would be too voluminous to present an analysis of all these studies. For illustration, however, we present a selected sampling of some of the typical problems we uncovered.

6.1. Research Studies

Chorba and New [1980] conducted an experimental study of decision maker learning in a competitive environment in order to identify which information system parameters facilitate learning. The authors used a strong experimental design and it was in general, a well planned study, with one major exception, the power of the study was low. For most of their statistical analyses their power was substantially below 50 percent (assuming a medium effect size). While they did uncover a number of statistically significant findings, their results are peppered with numerous non-significant findings such as "... no significant difference in the amount of data used between competitors that identified a correct strategy as opposed to those who did not" [pp.609]. They then attempt to interpret this as indicating that "... experience with a given decision making situation, rather than quality of performance of strategy, seems to be the operational determinant of the amount of data requested" [pp. 609]. While they may be correct they cannot claim this based on their statistics. The test they employed to examine this issue had a power of only 28 percent (assuming a medium effect size). Thus they were unlikely, from the onset of the study, to find any relationship. What was otherwise a well designed and executed study ends up with inconclusive findings, because the authors did not examine their power prior to determining what sample size they should employ.

Remus [1984] examined the impact of graphical versus tabular data presentations on decision making. He concluded that "... in the first 12 periods there was no significant advantage for either type of display" [pp. 538]. He based this statement on the result of a t-test which examined the differences in costs depending on whether the decision maker used a graphic or tabular decision aid. His t-test was not significant and he therefore concluded that no difference was demonstrated. This test, however, only had a power of 43 percent (assuming a medium effect size). While there may be no difference, he cannot claim this based on his analysis; his failure to demonstrate a difference may be a result of the study's low power.

In an attempt to learn the effect which program indentation has on comprehension, Miara et al. [1983] examined two styles of indentation: blocking and non-blocking. In addition they looked at four possible levels of indentation. The subjects included both novice and experienced programmers. While their study reported that experience level had an effect, no significant effects were found with non-blocked versus blocked indentation style or with the interactions. This finding puzzles them and they state in the discussion that "...we are not sure why this result occurred because we expected a significant difference in comprehension with the type of blocking used for control structures. It may be possible that comprehension scores for a longer and more complex program would show a greater difference with the type of blocking used for the control structures" [pp. 867]. A more probable explanation for their failure to find the expected result is that they had little power, ranging from 63 percent for the main effect to only 19 percent for the interactions (assuming medium effect sizes).

Shneiderman [1982] addressed the problem of what types and styles of documentation aids assist programmers in comprehending, debugging and modifying programs. One of his experiments tested program comprehension aids by comparing psuedocode and flowcharts for presenting control flow information and text versus graphics for presenting information on data structures. His results found no effects for the text versus graphic hypothesis or for the interactions. He concludes that for control information "... the form of presentation does not matter" [pp.62]. This is a serious misstatement based on his statistical analyses. The power of his analysis was well below 30 percent (assuming medium effect size). Thus he only had a one in three chance of detecting the phenomena, assuming that they do in fact exist. Once again the problem of power causes the author to erroneously claim that nonsignificant findings indicate no difference, rather than the more appropriate statement that no significant findings were demonstrated, possibly because of inadequate power.

Collectively these misstatements and inconclusive findings may result in MIS researchers prematurely abandoning what may be promising areas of research. By incorrectly concluding that the phenomenon under investigation has no effect or makes no difference, they are discouraging researchers from pursuing this direction in other studies. Thus the result of inadequate power may erroneously close off avenues of research that in fact may be important. Reflecting on these findings it appears that much future research effort may be necessary to determine which of the earlier negative findings were correct in their conclusions of no demonstrable, consequential effects, as opposed to those that were not found by tests that simply were not powerful enough to detect them.

6.2. The Problem of Significance

A related problem, but one which is different from the above, is that of interpreting significant findings when one has extraordinarily low power, for a posited medium effect size. On examining the studies we would occasionally find a test which had very low power, yet where many of the results were significant. From the standpoint of power, for these studies to have yielded strongly significant results, it is highly likely that the relationships explored must have been enormously strong. To translate this into Cohen's words: the relationship must have been quite obvious without any empirical testing.

An example of this was a paper by Harel and McLean [1985]. They compared a non-procedural language with a procedural one in terms of programmer productivity and program efficiency. They found strongly significant results which indicated that applications were written faster in the non-procedural language but that they were less efficient as they had a slower CPU execution time. Yet the power of their study was under 25 percent for these tests, assuming a medium effect size. While their study has many interesting points, the interesting findings are not the relationships detected by the statistical analyses. We doubt that many researchers were suprised that fourth generation languages created applications that ran slower yet were faster to write. There were many studies of this nature which asked questions to which the answers are quite obvious without any statistical testing. We believe this problem is directly attributable to our research tradition which demands statistical inference testing before a study is considered "publishable". Little is learned from testing hypotheses that have obvious relations or from applying statistical tests to studies which have very small samples and hence highly limited opportunities to detect any significant findings.

6.3. The Problem of Reliability

While the above picture may appear grim, the scenario worsens when the reliability of the studies is taken into account. The power figures presented in Table 4 assume measurement instruments which possess 100 percent reliability. The statistical power of a study sharply declines as the reliability of the instrument degrades. Schmidt, Hunter, and Urry [1976] investigated statistical power in criterion-related studies and found that in validation studies, criterion unreliability and restriction in range dramatically increased the necessary sample size required to maintain adequate levels of power.

Given that only a handful of the studies even addressed the issues of reliability with none discussing range restriction, we can assume that the actual power of the studies surveyed is substantially less than what was reported in Table 4. For example, Zmud's [1982] study was one of the few which reported reliability coefficients. His reliabilities ran from .69 to .76, well below the 1.00 levels assumed by Table 4. When researchers calculate the sample size necessary to achieve a particular statistical power level they should also consider the reliability of their instruments. The less certain that researchers are of the cleanliness of their instruments, the greater the sample size they will need.

7. CONCLUSIONS

Twenty years ago Cohen [1962] investigated the statistical power of abnormal-social psychology studies and found the average power of these studies to be unacceptablly low. Twenty years later we have found the average power of MIS studies to also be unacceptably low. This is an unfortunate situation, resulting in many MIS researchers drawing incorrect conclusions and making numerous misstatements. Fortunately, the problem of power is one that can, with diligence, be remedied. The following are our recommendations.

First, before any data is collected researchers must determine what sample size they require in order to achieve a .80 power level. If they are uncertain as to the appropriate effect size then they should employ the convention of medium effects. If they are unable to control their sample size they should still conduct a power analysis to provide perspective on the meaning of their findings.

Second, all published empirical studies should be required to report the power of their study under reasonable effect size assumptions. This would allow the reader to be wary of those studies with low power and would make explicit where it may be possible to interpret nonsignificant results as indicating little or no effect, distinct from inconclusive findings, possibly attributable to low power.

Finally, MIS researchers should be aware that there are other research methods available besides inference testing which may be more appropriate to exploring the phenomena of interest. There is much that can be learnt from employing exploratory, hypothesis-generating techniques that offer the richness of qualitative data analysis. Historical analyses, ethnographies, action research, phenomenology, ethnomethodology and critical research approaches offer much scope for insightful and valuable knowledge yet these are rarely employed by MIS researchers.

REFERENCES

1. Chonda, R. and J. New, Information Support for Decision-Maker Learning in a Competitive Environment: An Experimental Study, Decision Sciences, Vol.II, 1980, pp. 603-615.

2. Cohen, Jacob, The Statistical Power of Abnormal-Social Psychological Research: A Review, Journal of Applied Psychology, Vol.65, No.3, 1962, pp. 145-153.

3. Cohen, Jacob, Some Statistical issues in Psychological Research, in B.B. Woleman (ed) Handbook of Clinical Psychology, McGraw-Hill, New York, 1965, pp. 95-121.

4. Cohen, Jacob, Statistical Power Analysis for the Behavioral Sciences, (revised ed.), Academic Press, New York, 1977.

5. Cohen, Jacob and Patricia Cohen, Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences, Lawrence Erlbaum Assoc. Publishers, Hillsdale NJ, 1983.

6. Harel, E. and E.R. McLean, The Effects of Using a Nonprocedural Computer Language on Programmer Productivity, MIS Quarterly, June 1985, pp. 109-120.

7. Hays, W.L., Statistics, (3rd edition), Holt, Rinehart & Winston Inc., New York, 1981.

8. Keppel, G., Design and Analysis: A Researcher's Handbook, Prentice Hall, Inc., Englewood Cliffs NJ, 1973.

9. Miara, R.J, Musselman, J., Navarro, J. and B. Shneiderman, Indentation and Comprehensibilty, Communications of the ACM, Vol.26, No.11, 1983, pp. 861-867.

10. Remus, W., An Empirical Investigation of the Impact of Graphical and Tabular Data Presentations on Decision Making, Management Science, Vol.30, No.5, 1984, pp. 533-542.

11. Schmidt, F., Hunter, J., and V. Urry, Statistical Power in Criterion-Related Validation Studies, Journal of Applied Psychology, Vol.61, No.4, 1976, pp. 473-485.

12. Shneiderman, B., Control Flow and Data Structure Documentation: Two Experiments, Communications of The ACM, Vol.25, No.1, 1982, pp. 55-63.

13. Welkowitz, J., Ewen, R. and J. Cohen, Statistics for the Behavioral Sciences, (3rd ed)., Academic Press, New York, 1982.

14. Zmud, R., Diffusion of Modern Software Practices: Influence of Centralization and Formalization, Management Science, Vol.28, No.12, 1982, pp. 1421-1431.