A brief reminder about Sample Size

As published in Benchmarks RSS Matters, March 2012

http://web3.unt.edu/benchmarks/issues/2012/03/rss-matters

Jon Starkweather, PhD

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

http://www.unt.edu

http://www.unt.edu/rss

RSS hosts a number of “Short Courses”.
A list of them is available at:
http://www.unt.edu/rss/Instructional.htm

A brief reminder about Sample Size

We've all heard (or spoken) questions similar to those below. How many voters should I poll to get an idea of who will win the election? What sample size do I need to determine whether people prefer green M&M's over red? How many undergraduates should I collect data from to determine if my model of retention is meaningful or predictive? How many people should I survey to measure satisfaction with my new product? How many mice should I assign to each condition of my experiment? How many protein samples should I extract from each person in order to create a composite protein estimate of each person? These are good questions. However, easy answers do not often follow good questions. The above questions all relate to the issue of sample size and much has been said on the subject. In this issue I'll provide some highlights for your consideration.

This paragraph contains information you likely are aware of, but (alas); I'm compelled by my professional conscience to type it. Generally it is suggested that questions of sample size be addressed prior to proposing a study (e.g. as a student; prior to thesis/dissertation proposal & as a faculty/professional researcher; prior to IRB and grant application). Typically during discussions of study design or methodology the issue of sample size should be addressed – because sample size is directly linked to statistical power and external validity. Post hoc power estimates are virtually useless. Generally, it is recommended that an a-priori power analysis be computed (using a desired level of power, desired effect size, desired error rate, and known/proposed number of parameters, variables, or conditions); which will produce a sample size estimate which in turn gives the researcher a target sample size which is likely to achieve the specified levels of power and effect size for a given error rate and design. We (RSS) like to recommend using G*Power 3 (which is a free download[1]) or any one of several R packages designed for this task. In conducting a-priori power analysis, it is important to remember what statistical power actually is: the ability to detect an effect if one exists (in formula: power = 1 - ?). Or, if you prefer, as Cohen (1988) put it: "the power of a statistical test is the probability that it will yield statistically significant results" (p. 1).

The most general, and flippant, guideline for sample sizes often tossed around is "you need to have more cases/participants than you have parameters/variables/questions." The next most stringent phrase you are likely to hear, often associated with a 'step' from descriptive statistics to inferential statistics, is "you need to have at least 5 to 10 cases/participants for each parameter/variable/question." Next, often associated with a 'step' from fairly straightforward inferential techniques (t-test, ANOVA, linear [OLS] regression...) to multivariate statistical techniques is "you need at least 25 (up to 150) cases/participants for each parameter/variable/question." These types of heuristics, although they make nice quick sound-bite answers, are not terribly useful because; real consideration must be taken with respect to a variety of issues. The first issue to consider is the statistical perspective one is planning on taking with the data, will a Bayesian perspective be used or a Frequentist perspective. Generally speaking, Bayesian analyses handle small sample sizes better than analogous Frequentist analyses, largely because of the incorporation of a prior. A Bayesian perspective also allows one to use sequential testing; implementation of a stopping rule (Goodman, 1999a; Goodman, 1999b; Cornfield, 1966). Other considerations include, what types of hypothesis (-es) one is attempting to test, what type of phenomena is being statistically modeled, the size of the population one is sampling from (as well as its diversity), and (certainly not

[1] `http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/`

least) the type of analysis one expects to conduct. Some analyses inherently have more power than others (e.g., see discriminant function analysis vs. multinomial logistic regression). Furthermore, one must consider the assumptions of the analysis one is expecting to run. Often data collected does not conform to the assumptions of a proposed analysis and therefore, an alternative analysis must be chosen - one which will provide analogous statistics for addressing the hypothesis or research question posed; but, the alternative often has less power. Another consideration is this; it is well accepted that point estimates (e.g., mean, median, model parameters; such as regression coefficients) are fairly stable and fairly accurate even with relatively small sample sizes. The problem (again, well accepted) is that interval estimates (e.g., confidence intervals) will not be terribly accurate with small samples; often the standard errors will be biased. The only real answer is; larger samples are better than smaller samples...

Contrary to much of the above considerations; some modern methods (e.g., optimal scaling, resampling) can be used to overcome some of the pitfalls of a small sample. However, many people are suspicious of these modern methods and they can be quite controversial (e.g. if a journal editor or reviewer has never heard of optimal scaling, how likely do you think you are to get the study published in their journal?). These methods are genuinely controversial because they often assume a particular position or belief about something – for instance, people who use optimal scaling with survey data have particular beliefs about the characteristics and properties of survey measurement; which others, of equal professional respect, disagree with or hold opposing beliefs.

Lastly, with respect to sample size, using new measures/instruments (ones which have not been validated nor had their psychometric properties established/accepted) should motivate the collection of large samples. The larger sample can be divided into 2 or more subsamples so one subsample can be used for validation or confirmatory analysis, while the other subsample(s) can be used to fit the hypothesized models.

We (RSS) have a rule that the study author(s) or primary investigator(s) should be the one(s) to make decisions regarding what is done and we want those decisions to be as informed as possible by providing as much (often called too much) information as we can. Therefore, we will not provide 'easy' answers to questions of sample size. The amount of data collected for any empirical study should be based on critical thought, on the part of the study authors, directed toward the considerations mentioned in this article. The best two pieces of advice on the subject of sample size are; start to think about sample size very early (i.e. long before data collection begins) and collect as much data as you possibly can.

Until next time, don't play *The Lottery* with Shirley Jackson

References & Resources

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

Cornfield, J. (1966). A Bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association, 61*, 577 - 594. Available at JSTOR: http://www.jstor.org/stable/10.2307/2282772

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28*, 1-11. Available at: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/literature

Goodman, S. (1999a). Toward evidence-based medical statistics, 1: The p value fallacy. *Annals of Internal Medicine, 130(12)*, 995 - 1004. Available at: http://psg-mac43.ucsf.edu/ticr/syllabus/courses/4/2003/11/13/Lecture/read

Goodman, S. (1999b). Toward evidence-based medical statistics, 2: The Bayes factor. *Annals of Internal Medicine, 130(12)*, 1005 - 1013. Available at: http://psg-mac43.ucsf.edu/ticr/syllabus/courses/4/2003/11/13/Lecture/read

Herrington, R. (2002). Controlling False Discovery Rate in Multiple Hypothesis Testing. http://www.unt.edu/benchmarks/archives/2002/april02/rss.htm

Herrington, R. (2001). The Calculation of Statistical Power Using the Percentile Bootstrap and Robust Estimation. http://www.unt.edu/benchmarks/archives/2001/september01/rss.htm

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). London: Oxford University Press.

Price, P. (2000). The 2000 American Psychological Society Meeting. http://www.unt.edu/benchmar

This article was last updated on March 12, 2012.

This document was created using LaTeX