A BUSINESS RESEARCH: SIGNIFICANCE OF STATISTICAL POWER, LEVEL OF SIGNIFICANCE IN HYPOTHESIS TESTING

A.A Ghewari[1*] and A. Singh[2]
[1]Camp Education Society's Rasiklal M Dhariwal Institute of Management, Pune, MS, India
[2]STES Sinhgad Institute of Management and Computer Application, Pune, MS, India
asitaghewari@gmail.com[1*], anamikasingh.0502@gmail.com[2]

ABSTRACT

The purpose of writing this research paper is to describe how well management research use the statistical power with regard to studies specifically in testing of hypothesis. Interpreting statistical inferences mandates that researchers specify acceptable levels of statistical error. The most common approach is to specify the level of Type I error, generally represented as α. Formally defined, a Type I error is the probability of rejecting the null hypothesis when it is actually true. On average, the attention paid by researchers to the two types of statistical inference errors (Type I and Type II) is by far not equal. The belief is that the consequences of a false positive (Type I error) claim are more serious than those of a false negative (Type II error) claim. As a result, Type I errors are usually focused on more frequently and guarded against more stringently by researchers.

Keywords: *hypothesis testing; null hypothesis; statistical power; power analysis*

Introduction

Statistical Power surveys have been performed in areas such as management (e.g., Ferguson & Ketchen, 1999; management information systems (e.g., Baroudi & Orlikowski, 1989), education (e.g., Brewer, 1972), Cohen's (1962).

In general, the common finding is that deficient statistical power spread research in these diverse areas of study. Such findings could be observed to the fact that power issues and power analyses tend to receive inadequate attention by researchers. Cohen (1992) addressed this issue by stating, "It is not at all clear why researchers continue to ignore power analysis. Equal frustrations were noted by Sedlmeier and Gigerenzer (1989), who suggested that concerns about power in research are almost nonexistent. Nickerson (2000) suggested that such in attention might be attributable to statistical power not frequently being understood and, as a result, not often employed in research. Additional evidence was offered by Mone et al. (1996), who noted that the impact of power assessment surveys has been minimal and that calls for greater statistical power levels and usage have gone unheeded. In their study, Mone et al. surveyed the authors of a sample of studies contained in top-tier journals and found that almost two thirds of the respondents never used power analysis. Furthermore, the respondents stated that there is little call for greater usage of power analysis by journal editors or reviewers. Echoing these aforementioned concerns, authors of power assessment articles argue that insufficient statistical power may leave researchers unable to detect or reject false null hypotheses. In other words, researchers may actually fail to notice meaningful differences or effects as a result of low power. Cohen (1977) noted that such an occurrence is highly unfavorable to behavioral scientists because it is then reasonable to suggest that there is not an equitable chance of rejecting the null hypothesis, and, in general, behavioral scientists "typically hope to 'reject' [the null] hypothesis and thus 'prove' that the phenomena in question is in fact present" . Cohen and other authors of statistical power assessments basically conclude that a failure to reject a null hypothesis leaves readers wondering whether it is due to insufficient statistical power or truly due to the absence of the phenomenon.

Additional concerns about designing studies with low power are noted by Howard, Maxwell, and Fleming (2000) who suggested that such actions "tend to lead to a body of literature in which results appear to contradict one another" Researchers usually wish to demonstrate that the phenomenon in question is present (i.e., reject the null hypothesis in favor of the alternative hypothesis). However, there are instances in which researchers do have a priori, theoretically justified reasons to hypothesize formal, statistical null relationships (Cohen, 1990; Cortina & Dunlap, 1997; Cortina & Folger, 1998; Greenwald, 1975, 1993). Support for positing and testing null relationships between variables of interest is offered by Greenwald (1993), who noted that "scientific advance is often most powerfully achieved by rejecting theories. A major strategy for doing this is to demonstrate that relationships predicted by a theory are not obtained, and this would often require acceptance of a null hypothesis" . It was believed that theoretically based arguments that lead researchers to forecast null relationships between their research variables of interest are justified.

If power levels are not sufficiently high, then attainment conclusions of "no effect" may not be realistic and potentially lead to conflicting and/or invalid findings in the literature. Only if power levels are high can one come close to inferring that the null hypothesis is true when there is a failure of a test to establish statistical significance (Nickerson, 2000). Thus, power is particularly important for those testing null hypotheses because lack of power may in fact lead researchers to incorrectly affirm null hypotheses.

Statistical Power, Level of Significance and the Null Hypothesis

There are three main components that determine the level of statistical power of an inference test: the significance level (α), the sample size, and the effect size. The relationship between power and its three determinants is such that if one of the four elements (i.e., power, significance level, sample size, or effect size) is unknown, it can be calculated using the known values of the other three elements. Hence, researchers are often able to a priori determine statistical power levels of their tests. Researchers investigating a phenomenon typically hypothesize that a relationship between the investigated variables exists. Classical statistical inference tests posit a null hypothesis (Ho : the phenomenon under investigation is absent, or there is no—or at best a trivial—difference between the parameters being tested), which researchers contrast against the alternative hypothesis (Ha : the phenomenon is present, or there is a difference in the parameters being tested). Because researchers typically hope to reject the null hypothesis, they normally report the probabilities associated with the likelihood that such a conclusion is erroneous (i.e., α). However, when such tests are not significant or when one expects the null hypothesis to be upheld, it is critical to discuss the likelihood of rejecting the null hypothesis in favor of the alternative hypothesis if the alternative hypothesis is in fact true. Such a probability is better known as statistical power. Frequently, power is represented as 1 – β, where β is the probability of failing to reject the null hypothesis when it is actually false. Such an error is commonly referred to as a Type II error.

Significance Level

Interpreting statistical inferences mandates that researchers state acceptable levels of statistical error. The most common approach is to specify the level of Type I error, generally represented as α. Formally defined, a Type I error is the probability of rejecting the null hypothesis when it is actually true. On average, the attention paid by researchers to the two types of statistical inference errors (Type I and Type II) is by far not equal. The belief is that the consequences of a false positive (Type I error) claim are more serious than those of a false negative (Type II error) claim. As a result, Type I errors are usually focused on more frequently and guarded against more stringently by researchers (Baroudi & Orlikowski, 1989; Brewer, 1972; Chase & Chase, 1976; Cohen, 1977; Cowles & Davis, 1982; Greenwald, 1993;) It was suggested that researchers set their level of β to correspond to the traditional level of α when testing a non-null (i.e., alternative) hypothesis, which typically is set at the .05 level. Thus, when null hypotheses serve as the research hypotheses of interest, the researcher should opt for a β level of .05, which corresponds to a .95 power level; otherwise, statistical insignificance of the tests has no real significance. Because power is 1 – β, then at a power level of .80, β = .20, which means that there is a .20 probability of sustaining a false null. We argue that this represents a power level too low and a probability of Type II error too high to confidently affirm the null hypothesis. Such arguments are supported by Rossi (1990) who suggested, If power was high, then failure to reject the null can, within limits, be considered as an affirmation of the null hypothesis, because the probability of a Type II error must be low. Thus, in the same way that a statistically significant test result permits the rejection of the null hypothesis with only a small probability of error (alpha, the Type I error rate), high power permits the rejection of the alternative hypothesis with a relatively small probability of error (beta, the Type II error rate). (p. 646) We fully understand that some advocate setting β levels according to each situation and the overall cost of the error, but we firmly advocate a minimal power level of .95 (β = .05), if possible, for researchers to have confidence in their results and to guard against building a literature of contradictory results. This is particularly important for those testing null hypotheses.

Sample Size

As the number of observations in the sample increases, the reliability (i.e., precision) of the sample value approximating the population value also increases (Cohen, 1977). As a result of this greater reliability, a researcher has a higher probability of rejecting a false null hypothesis. Thus, as the sample size increases, so does the power of the study. Ideally, researchers should specify α, effect size, and the desired level of power and then determine the sample size needed in the study so that more valid conclusions can be drawn from the results of testing null hypotheses.

Effect Size

The effect size represents the magnitude or strength of the relationship between the variables in the population (Cohen, 1977). As previously argued, researchers can fail to reject the null hypothesis when the true relationship between two events of interest is determined to be trivial or inconsequential. In other words, affirmation of null hypotheses does not occur when the true value of a statistic equals exactly zero, but rather the relationship between variables of interest is so small as not to be usefully distinct from zero. Cohen (1977) and Sedlmeier and Gigerenzer (1989) argued that determination of a trivial effect is made when power (1 – β) is set at a high value and the sample size used is large enough so that the risk of Type II error (β) is relatively small and similar to that of the risk of Type I error, which is commonly set at the .05 level.

When conducting a power analysis as part of testing the null hypothesis, it is important to determine when an effect is large enough to be considered nontrivial. Lane, Cannella, and Lubatkin (1998) notes that, conceptually, a trivial effect implies a small effect size, as defined by the conventional values set forth by Cohen (1977, 1992). Cohen (1977, 1990) demonstrated that if a researcher considers an effect size ofr = .10 (a small effect size for a correlation according to Cohen) as negligible and wishes to test the null hypothesis (α = .05, power = .95, and β = .05), then a sample size of 1,308 is required. It is obvious that the use of small effect sizes places great demands on the sample sizes of studies. From this example, it appears that it takes an impractically large sample size to fail to reject the null hypothesis; however, “the procedure makes clear what it takes to say or imply from a failure to reject the null hypothesis that there is no nontrivial effect” (Cohen, 1990, p. 1309). So, with a small value for the effect size (i) and power set at a high value (so that β is relatively small), nonsignificance of results allows the researcher to properly conclude that the population effect size is no more than i (i.e., negligible), a conclusion significant at the specified level of β. Thus, drawing on the logic with which we reject the null hypothesis with risk equal to α, Cohen (1977) stated, The null hypothesis can be accepted in preference to that which holds that the effect size equals i with risk equal to β. Since i is negligible, the conclusion that the population effect size is not as large as i is equivalent to concluding that there is “no” (nontrivial) effect. (p. 16) For this power assessment, we felt that drawing on general approximations of small effect sizes for the statistical tests covered by Cohen (1977, 1992) was appropriate.

Recommendations for Testing Null Hypotheses

Recommendation 1: Researchers should report power for every standard statistical test. This allows the researcher to understand the risks associated with non significant findings.

Recommendation 2: Researchers should establish a β level of .05 or lower to confidently conclude that a unimportant effect exists between variables of interest (Cohen, 1977; Sedlmeier & Gigerenzer, 1989; Rossi, 1990). Thus, the risk of a Type II error will at least parallel to generally accepted levels for Type I errors.

Recommendation 3: Researchers try to include confidence intervals in their findings to provide further detail that the hypothesized null effect is not trivial due to sampling error (Cortina & Folger, 1998; Nickerson, 2000).

Recommendation 4: Add in in each experiment and analysis an additional independent variable that is recognized as having a relationship with the dependent variable (Cortina & Folger, 1998). Next, evaluate the relationship in each experiment/analysis to exemplify that the independent variable of interest has a zero or trivially nonzero relationship with the dependent variable while the additional independent variable has a significant nonzero rapport with the dependent variable.

Conclusion

Researchers do not want to conduct an study of low statistical power" (Rossi, 1990,). Also, Researcher wanted to determine whether researchers made power considerations when concluding support for a null hypothesis based on nonsignificant findings. Because presenting statistical null hypotheses in research is generally frowned upon by a majority of scholars, any time researchers present null hypotheses, they must do their highest to demonstrate that sufficient statistical power is present to assertively affirm or disaffirm the hypothesis. In addition, it is important to note that relegating the null hypothesis to a secondary status is unwarranted (Atkinson, Furlong, & Wampold, 1982; Cortina & Folger, 1998; Frick, 1995; Greenwald, 1993). As such, we, and others, believe that it is acceptable for a null hypothesis (i.e., a hypothesis of trivial effect or difference) to be offered on its own, if theoretically justified. In other words, "there is no suggestion that a null hypothesis must be used as a comparison value" (Cortina & Folger, 1998, p. 335; Greenwald, 1993). Probability of failing to reject false null hypotheses is greater than what is advocated in the power literature as it pertains to testing formal null hypotheses. It is hoped that the recommendations offered for future management research can prove to be of assistance for others testing null hypotheses.

References

1. Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.
2. Chow, S. L. (1996). Statistical significance: Rationale, validity, and utility. Thousand Oaks, CA: Sage.
3. Cohen, J. (1962). The statistical power of abnormal-social psychological research. Journal of Abnormal and Social Psychology, 65, 145-153.
4. Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). New York: Academic Press.
5. Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.
6. Cowles, M. P., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-558.
7. Ferguson, T. D., & Ketchen, D. J., Jr. (1999). Organizational configurations and performance:
8. Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 419-448). Hillsdale, NJ: Lawrence Erlbaum.

9. Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. Journal of Communication, 23, 251-265.
10. Lane, P. J., Cannella, A. A., & Lubatkin, M. H. (1998). Agency problems as antecedents to unrelated mergers and diversification: Amihud and Lev reconsidered. Strategic Management Journal, 19, 555-578.
11. Lawless, M. W., & Anderson, P. C. (1996). Generational technological change: Effects of innovation and local rivalry on performance. Academy of Management Journal, 39, 1185-1217.
12. Mazen, A. M., Graf, L. A., Kellogg, C. E., & Hemmasi, M. (1987). Statistical power in contemporary management research. Academy of Management Journal, 30, 369-380.
13. Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, 5, 241-301.
14. Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.