Choosing Sample Size

The purpose of this section is to describe three related methods for choosing sample size before data are collected -- the classical power method, the sample variation method and the population variation method. The classical power method applies to almost any statistical test. After presenting general principles, the discussion zooms in on the important special case of factorial analysis of variance with no covariates. The sample variation method and the population variation methods are limited to multiple linear regression, including the analysis of variance and covariance. Throughout, it will be assumed that the person designing the study is a scientist who will only be allowed to discuss results if a null hypothesis is rejected at some conventional significance level such as $\alpha = 0.05$ or $\alpha = 0.01$. Thus, it is vitally important that the study be designed so that scientifically interesting effects are likely to be be detected as statistically significant.

The classical power method. The term "null hypothesis" has mostly been avoided until now, but it's much easier to talk about the classical power method if we're allowed to use it. Most statistical tests are based on comparing a full model to a reduced model. Under the reduced model, the values of population parameters are constrained in some way. For example, in a one-way ANOVA comparing three treatments, the parameters are μ_1, μ_2, μ_3 and σ^2. The reduced model says that $\mu_1=\mu_2=\mu_3$. This is a *constraint* on the parameter values. The **null hypothesis** (symbolized H_0) is a statement of how the parameters are constrained under the reduced model. When a test of a null hypothesis yields a small p-value, it means that the data are quite unlikely if the null hypothesis is true. We then reject the null hypothesis -- that is, we conclude it's not true, and therefore that some effect of interest is present in the population.

The following definition applies to hypothesis tests in general, not just those associated with common multiple regression. Assume that data are drawn from some population with parameter θ -- that's the Greek letter theta. Theta is typically a vector; for example, in simple linear regression with normal errors, $\theta = (\beta_0, \beta_1, \sigma^2)$.

The **power** of a statistical test is the probability of obtaining signific
the true parameter values. That is, it is a function of θ.

The **power** of a statistical test is the probability of obtaining significant results. Power is a function of the true parameter values. That is, it is a function of θ.

a) The common statistical tests have infinitely many power values.
b) If the null hypothesis is true, power cannot exceed å; in fact, this is the technical definition of α. Usually, $\alpha = 0.05$.
c) If the null hypothesis is false, more power is good.
d) For a good test, power $\rightarrow$ 1 (for fixed n) as the true parameter values get farther from those specified by the null hypothesis.
e) For a good test, power $\rightarrow$ 1 as n $\rightarrow \infty$ for any combination of fixed parameter values, provided the null hypothesis is false.

Classical power analysis is used to select a sample size n as follows. Choose an effect -- a particular combination of parameter values that makes the null hypothesis false. If possible, select the weakest effect that would still be scientifically important if it were present in the population. If the null hypothesis is false in this way, we would like to have a high probability of rejecting it and obtaining significance. Choose a sample size n, and calculate the probability of significance (that is, calculate power) for that sample size and that set of parameter values. Increase (or decrease) n, calculating power each time. Stop when the power is what you want. A common target value for power is 0.80. My guess is that it would be higher, except that, for common tests and effect sizes, the sample would have to be prohibitively large.

There are only two difficulties with carrying out a classical power analysis in practice; one is conceptual, the other technical. The conceptual problem is that scientists often have difficulty choosing a configuration of parameter values corresponding to an effect that is scientifically interesting. Maybe that's not too surprising, because scientists usually think in terms of data rather than in terms of statistical models. It could be different if the statistical models were serious scientific models of what the scientists are studying, but usually they're quite generic.

The technical problem is that sometimes -- especially for statistical methods other than those based on common multiple regression -- it can be difficult to calculate the probability of significance when the null hypothesis is false. This problem is not really serious; it can always be overcome with some effort and

the right software. Once you move beyond multiple regression, SAS is not the right software.

Power for Factorial ANOVA. Considering this special case will provide a concrete example of the classical power method. It is also the most common example of power analysis.

The distributions commonly used for practical hypothesis testing (mainly the chi-square, t and F) are ones that hold when the null hypothesis is true. When the null hypothesis is false, these are no longer the distributions of the common test statistics; instead, they have probability distributions that migrate more into the rejection region (tail area, above the critical value) of the statistical test. The F distribution used for testing hypotheses in multiple regression is the central F distribution. If the null hypothesis is *false*, the F statistic has a non-central F distribution with parameters s, n-p and ϕ. The quantity ϕ is a kind of squared distance between the reduced model and the true model. It is called the **non–centrality parameter** of the non-central F distribution; $\phi \geq 0$, and $\phi = 0$ gives the usual central F distribution. The larger the non-centrality parameter, the greater the chance of significance -- that is, the greater the power.

The general formula for ϕ is best written in the notation of matrix algebra; it will not be given here. But the general idea, and some of its essential properties, are shown by the special case where we are comparing two treatment means (as in a two-sample t-test, or a simple regression with a binary independent variable). In this situation, the general formula for the non-centrality parameter of the non-central F distribution reduces to

$$\phi = \frac{(\mu_1 - \mu_2)^2}{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})} = \frac{\delta^2}{(\frac{1}{n_1} + \frac{1}{n_2})}, \qquad (4.3)$$

where $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$. Right away, it is possible to make some useful comments.

$$\phi = \frac{(\mu_1 - \mu_2)^2}{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})} = \frac{\delta^2}{(\frac{1}{n_1} + \frac{1}{n_2})}, \qquad (4.3)$$

where $\delta \equiv \frac{|\mu_1 - \mu_2|}{\sigma}$.

- The quantity δ is called **effect size**. It specifies how wrong the statement $\mu_1=\mu_2$ is, by expressing the absolute difference between μ_1 and μ_2 in units of the common within-cell standard deviation σ.

- For any statistical test, power is a function of the parameter values. Here, the non-centrality parameter (and hence, power) depends on the three parameters μ_1, μ_2 and σ^2 *only* through the effect size. This is quite wonderful; it does not always happen, even in the analysis of variance.

- The larger the effect size (that is, the more wrong the reduced model is -- in this metric), the larger the non-centrality parameter ϕ, and therefore the larger the probability of significance.

- If $\mu_1=\mu_2$, then $\delta=0$, $\phi=0$,the non-central F distribution becomes the usual central F distribution, and the probability of significance becomes exactly $\alpha=0.05$.

- The size of the non-centrality parameter depends on another quantity involving *both* n_1 and n_2, not just the total sample size $n = n_1+n_2$.

This last point can be illuminated by a bit of algebra. Let

- $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$
- $n = n_1 + n_2$
- $q = \frac{n_1}{n}$, the proportion of the sample allocated to Group One.

Then expression (4.3) can be re-written

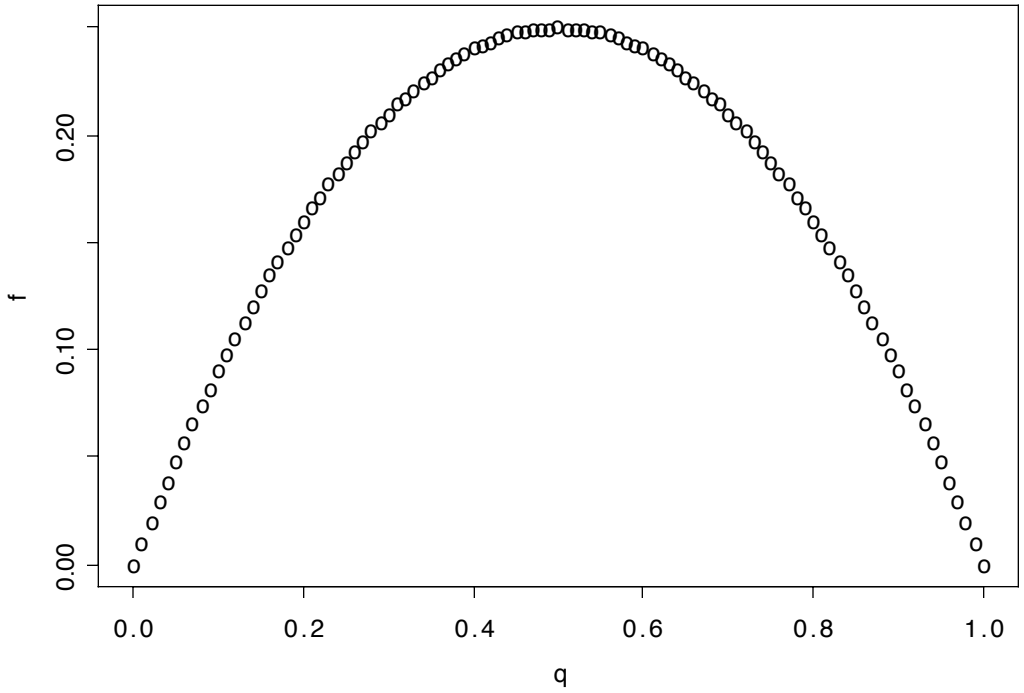
$$\phi = n\ q(1-q)\ \delta^2. \qquad (4.4)$$

Now it's clear.

- For any non-zero effect size and any (?) allocation of sample size to the two treatments, the greater the total sample size, the greater the power.
- For any sample size and any (?) allocation of sample size to the two treatments, the greater the effect size, the greater the power.
- Power depends not just on sample size and effect size, but on an aspect of *design* -- the allocation of sample size to the two treatments. This is a general feature of power in the analysis of variance and other statistical methods. It is important, but usually not mentioned.

Let's continue to pursue this interesting special case. For any given sample size and any non-zero effect size, we can maximize power by choosing q (the proportion of cases allocated to Group One) so that the function f(q) = q(1-q) is as large as possible. What's the best value of q?

This is a simple calculus exercise, but the following plot gives the answer by brute force. I just computed f(q) = q(1–q) for 100 equally spaced values of q ranging from zero to one.



So the best value of q is 1/2. That is, for comparing two means using the classical normal model, power is highest when the sample sizes are equal -- and this holds regardless of the total sample size or the magnitude of the effect.

This is a clear, simple example of something that holds for *any* classical ANOVA. The non-centrality parameter, and hence the power, depends on the total sample size, the effect, *and* the allocation of the sample to treatment combinations.

Equal sample sizes do not always yield the highest power. In general, the optimal allocation depends on the hypothesis being tested *and* the nature of the true effect. For example, suppose you have a design with 18 treatment combinations, and the test in question is to compare μ_1 with the average of μ_2 and μ_3. Further, suppose that $\mu_2 = \mu_3 \neq \mu_1$ (σ^2 can be anything); this is the effect. The optimal allocation is to give half the sample to Treatment One, split the other half any way at all between Treatments 2 and 3, and let n=0 for the other 15 treatments. This is why observations are not usually allocated to treatments based on a power analysis; it often advises you to put all your eggs in one basket.

In the analysis of variance, power analysis is used to select a sample size n as follows.

1. Choose an allocation of observations to treatments; usually, this is done without formal analysis, equal sample sizes being the most common choice.

2. Choose an effect. Your null hypothesis says that some collection of contrasts (of the treatment combination means) are all zero in the population. The "effect" you need to specify is that one or more of those contrasts is *not* zero. You must provide exact non-zero values, in units of the common within-treatment population standard deviation σ -- like, the difference between μ_1 and the average of μ_2 and μ_3 is minus 0.75σ. You don't need to know the numerical value of σ (thank goodness!), but you do need to be able to express differences between population means in units of σ. If possible, select the weakest effect that is still scientifically important.

3. Choose a desired power; again, a common choice is 0.80, but it's up to you.

4. Start with a modest but realistic value for the total sample size n. Increase it, each time determining the critical value of F, calculating the non-centrality parameter ϕ (you have enough information), and using ϕ to compute the probability that F will exceed the critical value. When that power becomes high enough, stop.

This is a rational strategy for choosing sample size. In practice, the hard part is selecting an effect. Scientists often can say what's a scientifically meaningful difference between means, but they usually have no clue about σ. Statisticians respond with the suggestion that σ^2 be estimated by MSE_F from similar studies. Scientists respond that there are no "similar" studies; the investigation being planned is new -- that's why we're doing it. In the end, the whole thing is based on so much guesswork that everyone feels uncomfortable. In my experience, this is what happens most of the time when people try to do a classical power analysis. Of course, there are exceptions; sometimes, everyone is happy.