By Hollis N. Erb

A Non-Statistical Approach for Calculating the Optimum Number of Animals Needed in Research

Hollis Erb is a Professor of Epidemiology at the Department of Clinical Sciences, New York State College of Veterinary Medicine. Send reprint requests to her at NYSCVM, Cornell University, Ithaca, NY 14853.

Reprinted from *ILAR News;* 32(1), Winter 1990, with permission.

It is unethical to use either too many or too few animals in research. Using many more animals than needed is an abuse of our privilege to use animals and a waste of research moneys. It also is wrong to use so few animals that either the data are unreliable or the smallest important effect goes unnoticed; in this case, both the animals and moneys are wasted. Calculating the optimum number of animals needed for a study using statistics can prevent this breach of ethics.

Since many people find statistics confounding, I do not present formulas for calculating sample size in this article. Most research communities employ a statistician to perform this function. Instead, this article provides a practical understanding of how to make decisions about what information to convey to the statistician who calculates the sample.

In order to do so, I will explain the issues underlying sample size calculations, including actual elements of sample size calculations, assessment of the kind of data, whether the purpose of analysis is description or comparison, common sample size mistakes, and ways to decrease the sample size.

Actual Elements of the Sample Size Calculations

The elements include such things as the alpha and beta errors, the smallest difference or effect that is worth detecting, the baseline or control group's count or measurement, and the typical variation that would be seen. They appear in the sample size formulas, and also are needed to find the right spot in a sample size table or nomogram.

Willingness to Make a False Positive or False Negative Statement

The alpha (type I; false positive) error is the chance that the investigator is wrong when he or she says there is a difference, effect, association, dependence, or correlation. This error is made when one wrongly concludes significance. A wrong conclusion could be made because the data are only a sample and, therefore, subject to random variation or because of measurement error in the data. The probability of this error occurring is the level of significance, or the *P* value. In contrast, a beta error (type II; false negative) occurs when one wrongly concludes non-significance and there really is a treatment effect, difference, or association. The beta error is especially important because it also indicates the power of the study. The power (power=1-beta) is the probability that a difference that truly exists will be detected by the study; larger beta means lower power.

One never can guarantee completely against alpha and beta errors, but their chances can be reduced with larger sample sizes. Alpha often is set at 0.05 or at 0.01- at a 5 or 1 percent level of significance—but this is arbitrary, and other alpha levels can be used. However, levels >10 percent will tend to alarm reviewers if one finds significance, as will levels <0.1 percent if one does not find significance. Some people have a rule of thumb that beta should be set to equal 2 alpha or 4 alpha, but this is nonsense. Beta should not be linked automatically and arbitrarily to alpha; beta should be set in its own right according to the situation under study. Beta should be smaller than alpha if committing a beta error is worse (is more costly; will cause more problems) than committing an alpha error. Beta should be greater than alpha if an alpha error would be worse than a beta error. If it is impossible to decide which error is worse, then set alpha roughly equal to beta.

The smaller that alpha and beta are, the greater the number of animals needed. Additional animals are the price paid for increased confidence that an error is unlikely. Alpha also must be specified for a descriptive study, if one wishes to calculate a confidence interval for the value being estimated. The formulas for confidence intervals include alpha. The interpretation

of alpha here is the probability that a large number of confidence intervals, formed from random samples similar to the one under study, would exclude the true population value being estimated.

Smallest Difference Worth Detecting

The smallest difference that is worth detecting must be specified. Clearly, any larger difference also would be worthwhile. Although phrased as a difference between groups, the term also could be thought of as the smallest worthwhile effect, correlation, change, odds ratio, and so on. The smallest difference must be large enough to make the research useful; this is very much a "why bother?" specification. In many instances, the worthwhile difference will be subjective ("We probably could convince people to substitute this drug as long as it was xx percent more effective."). However, it may be possible to quantify most of the costs associated with an intervention, especially in, for instance, food animal research. The new intervention will have to at least pay off these costs to be useful, so a starting place for setting the smallest worthwhile difference would be the difference that produces a benefit equal to the costs.

The smaller the smallest worthwhile difference is, the harder it will be to detect—or rule out—and the larger the sample size must be. This is intuitively obvious; it is easier to be certain that a big difference does not exist than it is to be certain that a small difference does not exist.

If the sample size is considerably larger than that needed to find the smallest worthwhile difference, then the extra animals were wasted. However, useful information still will have been obtained, whether or not significance was found. If the sample size is considerably smaller than that needed to find the smallest worthwhile difference and the results were nonsignificant, then all the animals were wasted because there was no useful information obtained.

This situation may occur more often than we realize. Frieman *et al.*[1] reviewed 71 clinical trials involving human patients; these trials had nonsignificant results and were published in reputable journals. The investigators did back-calculations, using the numbers in the 71 articles, and found that 67 out of 71 trials had insufficient patients to detect a 25 percent therapeutic benefit, and that 50 of 71 could not detect at 50 percent benefit. Although we cannot know for each trial what the exact costs and benefits were, it is unlikely that a 50 percent therapeutic improvement would not have been worthwhile in most of the trials. The implication is that most of the reviewed trials were a waste of time and money, and put patients at risk without justification.

Baseline or Control Value for the Outcome

The baseline or control value often will have to be specified in both descriptive and comparative studies. This value is the best guess of what the mean, correlation, or proportion will be in the descriptive study and the value expected in the baseline period or in the control group in the comparative study. The best guess might be based on pilot studies, a literature review, or expert opinion. The control value is important because the smallest difference worth detecting is tested for in relationship to the control value.

Typical Variation

There is always some variation or imprecision associated with counting or measuring. This variation can mask true differences between groups. The larger the ratio between the smallest difference worth detecting and the standard deviation of the outcome, the easier it will be to see that difference and the fewer the animals that will be needed. As the ratio gets smaller, one adds animals to get a "tighter" estimate of the effect; this decreases the standard deviation and makes the ratio (worthwhile difference to the standard deviation) larger again. If, however, the ratio gets smaller and one doesn't increase the sample size, the beta error increases, and there is a greater risk of missing a true difference.

Issues That Determine Which Sample Size Formula to Use

Type of Data

The type of data and the objectives of the research determine, in general terms, the statistical methods that should be used. Different statistical methods imply different methods for sample size estimations.

Data are either discrete (count type) or continuous (measurement type). Nominal discrete data are simply names without any intrinsic ordering, such as breeds, counties, or genders. Ordinal discrete data have intrinsic ordering and are common in scoring systems (*e.g.*, "-,+,++,+++" fluorescent antibody scores, or 0, 1, 2, 3 for "none, slight, moderate, severe" illness).

Discrete data can be counted and displayed as frequency distributions. Even if scores appear to be numeric, however, the numbers are just codes, and one cannot do arithmetic on them. Therefore, medians and modes can be calculated, but means are invalid, because the mean will change if the code is changed, even if the order in scheme is not changed. Nominal and ordinal data typically will be analyzed by methods such as chi-square tests, rank tests, or nonparametric correlations.

If data are not discrete, they are continuous. Height, age, daily milk yield, and rate of gain are examples of continuous data. With continuous data, one must decide whether or not they are Gaussian (normal; bell-shaped curved).

Gaussian data follow a bell-shaped curve when frequency is plotted against measurement. The bell-shaped curve is symmetric around the high point in the middle, which is the mean value. Gaussian data are the only data for which the mean and standard deviation (SD) are appropriate. The "mean

+\- 1, 2, or 3 SD" covers roughly the middle 67, 95, or 99 percent of the area under the bell-shaped curve—this works only with a curve that is symmetric with a particular shape around a central peak.

If various descriptive statistics are calculated, one can, in determining whether the data are Gaussian, make use of the fact that the bell-shaped curve is symmetric around a single peak. Because of this shape, for Gaussian data the mean, the median, and the mode will be similar, and complementary percentiles will be roughly equidistant from the median (*e.g.*, the difference between the values of the 25th and 50th percentiles will equal the difference between the values of the 50th and 75th percentiles). If the data do not at least roughly fit these rules of thumb, then the data probably are not Gaussian. As another alternative, one always can inspect either a graph of the frequencies plotted against the measurements or a histogram of the frequencies against ordered categories of the measurements to see if the graph looks like a bell-shaped curve. Finally, common sense often warns that a variable is unlikely to be Gaussian. For instance, one would expect the distribution of parities in a dairy herd would to look like a descending curve, with lots of first-parity heifers and few aged cows; distributions of liver enzymes in diseased animals often have very long tails on the right-hand (high) side.

If the continuous data are not Gaussian, then there are two options for data analysis. The first option—the one I prefer unless I need to do a multivariable analysis—is to use nonparametric methods (*e.g.*, rank sum instead of t-tests; Kendall's or Spearman's correlation instead of Pearson's correlation). The other option is to transform the data. Some transformations are fairly standard as "first-tries"—such as trying logs on data with a long rail to the right or square roots on data with a long tail to the left—but others may require help from a statistician.

There is an additional, special-case type of data: very large, essentially unbounded counts, such as numbers of somatic cells or bacteria in a ml of milk, or a red blood cell count on a hemogram. Although technically these are count-type (discrete) data, in practice the measurements take so many different values across such a large range that such data are treated as if they were continuous.

Description or Comparison?

The purpose of the analysis is description if one wants to know, for example, the proportion dying or the typical litter size. The purpose is comparison if one tests whether or not two—or more—values are equal.

If the purpose is description, then one needs to decide how precise the description must be to be useful. For discrete data, such as the proportion dying, this might mean specifying the maximum acceptable width of the confidence interval for the proportion. For continuous data, one might need to specify a maximum acceptable coefficient of variation, or an SD no larger than a certain number of units. The more precise the description must be, the greater the number of animals necessary.

If the purpose is comparison, rather than mere description, then there are at least three more questions to answer. One is: how many different groups will there be? The groups might be different breeds or different treatments. The more groups there are in any comparison, the greater the number of animals needed. This is because no matter what the number of groups, one will have to count or measure the outcome in each group with some reasonable precision in order to determine whether the groups differ.

A second question is: will the groups be of equal size? Most tables and sample size formulas assume equal sample sizes, partly because equal size is mathematically simpler. In fact, precision varies only with the square root of the sample size, so that it takes very large increases in sample size to make important changes in precision. A second and perhaps more compelling reason for equal sample sizes is that the investigator is supposed to behave as if he or she has no prior preference, prejudice, or bias for or against any particular treatment. Given this neutral stance, the sample sizes should be equal—or close enough to be within the range of slight imbalances that occur with simple randomization.

In some cases, however, unequal sample sizes may be justified. If there are large differences in treatment costs, one could increase the number of animals receiving the cheaper treatment and decrease the number receiving the costlier treatment. Or, if there are large differences in expected losses to follow-up, one could increase the number of animals in the group that might have higher losses (without decreasing animals in the other groups). Even so, Peto *et al.*[2] argued that the allocation ratio should vary by no more than 2:1. If one wishes to adjust sample sizes by costs, then a statistician probably is needed. If one wants to adjust for unequal losses to follow-up, it might be good enough to calculate the sample sizes needed if there were equal group sizes and no losses to follow-up; then, the sample size for each group can be inflated in proportion to the expected losses. For example, 25-33 percent of all dairy cows are culled each lactation; if information is needed from two successive lactations on each of 100 cows, then start with 150 animals. In general, expect to see losses to follow-up if the animals are followed for a very long time, if any of the treatments requires a difficult-to-manage manipulation (*e.g.*, a tricky surgical preparation), or if the animals are privately owned, as in a clinical trial with real patients.

The third preliminary question in a comparative study is: will a one-tailed test or a two-tailed test be used? One-tailed (one-sided) means that the investigator is concerned only about a difference in one direction (*e.g.*, do pigs on Diet I gain weight faster than pigs on Diet 2?). Two-tailed means that a difference in either direction would be important to discover (*e.g.*, do

pigs on Diet 1 gain weight at a different rate than pigs on Diet 2?).

Two-tailed tests imply greater prior neutrality on the part of the investigator and should be used if there is any doubt about whether the research question is one-tailed or two-tailed. Two-tailed tests also may be preferred by some editors or reviewers. Two-tailed tests require greater sample sizes, because one must check in both directions for the smallest difference worth detecting. Therefore, if a one-sided test can be justified unequivocally, use it. For example, it may have been shown already that Drug A1 is cheaper to make and more efficacious than its parent, Drug A. Obviously, Drug A1 will be more desirable than Drug A unless Drug A1 has significantly greater side effects. The test for side effects is clearly one-sided; it would be nice to know if Drug A1 has fewer side effects, but the real decision hinges on whether or not it has more.

Designation of the Experimental Unit

The calculated sample size refers to the number of independent experimental units, not to subsamples of those units. The experimental unit is the smallest divisible unit that can be assigned independently to a treatment group under the study's protocol. If the treatment is assigned and delivered to pregnant ewes, then it does not matter whether the ewes have triplets, twins, or singletons. The sample size is based on ewes even if the effect is measured in the lambs. If the treatment is delivered to nursing ewes, the same holds true. If the treatment is, instead, delivered to the lambs, and lambs of the same litter can receive different treatments, then lambs are the experimental unit. One especially must be aware of this problem whenever animals are managed in groups or where litters are used.

Doing the Sample Size Estimation

Estimating the sample size is difficult. Sample size formulas, tables, and computer programs exist for 2 by 2 chi-square tests, for unconditional odds ratios/relative risks, for simple Pearson correlations, for t-tests, and for one-way, Gaussian analysis of variance (ANOVA). There are no calculations specific for generalized r x *c* chi-squares, for multiple correlations, for multiple regressions, for multiple ANOVA, for nonparametric correlations, or for any other nonparametric test that relies on ranks.

So what does one do in one of the latter situations? Calculate the sample size using the closest, most analogous test for which a formula is available. Then, reinterpret this calculated size based on whether the real test will have adjustment for other variables (*i.e.*, will be multiple rather than simple) or whether the real test will be nonparametric rather than parametric.

For example, suppose the real test will be a multiple ANOVA (Gaussian). Calculate the sample size for each of the independent factors of primary interest as if each were to be tested in a one-way ANOVA. Primary interest would imply that these factors are central to the research hypothesis. Use the largest of the calculated sample sizes; this will be adequate for each of the primary factors. If the secondary factors in the model were carefully chosen, their inclusion should decrease the model's error term. This means that it should be easier to detect association between the primary variables and the outcome. However, each additional variable in the model removes at least one degree of freedom (one animal) from the sample size available to test the primary variables, so there is a trade-off.

As another example, suppose the real test will be Wilcoxon's rank sum test. The analogous test with a sample size formula is the t-test. Make a rough guess at a standard deviation, and calculate the t-test sample size. The nonparametric test will need a sample size at least as big, and probably a little bigger. Fudge the calculated sample size upwards, then go to a table that gives critical values for the rank sum test to make sure that—if everything goes well and there is a reasonable difference—statistical significance is achievable with the fudged-upwards sample size.

There are tables and nomograms that give confidence intervals on some descriptive statistics, such as correlations and binomial proportions. These tables can be used to get a rough idea of sample size by using the tables "backwards." Use the best guess of what the value is expected to be, combined with the confidence interval around that value that is small enough to be useful. The size of the confidence interval will be based on either the precision needed or the smallest difference worth detecting. The upper and lower confidence limits for the best-guess value each will have an associated sample size. The two sample sizes read from the table will bound a reasonable guess at the needed sample size. For example, the CRC Handbook[3] has nomograms that give 95 and 99 percent confidence belts on proportions. Suppose that alpha=0.05(95 percent confidence interval), that the best guess is that the proportion will be 40 percent, and that the confidence interval is to be within +/- 10 percent. Looking on the 95 percent nomogram at the intersections of the estimated proportion of 40 percent with the confidence limits of 30 and 50 percent brings one roughly to the sample size lines (belts) for n=100; 100 animals are needed.

There is one note of caution in using sample size formulas, tables, nomograms, and so on. Always be careful to read the description carefully to determine two things: whether the calculation is for a one-tailed test or a two-tailed test; and whether the calculated number is the total animals that will be needed or whether it is the number needed in each group. If the latter is the case, then the total number of animals needed will be the calculated number times the number of groups. If the table is for two-tailed tests, and a one-tailed test is desirable, look up the sample size for the alpha level that is twice as large as wanted.

Ways to Decrease the Needed Sample Size

Suppose the sample size calculations indicate a sample size that is not practical—too costly, for example, or perhaps not obtainable within the appropriate time frame. Before giving up and designing a different project, there are some other approaches to try, assuming that the declared alpha, beta, sidedness, and smallest difference worth detecting were the results of decisions that were well-thought-out.

Reduce the Number of Treatment Groups

Suppose four treatment groups were originally planned in a one-way ANOVA, for which the smallest difference equals 0.38 SD, alpha=0.10, and beta=0.05. The total animals needed would be 200 per group. If treatments are reduced to three, 45-50 fewer animals per group would be needed. This might be a reasonable approach if, for instance, the treatments were several different doses of the same drug, rather than several different drugs.

Change from a Dichotomous Outcome to a Continuous Outcome

Discrete data are less powerful than continuous data; it is easier to find significant differences with continuous data. Suppose the original outcome was the proportion of foals that became ill within the first 30 days of life. One might analyze instead the number of days ill per foal, the number of episodes of illness per foal, or the days to first illness per foal. An easy example of the savings in sample size is difficult to show, because changing from a discrete to a continuous outcome also changes the test that would be used to compare the groups of foals—but there should be a reduction in the number of foals needed.

Decrease the Variability in the Data

If one decreases the variability, it will be easier to tell that a difference is significant. Alternatively, it will be as easy to tell that a smaller difference is significant, or one could detect the same difference with the same error rates but with fewer animals.

One way to decrease variability is to decrease measurement error by improving the measuring tool or the measuring technique. Another way is to decrease the biologic variability, and this may be easier to manipulate than inventing a new tool.

Biologic variation also can be decreased by careful use of repeated measures. For example, before-and-after counts may be made to adjust for baseline, several measurements at one point in time might be taken to improve the precision of the estimate of the effect at that point, or cross-over designs—which may require help from a statistician—might be used so that each treatment is tested in each animal. However, the sample size still is based on the experimental unit, not on the number of measurements. Repeated measures are discussed by Shott[4].

Finally, as mentioned above, one may be able to add other variables to the model so that the residual variation—the error term—is decreased.

Some Recommended References

The following references are recommended for sample size tables and nomograms; numerous other useful references also exist.

- ▲ Beyer[3] for confidence limits on proportions;
- ▲ Beyer[3] for confidence limits on Pearson correlations;
- ▲ Aleong and Bartlett[5] for comparing two proportions;
- ▲ Fleiss[6] for 2 by 2 chi-square tests,
- ▲ Beyer[3] for t-tests between two means;
- ▲ Glantz[7] for t-tests and paired t-tests;
- ▲ Kastenbaum *et al.*[8] for one-way (Gaussian) analysis of variance; and
- ▲ Schlesselman[9] for odds ratios.

References

1. Frieman, J.A., Chalmers, T.C., Smith, H., and Kuebler, R.R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New Engl. J. Med.*; **299**: 690-694, 1978
2. Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., and Smith, P.G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br. J. Cancer*; **34**: 585-612, 1976.
3. Beyer, W.H. *Handbook of Tables for Probability and Statistics*, 2d ed. CRC Press, Boca Raton, FL, 1968.
4. Shott, S. Statistics in veterinary research. *J. Am. Vet. Med. Assoc.*; **187**: 138-141, 1985.
5. Aleong, J. and Bartlett, D.E. Improved graphs for calculating sample sizes when comparing two independent binomial distributions. *Biometrics*; **35**: 875-881, 1979.
6. Fleiss, J.L. *Statistical Methods for Rates and Proportions*, 2d ed. John Wiley & Sons, New York, 1981.
7. Glantz, S.A. *Primer of Biostatistics*, 2d ed. McGraw-Hill, New York, 1987.
8. Kastenbaum, M.A., Hoel, D.G., and Bowman, K.O. Sample size requirements:one-way analysis of variance. *Biometrika*; **57**: 421-430, 1970..
9. Schlesselman, J.J. *Case-Control Studies*. Oxford University Press, New York, 1982.