JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 3, No. 3, March 2007 “Can You Handle the Truth?”

How Many Are Enough? Statistical Power Analysis and Sample Size Estimation in Clinical Research

By Janet Houser

One of the most common questions asked of statisticians is, “How many are enough?” The answer to that question, unfortunately, is complex, and only part of it can be answered by the statistician. This is because power calculation is part mathematical calculation and part judgment call. The former requires knowledge about population characteristics; the latter requires choices about tolerable error and meaningful effect size. The two come together to assure that the sample is sufficient to detect a phenomenon of interest and provide the reader with confidence in the results.

The researcher controls a study’s power primarily by assuring that an adequate number of subjects are available to draw a definitive conclusion about the effects of a treatment. It would seem logical, then, to draw as large a sample as possible. On the other hand, adding subjects increases power, but also the duration and cost of the study. Samples can actually be too large, in which case treatment effects are statistically significant but of such small magnitude as to hold no clinical significance. So it is worth the time and effort to assure the sample size is large enough to detect differences, reasonable enough to be feasible, and small enough to be efficient. This requires a formal process of power analysis before the study begins, with a collaboration between researcher and statistician. Given that every study is different, and the specifics of a study sample cannot be known in advance, how do statisticians determine the number of study subjects required to achieve the necessary power for a study?

The Relationship of Power and Sample Size

Sample size is the primary way that a researcher assures a study has adequate power. Power is necessary to detect outcomes in experiments; it is the basis for assuring accurate conclusions about the effects of treatments. With insufficient power, a researcher may draw a conclusion that a treatment was ineffective – because no outcomes were detected – when, in fact, the treatment was effective but the size of the sample was too small to discover it. This is called a Type II error, and is an error of lost opportunity in healthcare. A treatment has the potential to be effective, but its benefits are unrealized.

The calculation of power is a mathematical process. It may be calculated prospectively (in advance, to determine how many subjects should be recruited) or retrospectively (after the fact) to determine how much power a sample possessed. In general, prospective power analysis is superior because it assures the investigator that he/she is recruiting enough subjects to make the study worthwhile.[1] However, accurate prospective power analysis requires that the researcher have good information about the variables and sample, either from previous reports or pilot studies.

Oftentimes, though, historical data is unavailable and pilot studies are not feasible. In these cases, researchers often use “rules of thumb” or best guesses to determine sample size and forge ahead with a study. If statistically significant results are found, then power is not an issue. If an outcome was, indeed, detected, then the sample was obviously large enough to detect it. On the other hand, if no statistically significant findings are reported, then a retrospective power analysis is called for. While determining in hindsight that a study had inadequate power does not help the researcher improve on the results, it does at least explain a lack of effect when it would otherwise be expected, so that there remains the potential that the treatment works.

An example of the appropriate use of power analysis appears in a study by Thomas (2005) comparing the effects of combined spinal-epidural technique with traditional epidural anesthesia technique in laboring mothers.[2] The primary outcome measure was the rate of manipulation required of the epidural catheter (the frequency with which the anesthesiologists had to reposition the catheter). The researchers set out to determine if there is a difference in the incidence rate of epidural manipulation in laboring mothers who receive combined spinal-epidural technique versus mothers who receive the traditional technique. A Type II error would have been committed if the authors concluded there was no difference in the outcomes of the two techniques – failing to reject the null hypothesis – when in fact there was. Because this study did, indeed, find there was no difference, a Type II error could be suspected if the authors had not conducted a prospective power analysis and achieved an adequate sample size.

The Calculation of Power

Power is calculated based on several elements. Some are judgments made by the researcher; others are calculated from a sample. These elements include:

- The statistical test that is planned
- The effect size that is detectable
- The acceptable level of power
- Specific characteristics of the sample

Different statistical tests require different methods of calculating power because the sensitivity of the tests varies. Power is the capacity to detect a difference; if the test being used is not very sensitive, then a larger sample will be needed to obtain adequate power. Tests in which the data meet the assumptions of a normal distribution (bell curve) generally are quite robust, meaning they can detect differences with high sensitivity, so their sample sizes may be relatively small. Because tests for non-normal data (e.g. Chi Square or the Mann-Whitney U) are relatively insensitive, they require larger samples. Multivariate tests often require larger samples because they are looking for multiple relationships and inter-relationships, and so have to be sensitive enough to detect them all.

In the Thomas study of epidural anesthesia, the outcome of interest was an incidence rate. Rates are not generally normally distributed until they reach very large sample sizes. The author, then, had to use a power analysis formula that is specific to non-normal data. The resulting sample size requirement was therefore larger than would be required for data expressed as mean averages or for continuous measures.

Effect size is a key consideration in power calculation. Just as a large object is easier to see than a small one, big effects are easier to detect in the data than small effects. To calculate power, the researcher must predetermine the size of effect considered clinically significant. A new drug that cures only 1% of the patients will never leave the pharmacy shelf, but one that cures 80% will fly off the shelves. The question of clinical significance becomes more interesting if a cure rate of only 10%, 20% or 30% is clinically significant. Large effects can be detected with small samples, while subtle effects require large samples. Detecting rare side effects requires very large samples.

Determining a meaningful effect size is a judgment call, but clinical expertise, historical studies, and patient responses may all be considered. In the Thomas study, the researcher used historical performance and consultation with experts to determine that a 15%

reduction in the epidural manipulation incidence rate would be clinically significant for the patients and a meaningful reduction in the workload of the anesthesiologists.

The acceptable level of power is also set by the researcher. Because power reflects Type II error, the researcher wants to minimize this potential problem. 80% power is considered the minimally acceptable level, while some researchers require much higher levels of power for a study.[3] Selecting a desirable power level is achieved by balancing the need to detect an outcome with the difficulty in obtaining large sample sizes.

These elements of power calculation – the specific statistical test, meaningful effect size, and level of power required – are determined by the researcher prior to the beginning of the study. Researchers normally report these elements in the methods section of an article with the discussion of sample size and power analysis.

Other elements of power calculation, though, are not selected but rather estimated by considering characteristics of the expected population. In general, a highly heterogeneous population is not represented well by a small sample, and so will have less power. Heterogeneity can be assessed statistically via the standard deviation of key variables. If prior measurement of outcome variables shows a great deal of variability as represented by a relatively large standard deviation, then larger samples will be needed. The chances are greater in a diverse population that some group will be under- or over-represented; larger samples can help ensure this does not happen. To establish population characteristics with any certainty, historical data or data from similar studies is required. Unfortunately, these figures are often unavailable for early-stage studies, as in the case of the Thomas study, in which historical data were unavailable.

When no historical data are available, accurate power analysis is problematic. The most common way to obtain data is to conduct a pilot study. There are many other advantages of conducting pilot studies prior to full-scale randomized trials: assessments can be tested, protocols refined, and procedures practiced relatively quickly and inexpensively with small numbers of subjects. However, if a pilot is not possible and no historical data available, the researcher can analyze data from the initial subjects to determine with limited certainty whether the sample size should be modified.

The Thomas study reported all of these considerations in the documentation of the prospective power analysis: "108 patients per group were required to detect a reduction in the epidural catheter manipulation rate from 32% to 15% with a power of 0.80 and an alpha of 0.05….[after a] previous study found a 31.4% overall incidence of epidural manipulation rate at out institution." The researcher has provided the reader with enough information to judge the trustworthiness of the power analysis process.

Effects on Power

Characteristics of the sample have a direct effect on power; highly diverse samples will require adjustments in sample size. But other elements of an experiment also affect power.

Adequate power is hard to achieve when results must be very accurate. Very high confidence levels, e.g. 99% or 99.9%, require very large samples or very focused research questions. Multiple subgroups also increase demands for power because each group is, in essence, its own clinical trial, albeit with the offsetting advantage of less heterogeneity within the subgroup.

On the other hand, some designs improve power. Research designs that involve repeated measures have more power. These designs have dependent samples; in other words, the same subjects are measured multiple times. This design increases the number of observations without an associated increase in subject recruitment.

When all of these elements have been considered and the critical decisions about error and effect size have been made, then an actual calculation is made. Researchers and statisticians can use the free, publicly available software in Table 1 to conduct the calculations. Alternatively, power can be calculated manually. Formulae and power tables can be found in most detailed research texts, such as the 2001 book by Hulley et al., "Designing Clinical Research." [4]

Table 1. Free Internet Sites for Calculating Power

http://calculators.stat.ucla.edu/powercalc/
A good site that requires the investigator to specify the parameters that are expected based on prior literature, and to have made decisions about level of statistical significance, required power, and directionality of the hypothesis. Can calculate power either prospectively or retrospectively.

http://www.cs.uiowa.edu/~rlenth/Power/#Download_to_run_locally
A series of downloadable Java applets that supports calculation of power for a variety of tests but is a bit complicated to use.

http://statpages.org/#Power
A directory of web-based power calculators that can handle over 50 types of calculations, including many that are uncommon.

When prior data are not available, there are other, less effective ways to estimate sample size. Rules of thumb can be applied. It is generally considered sufficient to have 15 subjects per tested variable, although some researchers estimate this number to be as high as 50. In general, samples with fewer than 30 subjects are not considered powerful enough to detect changes in an outcome variable. On the other hand, samples with more than 200 subjects generate only marginal improvements in power.

If the researcher has unlimited time and money and an accessible population, larger samples are almost always advantageous. The one statistical caveat is that as samples get larger and the results more precise, it becomes easier to detect inconsequential, clinically insignificant differences. For example, it is nice to know that an anti-hypertensive drug reliably lowers blood pressure by 0.1%, but it is not very useful for patients.

A study with insufficient power may lead the researcher to abandon a potentially useful treatment. Power analysis is the best method to avoid these serious errors.

References

1. Houlse, T., Penzien, D., & Houlse, C. 2005. Statistical power and sample size estimation for headache research: an overview and power calculation tools. Headache. 45:414-8.
2. Thomas, J. 2005. Dural puncture with a 27-guage Whitacre needle as part of a combined spinal-epidural technique does not improve labor epidural catheter function. Anesthesiology. 103(5):1046-51.
3. Portney, L. & Watkins, M. 2000. Foundations of Clinical Research: Applications to Practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall Health.
4. Hulley, S., Cummings, W., Browner, W., Grady, D., Hearst, N. & Newman, T. 2001 Designing Clinical Research, 2nd ed. Lippincott, Williams and Wilkins: Philadelphia.

Author

Janet Houser, Ph.D., is Associate Professor of Health Services Administration at Regis University. Contact her at 1.303.458.4061 or jhouser1392@excite.com.