STATISTICAL POWER ANALYSIS

Burak Keskin – Altan Aktas

Abstract

The power analysis is a method that enables to plan a highly valid and reliable research and to guarantee the validity, reliability and sensibility of the results of a research. Statistical power, in addition, is an approach which is used to evaluate to what degree the decisions obtained as a result of statistical tests used to test an aim are valid and reliable in terms of probability values. The power analysis is different from the other statistical methods to a great extent. Various statistical analyses begin the analysis with the existing data and focuses on the comments of the consequences. Nevertheless, power analysis is meaningful before data collection and significant in the process of setting the hypothesis.

In this context, the major purposes of this study are; to contribute researches in social sciences by creating awareness about what the power analysis means and to guide researchers about how the power analyses can be employed in a social science study.

Key words: Statistics, Statistical Power, Statistical Power Analysis

JEL Code: C12, C18, C19

Introduction

Many factors can influence the quality of research results. The research design, data collection procedures, data analysis techniques and research's perspectives all influence the accuracy of research results. When a statistical significance test is used, several additional factors will influence the accuracy of the reported test result. These factors include the significance level, sample size, effect size and statistical power. Power analysis is considered a newcomer to the field of statistics in comparison to statistical significance testing. The first systematic power analysis was conducted by Jacop Cohen on 70 articles in the 1962 volume of the *Journal of Abnormal and Social Psychology*. Accordingly, Cohen is recognized as the individual attributed with making researchers aware of the technique and the importance of power analysis. This article explains the concept of power and power analysis. Additionally, serious consequences of ignoring power are described at planning and interpreting stages of research.

1. The Concept of Statistical Power

When I stumbled on power analysis… It was as if I had died and gone to heaven. ~ Jacop Cohen

The development of the concept of statistical power is attributed to Jerzy Neyman and Egon S. Pearson, they introduced the concept in 1928 (Cohen, 1990). Since then, others have contributed to the concept of statistical power but most notable among writings on statistical power is Cohen's *Statistical Power Analysis for the Behavioral Sciences* (1988). It is considered the standard reference on statistical power in the social sciences (Clark, 1996). Cohen (1970) defined statistical power as the probability of rejecting a false null hypothesis. Cohen (1988) expanded the definition and stated "the power of a statistical test is the probability that it will yield statistically significant results." For example, if a researcher found that his/her study has a statistical power value of 0.50, the researcher had a %50 chance of rejecting a false null hypothesis. If the research were repeated 100 times, the researcher would reject a false null hypothesis in 50 studies.

Power analysis differs in important ways from other statistical approaches. Most statistical analyses begin with existing data, subject the data analysis and then focus on interpretation of the results. Power analysis is different. Power analysis does not involve existing data. In fact, power analyses are usually meaningful when conducted prior to data collection. In this manner, it is useful to think of power analysis as part of the hypothesis statement process. Another way power analysis differs from other statistical analyses is in term of interpretation. For most statistical procedures, texts devote considerable time to interpretation of result of computer output. In contrast, the output for power analysis is simple and requires little interpretation or discussion. Generally, output provides a single value, the power for the test of interest. The interpretation of output for such analyses does not involve much interpretation aside from an evaluation of whether our study is sensitive enough to detect our effects of interest given a particular sample size (Aberson, 2010).

Power analysis involves to specify four variables; the power of a statistical test, significance level (α), the effect size (ES) in the population and sample size (n). When any three of these four parameters are specified, the fourth can be determined. Thus, for any given statistical test, power is derivative of sample size, effect size and significance level. While numerous articles, tables or computer programs describe how to calculate power for various statistical tests, Cohen's power tables are recommended as the standard source for determining power or necessary sample size for obtaining a desired power level. Of the three

values (n, α and ES) required for computing power, ES is the most difficult to specify. ES can estimate upon similar studies, the minimum effect which would be of meaningful importance and Cohen's conventional values. Cohen's conventional definitions of effect sizes (small, medium, large) are probably the best known and most widely accepted guidelines (Clark, 1996).

Cohen (1992) noted that "my intent was that medium ES represent an effect likely to be visible to the naked eye of a careful observer". In this article, Cohen also pointed out that a medium ES "approximates to the average size of observed effects in various fields". He stated small ES "to be noticeably smaller than medium but not so small as to be trivial" and large ES "to be the same distance above medium as small is below it".

Effect size can be defined as the "degree to which a phenomenon exists" (Cohen, 1988). Actually, effect size is desired difference to be detected between the value specified in null hypothesis and the value specified in alternative hypothesis. If the ES is expressed in terms of standard deviation units, it is called standardized effect size or ES index (d). For example, in calculating the power of statistical tests for difference between means, the standardized effect size can be obtained by dividing the difference between estimated population means by the estimated population standard deviation or by the standard deviation of the sample if the population standard deviation is unknown (Deng, 2000).

Power analysis is beneficial in both phases of research design and interpretation. During the design phase, power analysis provides the capability of an a priori knowledge of whether the study has a reasonable chance of obtaining statistically significant results. In the even that power analysis detects low power, design parameters can be modified to increase power or the study can be abandoned entirely. Thus, in the case of extremely low power, power analysis contributes to a more efficient research design that results in savings of time, effort and money (Clark, 1996).

There is not set standard as to what constitutes adequate power. In regard to what is suggested as an acceptable level of power, recommendations range from at least 0.50 to as high as 0.95. Any level of power ranging between 0.70 and 0.85 might generally be satisfactory. Cohen's recommendation is that power should be at least 0.80.

The following sample is convenient in order to comprehend the importance of power analysis: An astronomer is interested in building a telescope to study a distant galaxy. A

critical factor in the design of the telescope is its magnification power. Seen through a telescope with insufficient power, the galaxy will appear as an indecipherable blur. But rather than figure out how much power he needs to make his observations, the astronomer foolishly decides to build a telescope on the basis of available funds. Maybe he does not know how much magnification power he needs, but he knows exactly how much money is in his equipment budget. So he orders the biggest telescope he can afford and hopes for the best.

In social science research the foolish astronomer is the one who sets sample sizes on the basis of resource availability. He is the one who when asked "how big should your sample be?", answers "as big as I can afford." Research constraints are a fact of research life. But if our goal is to conserve limited resources, it is essential that we begin our studies by asking questions about their power to detect the phenomena we are seeking. How big a sample size do I need to test my hypothesis? Assuming the phenomenon I am searching for is real, what are my chances of finding it given my research design? How can I increase my chances? My sample is only 50 or 30 or 200; do I have enough power to run a statistical test? Power analysis provides answers to these sorts of questions (Ellis, 2010).

1.1. Factors Affecting Statistical Power

Principal factors that affecting the statistical power of a study are significance level (α), sample size (n) and effect size (ES). These three factors are by no means the only considerations affecting the power of a study, but these are the most basic. Increasing any of these three factors will increase power on the condition that other parameters are fixed (Rossi, 2012). Additionally, the directionally of the hypothesis (one tail rather than two tail), the use of a more powerful statistical test (parametric rather than nonparametric), increasing the reliability of measurement on the dependent variable and research design have a direct effect on the level of statistical power. But of the factors specified in the related literature as affecting power, increasing sample size and specifying a higher risk level for type I error by relaxing alpha are two most common strategies suggested for increasing power (Clark, 1996; Mazen et. al., 1987; O'Keefe, 2007).

1.2. Types of Statistical Power Analysis

Power analysis typically takes two forms, priori power analysis and post-hoc power analysis. In priori power analysis, power is calculated before conducting significance tests

evaluating the hypothesis interest, whereas in post-hoc power analysis, power is calculated after conducting these significance tests (Cafri et. al., 2009).

1.2.1. Priori Power Analysis

Priori power analysis considered the ideal type of power analysis by most authors. In a priori power analysis, researchers specify the size of effect to be detected (i.e., measure of the "distance" between H_0 and H_1), the "α" level and the desired power (1-β) of the test. Given these specifications it is possible to compute the necessary sample size "n" (Erdfelder et.al., 1996). For the purposes of designing a research study, priori power analysis is of great utility.

1.2.2. Post-Hoc Power Analysis

Power analyses can be useful during the design stage of a study. But sometimes power analyses are run after the data have been analyzed and especially when insignificance results occurred. Unlike a priori power analysis that relies on estimates of effect size, post-hoc power analysis conveys the actual power in the study through the observed effect size rather than an estimated value. The advantage of this method is that the occurrence of a statistically nonsignificant finding may be evaluated. Statistically nonsignificance may occur because of insufficient power resulting from an insufficient sample size and/or a less than meaningful effect size (Balkin & Sheperis, 2011).

Nonsignificant results are a researcher's bane and running a power analysis prior to a study is no guarantee that results will turn out as expected. Priori analyses hing on anticipating the correct effect size, but if effects smaller than expected, then resulting power may be inadequate. Reassessing power based on the observed rather than the estimated effect size is sometimes done to determine actual power as opposed to planned power. If it can be shown that power was low, the researcher might conclude: "the results are not significant but that was because the test was not sufficiently powerful." However, if power levels are found to be adequate, then the researcher can conclude that the result was negative (Ellis, 2010). The post-hoc power analysis of nonsignificant results is sometimes painted as controversial by few authors.

1.3. The Problems Caused by Insufficient or Overpowered Studies

Power should be a primary consideration in many decisions in the design and implementation of a statistical study. Lack of a power may lead to erroneous decisions

concerning the null hypothesis. A deficiency in power increases the probability that a type II error will occur. In other words, low power increases the likelihood that a researcher will make a decision to retain a false null hypothesis. Such decisions leave possibly valuable research uncovered (DiLullo, 1997).

Studies that conducted with sufficient statistical power provide to be successful enough chance to researchers. Research which incorporates low power runs the same risk as "fishing for minnows with a tuna net: You probably won't catch any minnows, but you can't conclude your study there are none in the pond; everything just slips through the net." (Clark, 1996).

Awareness of the dangers associated with low statistical power is slowly increasing. A taskforce commissioned by the American Psychological Association recommended that investigators assess the power of their studies prior to data collection. Now it is not unusual for funding agencies and university grants committees to ask applicants to submit the result of priori power analysis together with their research proposals. Some journals also require contributors to quantify the possibility that their results are affected by type II errors, which implies an assessment of their study's statistical power. Despite these initiatives, surveys reveal that most investigators remain ignorant of power issues. The proportion of studies that merely mention power has been found to be in the 0-4% range for disciplines from economics and accounting to education and psychology (Ellis, 2010).

Some authors have cautioned against the possibility of "overpowered" research, which could be wasteful in terms of scarce funding resources or in the possibility of detecting trivially small effects. However, little extra power might not hurt in some situations. It is common for funding agencies to expect a minimum power of 0.80 for primary study outcomes.

Cohen, suggested to researchers $\beta/\alpha = 4$ proportion. Many researchers use generally 0.80 power level that suggested Cohen and Fisher's 0.05 significance level that known as a standard. These two ratios are called as "5-80" (five-eighty). This ratio has become a tradition for researchers and facilitated many things.

Researchers sometimes compare groups to see whether there are meaningful differences between them and, if so, to assess the statistical significance of these differences. The statistical significance of any observed difference will be affected by power of the statistical test. As statistical power increases, the cut-offs for statistical significance fall. Taken to an

extreme this can lead to the bizarre situation where two essentially identical groups are found to be statistically different. Field and Wright (2006) give the following SPSS-generated results to show how this situation might be:

t	df	Sig. (2-tailed)	Mean difference
−2.296	999998	.022	.00

The number in the last column tells us that the difference between two groups on a particular outcome is zero, yet this "difference" is statistically significant at the $p < 0.05$ level. How is it possible that two identical groups statistically different? In this case, the actual difference between the two groups was not zero but 0.0046, which SPSS round up to 0.00. Most would agree that 0.0046 is not a meaningful difference; the groups are essentially same. Yet this microscopic difference was judged to be statistically significant because the test was based on a massive sample of a million data-points. This situation demonstrates one of the dangers of running overpowered tests. A researcher who is more sensitive to the *p* value than the effect size might wrongly conclude that the statistically significant result indicates a meaningful difference.

What, then, is an appropriate level of statistical power? This is not an easy question to answer as it involves a trade-off between risk and return. If power is set a high level, say 0.90, then the chance of detected effects are greatly improved. But statistical power is little costly. To detect small effect, such as d = 0.20, using a nondirectional test with alpha level at 0.05 and beta level at 0.10 would require a sample of n = 858 (Ellis, 2010)

1.4. G-Power: A General Power Analysis Program

In the past, researchers have used imprecise methods to decide required sample sizes rather than using priori power analysis. These methods have many times shaped on budget of research, time limitation and recommendations of experts. Cohen (1992) said that "researcher may find power analysis too complicated and may simply avoid the issue." For that reason, power tables were created by Cohen in order to facilitate calculation of statistical power. In addition, computer software programs were created by many experts. G-Power is one of them. G-Power is a free software program to assist researchers in conducting power analyses either priori or post-hoc. G-Power can be download from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register web site for free.

Conclusion

The results of this study suggest that, more than 50 years after Cohen (1962) conducted his study, statistical power of studies are still low. In addition, researchers have not considered sample size, significance level and effect sizes in the light of a priori power analysis. It is clear that the concept of statistical power needs to be improved. Statistical power should be addressed in planning and analysis stages of the study. Researchers should perform a priori power analyses and include the following information: the alpha level, desired statistical power, the minimum effect size expected and the sample size needed. After the data are collected and statistically analyzed, the observed power and effect size of the study should be reported to help interpretations of the results. Editors and reviewers should assist in changing submission policies by expecting authors to explain their use of sample sizes, statistical power, effect size and alpha levels in their results.

References

Aberson, C. L. (2010). Applied power analysis for the behavioral sciences, New York: Routledge.

Balkin, R. S. & Sheperis, K. J. (2011). Evaluating and Reporting Statistical Power in Counseling Research, Journal of Counseling and Development: JCD; Summer 2011, 89, 268-272.

Cafri, G., Kromrey, J. D. & Brannick, M. T. (2009). A Sas Macro for Statistical Power Analysis in Meta-Analysis. Behavior Research Methods, 41(1), 35-46.

Clark, D. (1996). Statistical Power as a Contributing Factor Affecting Significance Among Dissertations in the School of Religious Education at Southwestern Baptist Theological Seminary. Doctoral Dissertation, USA.

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

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I Have Learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Deng, H. (2000). Statistical Power Analysis of Dissertations Completed by Students Majoring in Educational Leadership at Tennessee Universities. Doctoral Dissertation, East Tennesse State University, Tennessee, USA.

DiLullo, L. K. (1997). A Post Hoc Power Analysis of Inferential Research Examining The Relationship Between Mathematic Anxiety and Mathematic Performance. Doctoral Dissertation, Auburn University, Alabama, USA.

Ellis, P. D. (2010). The Essential Guide to Effect Size, Statistical Power, Meta-Analysis and Interpretation Research Results, Cambridge University Press.

Erdfelder, E., Faul, F. & Buchner, A. (1996). G Power: A General Power Analysis Program. Behavior Research Methods, Instruments & Computers, 28(1), 1-11.

Mazen, A. M., Graf, L., A., Kellog, K. E. & Hemmasi, M. (1987). Statistical Power in Contemporary Management Research. The Academy of Management Journal, 30(2), 369-380.

O'Keefe, D. J. (2007). Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses. Communication Methods and Measures, Lawrence Erlbaum Associates, Inc, I(4), 291-299.

Contacts

Burak KESKIN

Cankiri Karatekin University – Faculty of Economics and Administrative Sciences

Department of Business Administration

Cumhuriyet Mahallesi S. P. Ustegmen Erdem Ozturk Sokak 18100 CANKIRI

burakkeskiin@karatekin.edu.tr

Altan AKTAS

Cankiri Karatekin University – Faculty of Economics and Administrative Sciences

Department of International Relations

Cumhuriyet Mahallesi S. P. Ustegmen Erdem Ozturk Sokak 18100 CANKIRI

altanaktas@karatekin.edu.tr