The Perils of Insufficient Statistical Power
A Comparative Evaluation of Power and Sample Size Analysis Programs

Robert A. Yaffee
Connect Archive
New York University

30 January 1997

Contents

New statistical modules and packages for power and sample size analysis are making their appearance. Professors, researchers, and advanced graduate students interested in doing serious research and applying for grants need to know how to use these programs. For this reason, attention is directed to a few of the prominent and useful innovations in these tools of analysis. Three of these most needed of programs are Sample Power, by NYU Professor Jacob Cohen, StatPower, by Dr. James L. Bavery, and Power Analysis and Sample Size (PASS 6.0), by Jerry Hintze, all of which perform power and sample size analysis. The Statistics and Social Science Group at ACF has already acquired (StatPower from Scientific Software, Inc. and is in the process of obtaining Sample Power, from SPSS, Inc.

1 Skating on Thin Statistical Ice

To help researchers and doctoral students avoid the disasters and catastrophes of insufficient data, power analysis is necessary to be sure that the sample size of a study is large enough so that the statistical tests can actually detect the differences that they purport to find. If the sample size is too low, the standard statistical tests will not have the statistical power to detect differences that really exist. What happens in these cases is that no significant difference is found, although in reality such a difference exists. With the decline in sample size, the probability of acceptance of a false null hypothesis, sometimes referred to as beta, increases under such circumstances. False similarities plague the findings where differences in fact exist. This masking of differences is one of the reasons that studies with small sample sizes, lacking the proper preliminary power and sample size analysis, are treated with suspicion by statistical cognoscenti. If medical researchers are engaged in clinical trials of a drug, insufficient sample size that undermines assessment is criminal.

Moreover, this is the principal reason for which persons applying for large grants usually have to perform power and sample size analyzes for the statistical tests they are planning and show that they can detect small or medium effects with sufficient power (usually of 0.8). With a power of 0.8, the user has a 20% chance of making a type II error – obtaining a false negative result (in other words, the failure to detect a real difference when it exists) from insufficient sample size. For these reasons, it is important for researchers to understand the concept of statistical power and to know how and when to apply it.

2 Criminal Negligence and Wasting Resources

If the sample size of the study is too large, then the monetary, temporal, and scheduling costs of gathering the data may be so formidable that only the resourceful can do the research. If, however, students and researchers manage to muster the time and funds necessary to collect the data, they may develop much more power than they need. If the study is a clinical trial involving the random assignment of patients to an experimental group during the toxicity testing phase, then placing more subjects at risk than necessary, even though they might have given informed consent, is criminal. As the sample size increases, the size of the standard error decreases and power increases, for a given level of probability of alpha, the Type I error. A Type I error, rejection of a null hypothesis where there should have been no rejection, is in inverse proportion to the power. Too large a sample size means the waste of valuable resources in the process of the data collection. The key questions are how much sample size is enough, how much safety margin is needed, and how much is too much. In short, what is the optimal sample size?

3 Purpose of the Pilot Study

Standard social science research procedure involves a pilot study. From the pilot study, the researchers learn the cooperation rate in the target population. They glean a preliminary assessment of the prevalence rate of a trait, characteristic, or disease. From the power and sample size analysis, the researchers ascertain the needed sample size. By multiplying the prevalence rate by the cooperation rate and dividing this product into the needed sample size, the researchers can calculate how many persons they have to target for interviews or questionnaire administration to procure the proper sample size for their analysis. From their margin of error, they can determine how much of a safety margin they should allow in their calculations.

4 Sample Power, PASS, and Sample Power

Sample Power, developed from the older Power Analysis program by Michael Borenstein and Jacob Cohen and marketed by SPSS, Inc., is a user-friendly and delightfully simple program that handles much of the bread and butter basic statistics that undergraduates and some graduates need to do. This program computes sample size needed for a desired level of power or, alternatively, it calculates the power from a given sample size. The user is sometimes asked to include additional parameters explaining the nature of the analysis. For example, suppose the analyst wishes to compute the power of a chi-square test. The analyst specifies the alpha level, the number of degrees of freedom for the test, and the noncentrality parameter, whereupon the computer program computes the power. If the user does not know how to compute the noncentrality parameter, he may invoke the analysis assistant, which will compute the noncentrality parameter from the user specified effect size, sample size, and number of rows and columns in the table. From the noncentrality parameter calculation, the power will be computed for a given alpha level.

Similarly, Sample Power calculates power and sample size for a variety of basic t, proportions, and crosstabulation tests. For a number of t-tests, it tests whether the t=0 or t=specific value. It does this for one sample t-tests, two sample t-tests with the same variance, two sample t-tests with different variances, and paired t-tests. The program computes power for a number of tests of proportions as well. The program will test whether a proportion equals 0.5 or a specific value. It will test 2x2 independent samples chi-square or Fisher's exact tests, as will Jerry Hintze's Power Analysis and Sample Size Program ((PASS 6.0)), distributed by NCSS. Paired proportions power is tested for McNemar's test. The sign test power can be calculated as well. A table comes up for each of the tests allowing the user to input specific sample sizes for different cells and then to request a graph of power as a function of number of cases for the given effect size, alpha level, and number of tails for the statistical test under consideration.

For correlation and regression analysis, Sample Power assesses power for different sample sizes. For correlation analysis, the power analysis program will test the power for specific sample sizes, alpha levels, one- or two-tailed tests for a correlation that is equal to zero, equal to a specific value, or equal to each other. For regression analysis, Sample Power has almost the least capability. It can handle multiple correlation with and without partialling. (PASS 6.0) is slightly better and can handle one or two set regression analysis, with the first set consisting of covariates. Sample Power, which has even greater flexibility, will handle one set of independent variables; a model with a set of covariates and a set of independent predictors; a model with a set of covariates, a set of independent variables, plus a set of interactions; a polynomial regression; and a model with covariates and dummy variables. The user merely has to indicate the number of independent variables, the r square of the respective set of independent variables, and the sample size, after which the program will compute the power for each set.

For ANOVA designs, Sample Power and (PASS 6.0) are very good for basic cross-sectional analysis. These two programs as well as (StatPower, the Statistical Design Analysis System, developed by James L. Bavry, Ph.D and distributed by Scientific Software, perform the one-, two-, and three-way fixed effects anova. Given the number of levels of each factor and the effect size tested, Sample Power will perform power and sample sizes for one-, two-, and three-way fixed effects ANOVA and ANCOVA models. (PASS 6.0) can perform the ancovas as multiple regressions, but does not have procedures dedicated to them. It can perform power and sample size analysis on randomized block anova and a repeated measures design with one between and one within subjects effect. The advanced psychology researcher might prefer (StatPower which performs power and sample size analyzes for one- and two-way fixed effects, one- and two-way random effects, and general fixed effects designs. What is more, (StatPower handles one and two-way repeated univariate and multivariate repeated measures (mixed model) designs as well. Sometimes, the biostatistician will have special need for (PASS 6.0), which excels in its power and sample size analysis for logistic regression and in its power analysis for the log-rank test performed with Life-Tables survival analysis. It also provides for plotting of power for different proportions resulting from matched case/control studies or from bioe-quivalence of proportions resulting from clinical research.

As for distributional tests, (StatPower by far surpasses the other two in variety and range. (PASS 6.0) cannot handle any wide variety of distributions at all, while Sample Power can handle a few basic distributional tests for t, F, and Chi-squares. By comparison, (StatPower yields probabilities for whatever level of cumulative, inverse, and noncentral z, t, Chi-square, and F distributions may be found. Moreover, it produces probabilities for binomial and beta cumulative and inverse distributions, as well as logistic and Poisson cumulative distributions.

5 Beware the Type III Error

Which of these programs the analyst prefers depends on his needs. For the beginning student, Sample Power is probably the most useful. For the researcher more interested in experimental design, (StatPower is probably the most useful. For the biostatistician, (PASS 6.0) may be the most useful. Professor Robert Lee of Pace University and former chairperson of the New York Chapter of the *American Association of Public Opinion Research* (AAPOR), cautions against what he calls the Type III error, ”the failure to ask the right questions in the first place.” In this light, this author notes that Sample Power has no, and (StatPower and (PASS 6.0) have only limited, longitudinal capability. In directions for future research and development, there will be a growing need for analysis of power and sample size of time series data with time series tests and sparse data with Exact tests. While Exact tests are robust to errors of the alpha level, they need a power and sample size analysis to indicate the magnitude of the problem of the Type II error. There needs to be more theoretical development and statistical package implementation in both of these areas. It is hoped that future research and development in this area will fill these gaps of knowledge and capability. For assistance with these matters please e-mail me at robert.yaffee@nyu.edu or phone me at the ACF Statistics and Social Science Group, 998-3402.

6 Further Reading:

Researchers interested in pursuing the study of power analysis may refer to R. Goldstein, ”Power and sample size via MS/PC-DOS computers” in *The American Statistician*, 43:253-260, 1989. For a comprehensive list of power analysis programs refer to Len Thomas and Charles Kreb, ”A Review of Statistical Power Analysis Software,” in the forthcoming *Bulletin of the Ecological Society of America*, 78 (2). MacCallum, R. C., Browne, M. W., Sugawara, H. M. (1996). ”Power analysis and determination of sample size for covariance structure modeling” *Psychological Methods*, 1, 130-149.