Sample Size and Statistical Power

Determining Sample Sizes for Epidemiologic Studies

by Colleen Kaelin

This is the third article in our series in *Kentucky Epidemiological Notes and Reports* discussing biostatistical and epidemiologic research topics. The series is intended to provide biostatistical reviews for readers who do not often use statistics in their everyday practice and to generate discussion about biostatistical topics as they relate to areas of interest for those who commonly use these methods. In this issue, our topic is sample size and statistical power. We will describe the factors that affect statistical power, and discuss the methods used to calculate the sample size required to achieve the desired level of power.

One of the primary considerations in designing an epidemiologic study is determining how many subjects or how large a sample size is needed. A study that does not have enough subjects will not provide a satisfactory answer to the question posed by the researcher. In statistical terms, the power of a study is the ability of the study to demonstrate an association between an exposure and an outcome, given that the association exists *(Elwood, Mark. 2007. Critical Appraisal of Epidemiological Studies and Clinical Trials, Oxford Medical Publications, pg 258).* A study with greater statistical power will show an association more effectively than a study with less power. Even though larger studies generally have more power than smaller studies, there are other factors that may affect the sample size needed to achieve the minimum acceptable level of statistical power. We will discuss each of these factors in detail.

The first and most obvious factor affecting the statistical power and size of the study is the *strength of the association* between the exposure and the outcome in question. It makes both statistical and common sense that a stronger association requires a smaller number of subjects to demonstrate it than a weaker association. The typical measure of the strength of an association is the *risk ratio.* This ratio is the comparison between the risk of an outcome in the exposed group and the risk in the unexposed group, or the risk of an exposure in a group of cases versus a control group. This ratio can be directly obtained in cohort studies and clinical trials. In case-control studies, the odds ratio is used as an estimate of the risk ratio if the incidence of the outcome is more than 10 percent. The larger the difference in risk between the two groups, the easier it will be to detect *(Elwood, pg. 257).*

The second factor influencing the size of a study is *the frequency of the outcome or exposure under study.* The sample size needed to achieve the desired statistical power decreases as the frequency of the outcome or exposure approaches 50 percent. In a cohort study, the main consideration is the outcome in question, whereas in a case-control study, the primary consideration is the exposure *(Elwood, pg. 257).*

Another consideration affecting the sample size is the *significance level* set by the researcher as the "cut-off point" to determine if an association can be regarded as statistically significant. The significance level is also referred to by the Greek letter alpha (ά) and is usually set at the 0.05 or 95 percent level. This level indicates that there is a 5 percent chance the study will show an association where none exists. As the significance level decreases, the sample size necessary to achieve an acceptable answer to the question under study increases. The *power* of a study is designated by the Greek letter beta (β), and is a measure of how often a study will fail to show an association where an association actually does exist. The typical value of power designated by researchers is 80 percent *(Elwood, pgs. 257-258).*

There are several studies in which it is impractical to have all study groups be an equal size. As an example, a study published in the *Journal of the American Medical Association* (JAMA) compared the use of mental health services among veterans deployed to Iraq versus those deployed to Afghanistan versus other locations. The results were based on the Post-Deployment Health Assessment (PDHA), which was mandated for all military personnel returning home from any deployment. The study included 222,620 veterans who participated in Operation Iraqi Freedom; 16,318 veterans who participated in Operation Enduring Freedom in Afghanistan; and 64,967 veterans who participated in other military operations. Of the surveys analyzed, 73.3 percent were from veterans who had served in Iraq, 21.4 percent were from veterans who had served in other locations, such as Kosovo and Bosnia, while only 5.4 percent were from veterans who had served in Afghanistan. The authors concluded that the prevalence rate of mental health problems was consistently higher among veterans deployed to Iraq (19.1 percent) than among veterans deployed to Afghanistan (11.3 percent) and other locations (8.5 percent). They also reported that veterans of Iraq used inpatient and outpatient mental health services at higher rates after deployment and were significantly more

likely to leave military service than the veterans from the other groups. But what effect does the disparity in the number of subjects in the three groups in the study have on the power of the study? Using the statistical power function of the Open Epi program, which we will discuss in detail later, the risk of mental health problems for veterans serving in Iraq was 1.7 times greater than for veterans serving in Afghanistan. The statistical power of the study, according to the program, was 100 percent. When comparing Iraq veterans with the veterans serving in other locations besides Afghanistan, the risk ratio for mental health problems increases to 2.2, but the statistical power remains at 100 percent. Comparing veterans of Afghanistan with veterans of all locations besides Iraq produces a risk ratio of 1.3, while the statistical power remains at 100 percent. It seems that with such a large sample size, the statistical power of the study was unaffected by the disparity in the size of the three groups. *(Hoge, Charles W. M.D., et. al. Mental Health Problems, Use of Mental Health Services, and Attrition from Military Service After Returning from Deployment to Iraq or Afghanistan; Journal of the American Medical Association, March 1, 2006. vol. 295; no. 9; pgs 1023-1032.)*

Another study of mental health and military service compared a new therapist-assisted, Internet-based, self-management cognitive behavior therapy versus Internet-based supportive counseling to see which treatment was more effective. Subjects included service members who exhibited symptoms of Post-Traumatic Stress Disorder (PTSD) after the September 11, 2001, attack on the Pentagon, as well as Iraq or Afghanistan veterans. Twenty-four subjects were assigned to self management cognitive behavior therapy, while twenty subjects were assigned to supportive counseling. The authors concluded that self-management cognitive behavior therapy led to greater reductions in PTSD, depression, and anxiety scores at 6 months evaluation. How much statistical power does this study have, given the small number of subjects?

As you can see from the results produced by the Open Epi Web site, the statistical power of the study at a 95 percent confidence level, based on the post-treatment scores for total Post-Traumatic Stress Disorder symptoms, is only a little more than 28 percent, not nearly the 80 percent most researchers aim for in their initial study design. By using the Open Epi website, we can estimate that it would take at least 93 individuals in each sample group to achieve the desired statistical power for this study at a 95 percent confidence interval. *(Litz, Brett T., PhD. et. al. A Randomized, Controlled Proof-Of-Concept Trial of an Internet-Based, Therapist-Assisted Self-Management Treatment for Posttraumatic Stress Disorder. American Journal of Psychiatry; 164:11, November 2007. pgs. 1676-1683).*

Input Data

Two-sided Confidence Interval	95%		
	Group 1	Group 2	Mean Difference [1]
Mean PTSD post-treatment Test score	14.86	20.00	-5.14
Sample size	24	21	
Standard deviation	13.35	11.50	
Variance	178.223	132.25	

Power based on Normal approximation method	28.36%

[1] Mean difference= (Group 1 mean) - (Group 2 mean)

Results from OpenEpi, Version 2, open source calculator--PowerMean

In addition to the factors mentioned above, *controlling for confounders* may affect the size of the study sample. The issue of confounding has been addressed in the two previous biostatistical research topic articles, but now we will address the effect of controlling for confounders on sample size. Some methods for controlling are used in the design phase of a study, some in the analysis phase, and other methods can be used in both phases. One of the methods used in the analysis phase is *stratification,* which is a term for dividing the sample group into separate classifications based on some possible confounding factor. For example, if we were studying the effect of age on the risk of lung cancer, it would make sense to stratify the sample into smokers and non-smokers, to ensure that the effect of smoking was accounted for and did not obscure the risk we were trying to examine. Many studies stratify the results by age, gender and other factors that are known to affect the risk of several common diseases. The result is that each category or subgroup becomes smaller as the study is stratified by more and more specific potential confounders. Therefore, a researcher who intends to stratify the sample group in analysis to control for potential confounders will need to increase the size of the study. The only method of controlling for confounders that could decrease the size of a study is if the cases and controls are individually matched to each other by the suspected confounder *(Elwood, pg. 257).*

Factors That Decrease Sample Size	Factors That Increase Statistical Power
Lower desired statistical power	Larger sample size
Larger meaningful difference (effect size)	Larger meaningful difference (effect size)
Smaller standard deviation	Smaller standard deviation
Less stringent significance criterion	Less stringent significance criterion

Sample Size Estimation: A Glimpse Beyond Simple Formulas, John Eng, M.D. (*Radiology* 2004; 230:606-612.)

Formulae to Estimate Sample Size

For Comparative Studies

When the outcome variable of a comparative study is a continuous value for which means are compared, the appropriate sample size is given by:

$$N = \frac{4\sigma^2(z_{crit} + z_{pwr})^2}{D^2}, \qquad (1)$$

where N is the total sample size (i.e., the total of the two comparison groups), D is the smallest meaningful difference between the two means being compared, σ is the Standard Deviation (SD) of each group, and z_{crit} and z_{pwr} are constants determined by the specified significance criterion and desired statistical power respectively. Since z_{crit} and z_{pwr} are independent of the properties of the data, sample size depends only on the ratio between the smallest meaningful difference and the SD.

Sample Size Estimation: A Glimpse Beyond Simple Formulas, John Eng, M.D. (*Radiology* 2004; 230:606-612.)

TABLE 1
Standard Normal Deviate (z_{crit}) Corresponding to Selected Significance Criteria and CIs

Significance Criterion*	z_{crit} Value†
.01 (99)	2.576
.02 (98)	2.326
.05 (95)	1.960
.10 (90)	1.645

* Numbers in parentheses are the probabilities (expressed as a percentage) associated with the corresponding CIs. Confidence probability is the probability associated with the corresponding CI.
† A stricter (smaller) significance criterion is associated with a larger z_{crit} value. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula z_{crit} = NORMSINV(1_(P/2)), where P is the significance criterion.

TABLE 2
Standard Normal Deviate (z_{pwr}) Corresponding to Selected Statistical Powers

Statistical Power	z_{pwr} Value*
.80	0.842

.85	1.036
.90	1.282
.95	1.645

* A higher power is associated with a larger value for z_{pwr}. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula z_{pwr} = NORMSINV(*power*). For calculating power, the inverse formula is *power*=_NORMSDIST(z_{pwr}), where z_{pwr} is calculated from Equation (1) or Equation (2) by solving for z_{pwr}.

Sample Size Estimation: How Many Individuals Should Be Studied? John Eng, M.D. (*Radiology* 2003; 227:309–313.)

Formula for Cohort or Trial Design

$$N = \frac{(p1q1 + p2q2) \times K}{(p1 - p2)(p1 - p2)}$$

Where
p1 = frequency of outcome in group 1
q1 = 1-p1

And
p2 = frequency of outcome in group 2
q2 = 1-p2

And K = (z_{crit} + z_{pwr}) squared

(*Elwood, pg. 261*)

Use the same formula for case control studies, where p1 = frequency of exposure in the case group, and p2 = frequency of exposure in controls.

Computer Programs and Web Sites to Determine Sample Size Estimation

EPI INFO
In the Epi Info home page choose Utilities from the top toolbar. Choose Statcalc, then Sample Size and Power. Select from Population Survey, Cohort / Cross Sectional, or Unmatched Case Control. Enter the necessary information and the program will compute the result.

OPEN EPI
Go to the Open Epi Web site.

From the menu on the left side, expand the Sample Size folder or the Power folder.

Select from Proportion, Unmatched Case-Control, Cohort/Randomized Clinical Trial, Mean Difference, or Cross Sectional. Click on "Enter" and follow the instructions to calculate the result.

If you wish to calculate the statistical power of a study, choose from the options in the Power folder underneath the Sample Size folder.

It is always advisable to consult a biostatistician during the design phase of a study to help determine what formulae and programs may be useful to determine an appropriate sample size.

About the Author

Colleen Kaelin, MSPH, is the staff epidemiologist for the Kentucky Department for Public Health's Division of Public Protection and Safety.

Last Updated 9/8/2009