# Choosing Sample Size

Lawrence J. Brunner

January 14, 2004

**Abstract**

Four methods for selecting sample size are described: the Classical Power Method, the Sample Variation Method, the Population Variation Method, and the Dummy Population Method. The Classical Power Method, the Population Variation Method, and the Dummy Population Method are variants of Power analysis; the Sample Variation method is something else.

The discussion of the Classical Power Method treats the case of fixed-effects multiple regression in considerable detail, and both the Sample Variation Method and the Population Variation Method are limited to this important special case. The Dummy Population Method is a way to do power analysis by simulation in situations where the usual parametric assumptions are not satisfied, or are too difficult for the practitioner to think about. It is the most general of the methods discussed here, but it also requires more effort because it is computer-intensive. Computed examples of all the methods are provided, with SAS and S code.

## 1 Introduction

The purpose of this technical report is to describe four related methods for choosing sample size *before* data are collected — the classical power method, the sample variation method, the population variation method and the dummy population method. The classical power method applies to almost any statistical test. After presenting general principles and giving an example of calculating power by simulation, the discussion focuses on the important special case of linear models with i.i.d. normal errors, including multiple regression and the factorial analysis of variance and covariance.

In practice, the classical power method is often difficult to apply, because power is the probability of rejecting the null hypothesis *as a function of the true parameter values.* The classical power method is based on calculating such a probability for some particular set of parameter values deemed scientifically meaningful by a scientist or subject-matter expert, and then, for that configuration of parameter values, adjusting the sample size until an acceptably high probability of rejecting the null hypothesis is achieved. Perhaps because scientists tend not to think in terms of the parameters of statistical models, the process tends to break down — even in consulting situations, where expert advice is available.

The sample variation method, the population variation method and the dummy population method are all attempts to avoid the problem of specifying parameter values. The sample variation method and the population variation method apply only to the normal linear model, while the dummy population method is more general. It is also more computer-intensive, and usually requires a bit of custom programming.

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, the objective is to design the study be designed so that scientifically interesting effects can be detected as statistically significant.

Two important problems are deliberately excluded from this discussion. One is choosing sample size so that a confidence interval will have a desired width. The main *practical* example is the estimation of a percentage or proportion when sampling from a finite population under various sampling plans, like stratified random sampling or cluster sampling. A good introductory reference is Sharon Lohr's *Sampling: Design and analysis* [2].

The second problem not discussed here is the *estimation* of statistical power after data have already been collected. This will be treated in a separate document. For now, it's enough to say *watch out*! Estimates of power can be quite inaccurate. It is an excellent example of why you should beware of point estimates that are not accompanied by some margin of probable error.

This report does not contain anything really new; everything in it is well known, in some sense. Classical power analysis makes an appearance in most treatments of experimental design – for example [3]. But examples are typically limited to the one-way analysis of variance, and the treatment tends to be superficial. Cohen's *Statistical power analysis for the behavioral sciences* [1] contains what I call the Population Variation Method. Cohen just calls it "statistical power analysis." For most psychologists, Cohen's book is the Bible of power analysis. They have no idea how unusual it is from the standpoint of a professional statistician. As for the Sample Variation Method and the Dummy Population Method, I have never seen them in print anywhere, but they are so simple that many people must know about them.

## 2 The Classical Power Method

We begin with the following general hypothesis-testing framework. The data are $Y$; the distribution of $Y$ depends on the parameter $\theta$, and we are testing a null hypothesis $H_0$ using a size $\alpha$ test that corresponds to the critical region $C$.

$$\begin{array}{l}
Y = (Y_1, \ldots, Y_n)', \text{ where } n \text{ is the sample size.} \\
Y \sim P_\theta,\ \theta \in \Theta, \\
H_0 : \theta \in \Theta_0 \text{ v.s. } H_A : \theta \in \Theta \cap \Theta_0^c, \\
\text{and critical region } C = \{y : H_0 \text{ is rejected when } Y \in C\} \\
\text{with } P_\theta\{Y \in C\} \leq \alpha \text{ for all } \theta \in \Theta_0
\end{array} \qquad (1)$$

Elements of the random quantity $Y$, elements of the corresponding set of fixed values $y$, and the parameter $\theta$ can be very large vectors. For example, in testing the parameters of a multivariate normal we might have $Y = \mathbf{Y}_1, \ldots, \mathbf{Y}_n$, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f(y;\theta)$ might be the product of $n$ multivariate normal densities.

The *power* of the test is defined as the probability of rejectiong the null hypothesis. Symbolically, it is $P_\theta\{Y \in C\}$. Note that

- Power is a function of the true parameter value $\theta$; the common statistical tests have infinitely many power values.

- If $H_0$ is true, power cannot exceed $\alpha$; in fact, this is the definition of $\alpha$.
- If the null hypothesis is false, more power is good.
- For a good test, power approaches one (for fixed sample size $n$) as the true parameter values get farther from those specified by the null hypothesis.
- For a good test, power goes to 1 as $n \rightarrow \infty$ for fixed parameter values, provided $H_0$ is false. Such tests are called *consistent*.

In the Classical Power Method, the sample size $n$ is selected as follows. $H_0$ can be false in infinitely many ways, one for each value of $\theta$. Select one that is interesting; if possible select a value of $\theta$ corresponding to the weakest effect that would still be scienticically important. If $H_0$ is false in this way, we would like to have a high probability of rejecting it. Calculating the power $P_\theta\{Y \in C\}$ repeatedly, increasing $n$ until the power is as high as you desire; that is the sample size you need.

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 it's often difficult to choose a configuration of parameter values corresponding to an effect that is scientifically interesting. Who's going to pick the paramter value? Not the statistician! Statisticians don't generally know enough about the subject matter to decide what is important; they have to ask somebody, usually the client who brought them the request in the first place. This is particularly true because the parameter values that are most useful in a power analysis are typically in a kind of grey area, just barely far enough away from the null hypothesis to be meaningful. The statistician really has to ask the client.

But the client doesn't know either. He or she knows a lot about the subject matter, but seldom has much familiarity or comfort with the statistical model. In a better world, the statistical model would be a true theoretical model for the data, and the scientist planning data collection would be intimately familiar with its parameters and their meaning. But it's not a better world. At least in the biological and social sciences (with the possible exception of Economics), the statistical models in common use are deliberately quite generic, and the tests associated with them are seen as semi-arbitrary conventions for deciding when it is permissible to discuss and publish a set of results. To do a power analysis, one must take the model and its parameters seriously, and go beyond a consideration of whether some of the parameters are zero or not. It's not easy.

There is sometimes appears to be technical difficulty, too; it is this. Especially for statistical methods other than those based on common multiple regression, it can be difficult to calculate the probability of rejecting the null hypothesis when it is false. This problem is actually not very serious. The desired probability can almost always be approximated by simulation. Here is an example.

## 2.1 An Example of Power Analysis by Simulation

Suppose we want to test whether a collection of variables are uncorrelated. That is, we want to know if the off-diagonal elements of a correlation matrix (or equivalently, a variance-covariance matrix) are all zero. Say we have 5 variables, we are willing to assume multivariate normality, and we are going to use a large-sample likelihood ratio test. Would $n = 50$ be large enough?

As usual, the answer to such a question is another question: Large enough for what? It might mean "Large enough for the large-sample test to have approximately size $\alpha$", or it might mean "Large enough to detect a reasonable effect with acceptably high probability." Both questions can be answered by simulation; the second one is a standard question about power, and that is the one we will consider.

Since the data in this case will be assumed multivariate normal, the parameter is $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a vector of five elements, and $\boldsymbol{\Sigma}$ is a $5 \times 5$ symmetric positive definite matrix. We need to pick a particular value of $\theta$ corresponding to a reasonable and scientifically interesting effect, and then calculate the probability of rejecting the null hypothesis for that value of $\theta$. Here is one possibility, out of infinitely many.

Suppose there are two "clusters" of variables. In the population, three of the variables are correlated with one another at about level $\rho = 0.30$. The other two variables are correlated at about 0.15, and the three variables in the first group are uncorrelated with the two variables in the second group. Again, there are infinitely many variance-covariance matrices with this structure. We'll just pick one.

$$\boldsymbol{\Sigma} = \left[ \begin{array}{ccccc} 1.42 & 0.42 & 0.42 & 0.00 & 0.00 \\ 0.42 & 1.42 & 0.42 & 0.00 & 0.00 \\ 0.42 & 0.42 & 1.42 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.18 & 0.18 \\ 0.00 & 0.00 & 0.00 & 0.18 & 1.18 \end{array} \right] \qquad (2)$$

By a sufficiency argument and the well-known independence of the sample mean vector $\overline{\mathbf{X}}$ and the sample variance-covariance matrix $\mathbf{S}$, the value of $\boldsymbol{\mu}$ we use does not matter; we'll just let $\boldsymbol{\mu}$ be a vector of zeros. Thus, we have chosen a $\theta$.

Denoting the sample variance-covariance matrix by $\mathbf{S}$, and the $j$th sample variance (diagonal element of $\mathbf{S}$) by $s_j^2$, the large-sample likelihood ratio test statistic for $k$ variables may be written

$$G = n(\sum_{j=1}^{k} \log s_j^2 - \log |\mathbf{S}|), \qquad (3)$$

where $|\mathbf{S}|$ refers to the determinant of $\mathbf{S}$. Under the null hypothesis that $\boldsymbol{\Sigma}$ is diagonal, $G$ has a chisquare distribution with $\frac{1}{2}k(k-1)$ degrees of freedom (the number of unique off-diagonal elements). Here, the degrees of freedom equal 10.

Using the usual $\alpha = 0.05$ significance level, how do we calculate the power — that is, the probability that the chisquare test just described will yield $p < 0.05$? It would be very hard or impossible to get an exact answer analytically, but it's easy to get an approximate answer by simulation. It may be done like this.

1. Randomly generate $n = 50$ observations from a multivariate normal with mean vercor zero and variance-covariance matrix (2).

2. Compute the chisquare statistic $G$ using (3).

3. Check whether $G$ is significant at the 0.05 level.

Do this $M$ times ($M$ is the "Monte Carlo sample size"). By the Law of Large Numbers, the proportion of tests that are significant converges to the probability of significance, or power.

The number we get, of course, will just be an *estimate* of the power. How accurate is the estimate? We'll accompany every Monte Carlo estimate of a probability with a confidence interval. Here's the formula. It's based on the normal approximation to the binomial, not bothering with a continuity correction.

$$\widehat{P} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{P}(1-\widehat{P})}{M}} \tag{4}$$

This formula is implemented in the S function `merror` for "margin of error."

```
merror <- function(phat,m,alpha) # (1-alpha)*100% merror for a proportion
    {
    z <- qnorm(1-alpha/2)
    merror <- z * sqrt(phat*(1-phat)/m)  # m is (Monte Carlo) sample size
    merror
    }
```

The Monte Carlo estimate of the probability is denoted by $\widehat{P}$, the quantity $M$ is the Monte Carlo sample size, and $z_{1-\alpha/2}$ is the value with area $1 - \frac{\alpha}{2}$ to the left of it, under the standard normal curve. Typically, we will choose $\alpha = 0.01$ to get a 99% confidence interval, so $z_{1-\alpha/2} = 2.575829$.

How should we choose $M$? In other words, how many data sets should we simulate? It depends on how much accuracy we want. Since our policy is to accompany Monte Carlo estimates with confidence intervals, we will choose the Monte Carlo sample size to control the width of the confidence interval. The following table can be useful.

Table 1: Monte Carlo Sample Size Required to Estimate Power with a Specified 99% Margin of Error

| | Power Being Estimated | | | | | |
|---|---|---|---|---|---|---|
| Margin of Error | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.99 |
| 0.10 | 140 | 125 | 107 | 85 | 60 | 7 |
| 0.05 | 558 | 498 | 425 | 339 | 239 | 27 |
| 0.01 | 13,934 | 12,441 | 10,616 | 8,460 | 5,972 | 657 |
| 0.005 | 55,734 | 49,762 | 42,464 | 33,838 | 23,886 | 2,628 |
| 0.001 | 1,393,329 | 1,244,044 | 1,061,584 | 845,950 | 59,7141 | 65,686 |

Now we have all the information we need. Let's choose $M = 10,000$, to get a margin of error equal to 0.01 if the power is a respectable 0.80. Here's some S code.

```
# cvm.R
# Power for test of zero correlation for an entire matrix
# (Large Sample LR Test)
# Execute with   source("cvm.R")
#
M <- 10000
sim <- numeric(M)
set.seed(32448)
n <- 50 ; v1 <- .42 ;       v2 <- .18
```

```
          s1 <- sqrt(v1) ; s2 <- sqrt(v2)
G <- function(datamat)
   {
   nn <- dim(datamat)[1] ; kk <- dim(datamat)[2] ; df <- kk*(kk-1)/2
   G <- numeric(3)
   names(G) <- c("Chisq","df","P-value")
   S <- var(datamat)
   G[1] <- nn * ( sum(log(diag(S))) - sum(log(eigen(S)$values)) ) #$
   G[2] <- df
   G[3] <- 1 - pchisq(G[1],df)
   G
   } # End function G
merror <- function(phat,m,alpha=0.01) # (1-alpha)*100% merror for a proportion
    {
    z <- qnorm(1-alpha/2)
    merror <- z * sqrt(phat*(1-phat)/m)  # m is (Monte Carlo) sample size
    merror
    }

for(j in 1:M)
   {

e1 <- rnorm(n,0,s1) ; e2 <- rnorm(n,0,s2)
x1 <- rnorm(n)+e1 ; x2 <- rnorm(n)+e1 ;  x3 <- rnorm(n)+e1 ;
x4 <- rnorm(n)+e2 ;  x5 <- rnorm(n)+e2
dat <- cbind(x1,x2,x3,x4,x5)
# print(G(dat))
print(j)
sim[j] <- G(dat)[3] < .05

    }
poww <- length(sim[sim==1])/M
cat("Power = ", poww ,"\n")
cat("Plus or Minus 99% Margin of error: ",merror(poww,M),"\n")
```

Here's the output.

```
Power =  0.6987
Plus or Minus 99% Margin of error:  0.01181849
```

## 2.2 The Normal Linear Model

In this section, we are concerned concerned with the usual univariate linear model with independent normal errors; note that this model subsumes all the usual fixed-effects factorial analysis of variance and covariance methods. Here is the model.

$$\mathbf{Y} \; = \; \mathbf{X}\boldsymbol{\beta} \; + \; \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{X}$ is an $n \times r$ matrix of known constants, $\boldsymbol{\beta}$ is a $r \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, and $\sigma^2 > 0$ is an unknown constant. It will be assumed that the rank of $\mathbf{X}$ is $r$, so the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'X)^{-1}\mathbf{X}'Y$, and the MLE of $\sigma^2$ is $\hat{\sigma}^2 = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\beta})'(\mathbf{Y}-\mathbf{X}\widehat{\beta})}{n}$. The Mean Squared Error ($MSE$) is $\frac{n\hat{\sigma}^2}{n-r}$.

We will consider linear null hypotheses of the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h}$, where $\mathbf{C}$ is a $q \times r$ matrix of row rank $q \leq r$. The $F$ statistic for testing this null hypothesis is

$$F^* = \frac{(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{q\,MSE} \tag{6}$$

Under the model assumptions and null hypothesis, $F^*$ has an $F$ distribution with $q$ and $n-r$ degrees of freedom. This distribution is the familiar "central" $F$ distribution; it holds when the null hypothesis is true. When $H_0$ is false, $F^*$ has a *noncentral* $F$ distribution with parameters $q$, $n-r$ and $\phi$. The quantity $\phi$ is a sort of squared distance between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$, and is measure of how wrong the null hypothesis is. It is called the *noncentrality parameter* of the noncentral $F$ distribution. Because $\phi$ is a squared distance, we always have $\phi \geq 0$. If $\phi = 0$, the noncentral $F$ distribution reduces to the usual central $F$. A useful formula for $\phi$ is

$$\phi = \frac{(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})}{\sigma^2} \tag{7}$$

This expression is taken from Sen and Srivastava (provide reference with page). Other texts define the noncentrality parameter in a variety of different ways. If you are looking at textbook examples or power tables, you need to be very careful about which definition is being used. Because power analysis is best done with software rather than tables, it's helpful to know that SAS, SPlus and R use (7) as their definition of the noncentrality parameter.

Letting $F^*$ represent a random variable with a noncentral $F(q, n-r, \phi)$ distribution and denoting the $1-\alpha$ quantile of the *central* $F$ distribution by $F_{1-\alpha}(q, n-r)$, the power of the $F$-test is $Pr\{F^* > F_{1-\alpha}(q, n-r)\}$. This probability — that is, the power — is an increasing function of $\phi$. As $\phi$ increases, power approaches one.

Remember the recipie for selecting a sample size with the classical power method. First, pick a specific parameter value representing a scientifically meaningful way in which the null hypothesis could be false. Then, assuming that parameter value, increase the sample size until power is acceptably high. How does this work in the present case?

First, notice that though the parameter $\theta = (\boldsymbol{\beta}, \sigma^2)$ is of dimension $r+1$, power depends on $\theta$ only through $\phi$, a one-dimensional function of $\theta$ and the $\mathbf{X}$ matrix. For any given $\mathbf{X}$ matrix, infinitely many values of $\theta$ are represented by a single noncentrality parameter $\phi$. This is a valuable simplification.

Second, the sample size $n$ is hidden just under the surface in formula (7), because $n$ is the number of rows in the $\mathbf{X}$ matrix. Thus, $\phi$ depends on $n$ through $\mathbf{X}'\mathbf{X}$. For example, if the regression model has an intercept, the upper left-hand element of the $\mathbf{X}'\mathbf{X}$ matrix is $n$.

Now we will consider a set of examples, moving from the simple to the more complex and realistic. We'll start with the comparison of two means (the two-sample $t$-test), then do the comparison of more than two means (one-way ANOVA), multifactor ANOVA, and finally general linear regression models in which one or more independent variables may be random. I believe that the discussion accompanying the simpler examples is informative about power analysis in general, or I would not have included it. However, a reader who wants to just *do* a power analysis for some factorial ANOVA design with fixed effects can skip to Section 2.2.3 on factorial experiments and look at the matrix approach.

### 2.2.1 Comparing Two Means

At this point, a simple example may be useful. We will fit a power analysis for a two-sample $t$-test into the multiple regression framework we have been developing. This will clarify the meaning of some of the notation, and also reveal features shared by more complicated and realistic cases.

Suppose we have a random sample of size $n_1$ from a normal distribution with mean $\mu_1$ and variance $\sigma^2$, and independently, a second random sample from a normal distribution with mean $\mu_2$ and variance $\sigma^2$. We wish to test $H_0 : \mu_1 = \mu_2$ versus the alternative $H_a : \mu_1 \neq \mu_2$.

We'll do it with dummy variable regression, letting $x_i = 1$ if observation $i$ is from population one, and $x_i = 0$ if observation $i$ is from population two. The model is the usual simple regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Normal$(0, \sigma^2)$ random variables, and $n = n_1 + n_2$.

Clearly, he have $\mu_1 = \beta_0 + \beta_1$, $\mu_2 = \beta_0$, and we are testing $H_0 : \beta_1 = 0$. In matrix notation,

$$\mathbf{C} = \left[ \begin{array}{cc} 0 & 1 \end{array} \right], \boldsymbol{\beta} = \left[ \begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right], \text{ and } \boldsymbol{h} = [0].$$

The $\mathbf{X}$ matrix has $n = n_1 + n_2$ rows; The first column is all ones, and the second column has $n_1$ ones and $n_2$ zeros. The $\mathbf{X}'\mathbf{X}$ matrix is

$$\left[ \begin{array}{cc} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{array} \right] = \left[ \begin{array}{cc} n_1 + n_2 & n_1 \\ n_1 & n_1 \end{array} \right],$$

and

$$(\mathbf{X}'\mathbf{X})^{-1} = \left[ \begin{array}{cc} \frac{1}{n_2} & -\frac{1}{n_2} \\ -\frac{1}{n_2} & \frac{1}{n_1} + \frac{1}{n_2} \end{array} \right].$$

Therefore, in this case the noncentrality parameter in (7) equals

$$\begin{aligned} \phi &= \frac{\beta_1^2}{\sigma^2} \left( \frac{n_1 n_2}{n_1 + n_2} \right) = n \left( \frac{\mu_1 - \mu_2}{\sigma} \right)^2 \frac{n_1}{n} \frac{n_2}{n} = nf(1-f) \left( \frac{\mu_1 - \mu_2}{\sigma} \right)^2 & (8) \\ &= nf(1-f)d^2, & (9) \end{aligned}$$

where $f = \frac{n_1}{n}$, and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$. Now some comments are in order.

- The quantity $d = \frac{|\mu_1 - \mu_2|}{\sigma}$ is called the *effect size*. The effect size specifies how wrong the null hypothesis $\mu_1 = \mu_2$ is, by expressing the absolute difference between $\mu_1$ and $\mu_2$ in units of the common within-cell standard deviation $\sigma$.

- The notation $d$ for effect size is consistent with Cohen's [1] *Statistical power analysis for the behavioral sciences.* Cohen uses a boldface $\mathbf{d}$ to refer to effect size for the two-sample $t$-test.

- For any statistical test, power is a function of the parameter values. Here, the noncentrality parameter (and hence, power) depends on the three parameters $\mu_1$, $\mu_2$ and $\sigma^2$ only through the effect size $d$.

- If the effect size is zero — that is, if the null hypothesis of equal means is true — the noncentral $F$ distribution becomes the usual central $F$ distribution, and the probability of significance (power) becomes exactly $\alpha$. For any non-zero effect size and any allocation of sample size to the two treatments (of course neither sample size may be zero), 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.

- $f = \frac{n_1}{n}$ is the fraction of cases allocated to Treatment One. Thus, we see that 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 often ignored.

- What allocation of sample to treatments results in the greatest power? It's a simple calculus exercise to show that the function $g(f) = f(1-f)$ is maximized when $f = \frac{1}{2}$. This tells us that 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 noncentrality parameter, and hence the power, depends on the total sample size, the effect, and the allocation of the sample to treatment combinations.

- When there are more than two treatments, 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.

Here's how to select a sample size using the classical power method for this two-sample case.

1. Let $f = \frac{1}{2}$ (equal sample sizes), because this maximizes power for any sample size and any effect size in the present case.

2. Choose an effect size; if possible, select the smallest effect that is still scientifically important. You don't need to specify values of $\mu_1$, $\mu_2$ and $\sigma$ individually, but you do need to be able to express a *difference* between population means in units of the population standard deviation.

3. Choose a desired power; 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 noncentrality parameter $\phi$, and using $\phi$ to compute the probability that $F^*$ will exceed the critical value. When that power becomes high enough, stop.

So, suppose we want to be able to detect a difference between population means of half a standard deviation, with probability 0.80. With $f = \frac{1}{2}$ and $d = \frac{1}{2}$, the noncentrality parameter for the two-sample case (formula 8) equals $\frac{n}{16}$.

Most books on experimental design have some nasty power tables that only apply to limited cases anyway. It is better to use a computer. Again, the definition of $\phi$ given in (7) is the one used by SAS, SPlus and R.

Here is some R input and output. It was convenient to start with a likely range of $n$ values, generate the power for each $n$, and then do the search by eye; that way it's two lines of code. The required sample size is $n = 128$.

```
> n <- seq(from=120,to=140,by=2) ; phi <- n/16 ; ddf <- n-2
> cbind(n,pf(qf(.95,1,ddf),1,ddf,phi,FALSE))
        n
 [1,] 120 0.7752659
 [2,] 122 0.7820745
 [3,] 124 0.7887077
 [4,] 126 0.7951683
 [5,] 128 0.8014596
 [6,] 130 0.8075844
 [7,] 132 0.8135460
 [8,] 134 0.8193475
 [9,] 136 0.8249920
[10,] 138 0.8304825
[11,] 140 0.8358223
```

Here is how one might do the same thing with SAS. Note that this program does not produce a list file; it writes on the log file using `put` statements.

```
/*****************  fpow.sas *********************/
options linesize = 79 pagesize = 35 noovp formdlim='-';
data fpower;        /* Replace alpha, q, r and wantpow below           */
     alpha = 0.05;  /* Signif. level for testing H0: C Beta = h        */
     q = 1;         /* Numerator df = # rows in C matrix               */
     r = 2;         /* There are r beta parameters                     */
     wantpow = .80; /* Find n to yield this power                      */
     power = 0; n = r+2; oneminus = 1-alpha;  /* Initializing ... */
     do until (power >= wantpow);
        n = n+1 ;
        ncp = n/16;
        df2 = n-r;
        power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
     end;
```

```
    put ' ';
    put ' *********************************************************';
    put '   A sample size of ' n 'is needed to have probability ' ;
    put '   ' wantpow ' of rejecting H0';
    put ' *********************************************************';
    put ' ';
    put ' ';
```

And the log file contains the following output (along with a load of other stuff, of course):

```
 *********************************************************
   A sample size of 128 is needed to have probability
   0.8  of rejecting H0
 *********************************************************
```

### 2.2.2 Comparing $r$ Means: One-way ANOVA

The preceding example was somewhat informative; but unfortunately, almost nobody designs a study with just two treatments. A little better is the standard one-factor analysis of variance. That is, there are $r$ treatments, and the response variable is assumed normally distributed within treatments. The means of these $r$ normal distributions may differ, but they all have the same variance $\sigma^2$. The null hypothesis is $H_0 : \mu_1 = \ldots = \mu_r$. For this case, the noncentrality parameter in (7) reduces to

$$\phi = \frac{\sum_{k=1}^{r} n_k(\mu_k - \mu_.)^2}{\sigma^2}, \tag{10}$$

where $\mu_. = \sum_{k=1}^{r} \frac{n_k}{n}\mu_k$.

Written in this form, the noncentrality parameter clearly reflects the dispersion of treatment means in the population; it equals zero if and only if they are all the same. Expression (10) is instructive in several ways.

By bringing $\sigma^2$ up into the numerator, (10) may be rewritten

$$\phi = \sum_{k=1}^{r} n_k \left(\frac{\mu_k - \mu_.}{\sigma}\right)^2,$$

so the noncentrality parameter is based on deviations from the average treatment mean *relative* to the within-treatment standard deviation. This is like the two-sample case, and in fact is a general feature of power analysis under the normal model. It is very difficult to avoid specifying differences among means relative to the error standard deviation. This can make power analysis quite difficult to carry out in practice, because typically nobody has the faintest idea what the error standard deviation is.

Another way to re-write (10) is

$$\phi = n\frac{\sum_{k=1}^{r} \frac{n_k}{n}(\mu_k - \mu_.)^2}{\sigma^2} = n\frac{\sum_{k=1}^{r} f_k(\mu_k - \mu_.)^2}{\sigma^2}, \tag{11}$$

where $f_k = \frac{n_k}{n}$. Notice how division and multiplication by $n$ allows us to write the noncentrality parameter as the product of the sample size and a quantity that is either a constant or settles down rapidly to a constant as the sample size increases. This trick is always useful, even in the most general setting.

In the two-sample case, the noncentrality parameter was the product of sample size, effect size and the configuration of relative sample sizes. Therefore, it was possible to optimize relative sample size (obtaining equal sample sizes as the solution) for *any* non-zero effect size. However, (11) tells us that once we get beyond two groups, effect and design are mixed together in a way that's impossible to separate.

Should the quantity $\frac{\sum_{k=1}^{r} f_k(\mu_k-\mu_{.})^2}{\sigma^2}$ still be called *effect size*? The answer is a reluctant "Yes." The affirmative is reluctant because

- It is not a pure effect size of the sort we saw in the two-sample case; it is "size" in a metric strongly influenced by the design — in particular, by the allocation of relative sample size to treatments.

- In the case of two treatments, it does not reduce to $d = \frac{|\mu_1-\mu 2|}{\sigma}$, the quantity called "effect size" in Section 2.2.1. In the two-sample case, the quantity we are now calling effect size equals $d^2/4$.

- Really, "size" should be the *distance* between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$, and the quantity we are discussing is a squared distance. But this is pedantic, and the better choice here is simplicity over precision.

Anyway, for the one-factor ANOVA model, the term *effect size* will refer to

$$\frac{\sum_{k=1}^{r} f_k(\mu_k-\mu_{.})^2}{\sigma^2}.$$

This is a special case of expression (16), which applies to multiple regression in general.

Clearly, the noncentrality parameter (and hence the power) will be increased by allocating a larger fraction of the observations to the treatment whose population mean is most different from the average. If nothing else, this teaches us that equal sample sizes do not always yield optimal power except in the two-sample case.

Still, if one had a good prior idea of which treatments would be most different, one could increase power by allocating more of the sample to those treatments expected to be most different from the average. For a given configuration of population treatment means, one could optimize the relative sample sizes first, and then find the total sample size needed to get the desired power. One could even replace relative sample sizes by continuous quantities and do the optimization numerically for a chosen pattern of treatment means.

It's too bad, but this approach is not productive most of the time. It can be proved that for any set of treatment means in a one-factor design, the noncentrality parameter is maximized by giving *all* the observations to the two treatments whose means are most different, and none to the others. Then, of course, we know that the two sample sizes should be equal. This solution is just not interesting, assuming that one has a good reason to collect data on more than two treatments.

Therefore, what people do most of the time is just close their eyes and use equal sample sizes. This is reasonable enough, and in designs with more than one independent variable, it has the advantage that independent variables are unrelated, and therefore treatment effects are not confounded with one another.

In the case of equal sample sizes, $f_k = \frac{1}{r}$ for $k = 1, \ldots, r$, $\mu_.$ is the arithmetic mean of the treatment means, and (11) becomes

$$\phi = n \, \frac{\frac{1}{r}\sum_{k=1}^{r}(\mu_k - \overline{\mu})^2}{\sigma^2}. \tag{12}$$

Here, $\phi$ is the sample size multiplied by an effect size that looks like the ratio of two variances. One might think of it as a signal-to-noise ratio; this might make it easier to choose a meaningful effect size in some fields.

**Computed Examples** Suppose we have four treatments, and that the four population treatment means are equally spaced, one-quarter of a standard deviation apart. We'd like to be able to detect the differences among treatment means with probability 0.80, using the conventional significance level of $\alpha = 0.05$. We'll use equal sample sizes.

Without loss of generality, we'll let the four population treatment means be 0, 0.25, 0.50 and 0.75. Using R as a calculator, and remembering that the `var` function divides by the number of observations minus one, we'll calculate the effect size in (12) as

```
> 3 * var(c(0,.25,.5,.75)) / 4
[1] 0.078125
```

Here is `fpow2.sas`, a SAS program for finding the required $n$. Like `fpow.sas`, it writes on the log file, producing no list file. This program is quite general. It may be used for any version of the multiple regression model (5). Just edit the file to new substitute values for some of the parameters. The main shortcoming this program is that you must calculate the effect size yourself.

```
*****************  fpow2.sas ************************************************/
options linesize = 79 pagesize = 100  noovp formdlim='-';   /*               */
data fpower;              /* Replace alpha, q, r, effsize and wantpow below  */
     alpha = 0.05;        /* Signif. level for testing H0: C Beta = h        */
     q = 3;               /* Numerator df = # rows in C matrix               */
     r = 4;               /* There are r beta parameters                     */
     effsize = 0.078125;  /* Effect size is ncp/n                            */
     wantpow = .80; /* Find n to yield this power                            */
     power = 0; n = r+2; oneminus = 1-alpha;  /* Initializing ...            */
/*****************************************************************************/
     do until (power >= wantpow);
        n = n+1 ;
        ncp = n * effsize;
        df2 = n-r;
        power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
     end;
     put ' ';
     put ' ***************************************************************';
     put '   With ' r ' beta parameters, testing H0 of ' q ' linear';
     put '   restrictions on the betas and an effect size of ' effsize ',';
     put '   A sample size of ' n 'is needed to have probability ' ;
     put '   ' wantpow ' of rejecting H0 at significance level alpha = ' alpha;
     put ' ***************************************************************';
     put ' ';
     put ' ';
```

Here is the output; it appears on the log file.

```
*****************************************************************
   With 4  beta parameters, testing H0 of 3  linear.
   restrictions on the betas and an effect size of 0.078125 ,
   A sample size of 144 is needed to have probability
   0.8  of rejecting H0 at significance level alpha = 0.05
 *****************************************************************
```

Now we'll do the same thing with S. Below is a listing of the file `fpow2.R`. All it contains is the definition of the function `fpow2`. The name is chosen for similarity to the SAS program that does the same job. There is no S function `fpow` or `fpow1`.

```
fpow2 <- function(r,q,effsize,wantpow=0.80,alpha=0.05)
#############################################################################
# Power for the general multiple regression model, testing H0: C Beta = h   #
#       r       is the number of beta parameters                            #
#       q       Number rows in the C matrix = numerator df                  #
#       effsize is ncp/n, a squared distance between C Beta and h           #
#       wantpow is the desired power, default = 0.80                        #
#       alpha   is the significance level, default = 0.05                   #
#############################################################################
    {
    pow <- 0 ; nn <- r+1 ; oneminus <- 1 - alpha
    while(pow < wantpow)
        {
        nn <- nn+1
        phi <- nn * effsize
        ddf <- nn-r
        pow <- 1 - pf(qf(oneminus,q,ddf),q,ddf,phi)
        }#End while
    fpow2 <- nn
    fpow2  # Returns needed n
    }      # End of function fpow2
```

The function `fpow2` is made available to R by executing `fpow2.R` with the `source` command. Then the function is invoked, yielding the same answer of $n = 144$ that we got from SAS. Notice how the last two parameters are not specified; this has the effect of using the default values of 0.05 for the significance level and 0.80 for the desired power.

```
> source("fpow2.R")
> fpow2(r=4,q=3,effsize=0.078125)
[1] 144
```

Here is another one-way ANOVA example; this one is quick. Viewing the noncentrality parameter as a signal-to noise ratio, suppose we wanted to detect a ratio of $\frac{1}{10}$ with probability 0.80. For comparison, in the two-sample case, half a $\sigma$ difference between $\mu_1$ and $\mu_2$ corresponds to a signal $\frac{1}{16}$ as great as the background noise. Suppose we have equal sample sizes, and this time there are six treatments.

```
> source("fpow2.R")
> fpow2(r=6,q=5,effsize=0.10)
[1] 134
```

### 2.2.3 Factorial Experiments

This section is concerned with designed experiments having one or more categorical independent variables. All of the independent variables in the study must be manipulated; there can be no covariates or *measured* categorical independent variables. This case is limited, but still of practical interest. There are two main ways to do it. One is to calculate the effect size from some special-purpose formula, and then use the SAS program `fpow2.sas` or the S function `fpow2`. The key to getting such special-purpose formulas is given in the *Substitution Method* below.

The other way is to express one's hypothesis of interest as a collection of contrasts of the cell means, and to specify a set of non-zero values for one or more of these contrasts (in units of the common within-treatment standard deviation). With this information, the effect size can be calculated by the same program that searches for the required sample size. This method is explained in the *Matrix Approach*, which comes directly after the substitution method. The matrix approach is recommended for readers who want to just get in and do a power analysis quickly — but they must be comfortable with specifying effects in terms of contrasts.

**The Substitution Method** There has been some real benefit to using the special-purpose formulas (10) and (11) for the noncentrality parameter in the one-way case. Now we'll see where these formulas came from, and how to get similar expressions for any factorial design with fixed effects.

If expression (10) for the noncentrality parameter in one-way ANOVA looks familiar, it's no accident. The numerator is the standard elementary formula for the Between-Groups sum of squares in a one-way ANOVA, except with $\mu$ values substituted for $\overline{Y}$ values. This happens because formula (6) for the general $F$ statistic, and formula (7) for the noncentrality parameter are so similar. Basically, any re-expression of

$$(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{h}),$$

as a function of the sample cell means corresponds to a re-expression of

$$(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h}).$$

as a function of the population cell means — and it's the same function.

Therefore, to obtain a formula for the noncentrality parameter, all you have to do is locate a convenient formula for the $F$-test of interest. In the expression for the numerator sum of squares, replace sample cell means by population cell means. Then divide by $\sigma^2$.

Table 1: Cell Means in a 3 by 2 design

| | Level of B | | |
|---|---|---|---|
| Level of A | 1 | 2 | Average |
| 1 | $\mu_{11}$ | $\mu_{12}$ | $\mu_{1.}$ |
| 2 | $\mu_{21}$ | $\mu_{22}$ | $\mu_{2.}$ |
| 3 | $\mu_{31}$ | $\mu_{32}$ | $\mu_{3.}$ |
| Average | $\mu_{.1}$ | $\mu_{.2}$ | $\mu_{..}$ |

The result is a formula for the noncentrality parameter. This principle yields (10) without effort, and it applies to *any* $F$-test in *any* fixed-effects factorial ANOVA model.

The substition principle follows easily from a more general result in Scheffé's [4] *The analysis of variance*. The following is nearly a direct quote of his "Rule 1" on Page 39, except that the notation of the present report is employed.

> ***Scheffé's Rule 1****: Under the general linear regression model (5), the value of the noncentrality parameter $\phi$ of the noncentral $F$-distribution of the statistic $F^*$ may be calculated as follows: If, in the sum of squares in the numerator of $F^*$ each observation $Y_i$ is replaced by its expected value $E[Y_i]$, the result is $\phi\,\sigma^2$.*

Scheffé's Rule 1 is valid for *any* multiple regression model of the form (5), though its usefulness for sample size selection would seem to apply mostly to factorial ANOVA. Depending on what formula you are using for the $F$ statistic, it may sometimes be easier to use Scheffé's Rule than to substitute population means for sample means.

Here is an example. Suppose we have a two-factor design. Factor A has three levels, and factor B has two levels. The table of *population* cell means looks like this. Dots in the subscripts indicate averaging over the indicated dimension, so that for example, $\mu_{.2} = (\mu_{1,2} + \mu_{2,2} + \mu_{3,2})/3$.

Suppose we are interested in the interaction. Looking up the formula for SSAB in [3] and assuming equal sample sizes, we find the following on p. 822:

$$SSAB = n \sum_i \sum_j (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2. \tag{13}$$

The first remark about formula 13 is a warning. Here, the symbol $n$ refers to the common sample size in each of the six cells, not the total sample size. This illustrates an unpleasant feature of the Substitution Method. When you look up special-purpose formulas, you must take some care to know what the notation means. It's best to use a familiar text if you can.

Anyway, the substitution principle says that to get the noncentrality parameter, we just substitute population means for sample means and divide by $\sigma^2$. Writing the noncentrality parameter as sample size times effect size (now $n$ refers to the total sample size, as it should) we obtain

$$\phi = n \times \frac{1}{6} \frac{\sum_{i=1}^{3} \sum_{j=1}^{2} (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..})^2}{\sigma^2}. \tag{14}$$

Looking at (14), it's hard to see how *anyone* from *any* discipline could say what a reasonable effect size is. As far as I can tell, the only solution is to make up an effect, and apply the formula to it. Suppose that for A=1 and A=2, the population mean of $Y$ is a quarter of a standard deviation higher for B=2, but if A=3, the population mean of $Y$ is a quarter of a standard deviation higher for B=1. Of course there are infinitely many sets of means satisfying these constraints, even if they are expressed in standard deviation units. But they will all have the same effect size, and they will have the same pattern of non-parallel slopes when graphed. One such pattern is the following.

| | Level of B | | |
|---:|---:|---:|---:|
| Level of A | 1 | 2 | Average |
| 1 | 0.000 | 0.250 | 0.125 |
| 2 | 0.000 | 0.250 | 0.125 |
| 3 | 0.000 | -0.250 | -0.125 |
| Average | 0.000 | 0.083 | 0.042 |

Here us a bit of S code to calculate effect size for these values, using (14); remember, the noncentrality parameter is the product of sample size and effect size. The calculation assumes that population means are already expressed in units of the common within-cell standard deviation $\sigma$, and it preserves more decimal places of accuracy than the table above.

```
tab <- numeric(12) ; dim(tab) <- c(4,3)
tab[1,2] <- .25 ; tab[2,2] <- .25 ; tab[3,2] <- -.25
tab[4,1] <- mean(tab[1:3,1]) ; tab[4,2] <- mean(tab[1:3,2])
for(i in 1:4) tab[i,3] <- mean(tab[i,1:2])
tab
effsize <- 0
for(i in 1:3)
    {
    for(j in 1:2)
        {
        effsize <- effsize + (tab[i,j]-tab[i,3]-tab[4,j]+tab[4,3])^2
        } # Next j
    } # Next i
effsize <- effsize/6
cat("Effect size is ",effsize,"\n")
```

The effect size is 0.01388889. Now we'll use a modification of `fpow.sas` to search for a total sample size yielding power 0.80 assuming the 5% significance level. Notice that the program has been modified to increase in increments of six, since since there are six cells in the design and we are assuming equal sample sizes.

```
*****************  fpow2b.sas ************************************************/
options linesize = 79 pagesize = 100  noovp formdlim=’-’;   /*               */
data fpower;              /* Replace alpha, q, r, effsize and wantpow below  */
     alpha = 0.05;        /* Signif. level for testing H0: C Beta = h        */
     q = 2;               /* Numerator df = # rows in C matrix               */
     r = 6;               /* There are r beta parameters                     */
     effsize = 0.01388889;/* Effect size is ncp/n                            */
     wantpow = .80; /* Find n to yield this power                            */
     power = 0; n = r+6; oneminus = 1-alpha;  /* Initializing ...            */
/*****************************************************************************/
     do until (power >= wantpow);
        n = n+6 ;
        ncp = n * effsize;
        df2 = n-r;
        power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
     end;
     put ' ';
     put ' ************************************************************';
     put '   With ' r ' beta parameters, testing H0 of ' q ' linear';
     put '   restrictions on the betas and an effect size of ' effsize ',';
     put '   A sample size of ' n 'is needed to have probability ' ;
     put '   ' wantpow ' of rejecting H0 at significance level alpha = ' alpha;
     put ' ************************************************************';
     put ' ';
     put ' ';
```

Here is a piece of the log file, containing the output.

```
 **************************************************************
   With 6  beta parameters, testing H0 of 2  linear
   restrictions on the betas and an effect size of 0.01388889 ,
   A sample size of 702 is needed to have probability
   0.8  of rejecting H0 at significance level alpha = 0.05
 **************************************************************
```

Just for completeness, we'll do it with the S function `fpow2` as well. Note that unlike the SAS program, the S function has *not* been modified to yield a sample size that is a multiple of six.

```
> source("fpow2.R")
> fpow2(r=6,q=2,effsize=0.01388889,wantpow=0.80,alpha=0.05)
[1] 697
```

To preserve equal sample sizes, we would need to increase this value to the next multiple of six, yielding $n = 702$, the same answer we got from `fpow2b.sas` above. To attain the desired power, we need 117 observations in each cell.

Here is what we have learned from our examination of the Substitution Method. For any $F$-test in a factorial ANOVA, we can use a formula for the numerator sum of squares in the $F$-test to obtain a similar formula for the non-centrality parameter – and hence for effect size. Such formulas may provide valuable insight; they certainly do in the cases of the two-sample $t$-test and the one-way ANOVA. However, as a way of doing power analysis in proctice, there are some disadvantages to this approach.

- It is necessary to locate a special-purpose formula for the $F$-ratio's numerator sum of squares. For more obscure hypotheses, this can be frustrating and time-consuming.
- Once you have located a formula for the numerator sum of squares, you must take care to understand the notation. This also can be expensive in terms of time.
- Once you have a special-purpose formula for the noncentrality parameter, it may or may not be easy to interpret. If the formula is difficult to interpret (as in the example of a two-factor interaction), it may not be obvious what a reasonable effect size is. Frequently, it will be necessary to make up a plausible effect, and then use the formula to calculate the size of the effect.
- Calculating effect size from a special-purpose formula may be cumbersome for larger designs, and may involve careful work with a calculator or a bit of programming. Errors are possible, including subtle difficulties with rounding error.

For practical use, the matrix approach below may be better in many cases.

**A Matrix Approach** Special-purpose formulas like (10) may yield insight, but they are not always the most convenient way to calculate the noncentrality parameter in practice. Sometimes, it is preferable to use the general expression (7), and let software compute the noncentrality parameter as part of the power calculation. Here is an approach that is specialized for fixed-effects factorial ANOVA. Sample sizes need not be equal or proportional. For unbalanced designs, the approach assumes $F$-tests that are based on the equivalent of SAS's Type III sums of squares, in which every effect is tested controlling for all other effects in the model.

Consider a one-way design with $r$ cells. For multi-factor designs, create a single categorical independent variable whose values are the treatment combinations. That is, there will be $r$ treatment combinations. Tests for main effects, interactions and so on will correspond to tests on collections of *contrasts* of the $r$ cell means. More detail will be given presently.

We will use a dummy variable coding scheme with $r$ indicator dummy variables and no intercept. Specifically, let

$$x_{i,j} = \left\{ \begin{array}{ll} 1 & \mbox{if observation } i \mbox{ is in treatment group } j \\ 0 & \mbox{otherwise} \end{array} \right.$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, r$. Then the expected value of the response variable for an observation in treatment group $j$ is

$$E[Y_i] = \beta_1 x_{i,1} + \ldots + \beta_r x_{i,r} = \beta_j$$

This is the well-known *cell means* dummy variable coding scheme. See [3] for a detailed exposition. One advantage of using this dummy variable setup is that the $\mathbf{X}'\mathbf{X}$ matrix takes a particularly simple form; it is diagonal, with treatment sample sizes on the main diagonal. Denoting the number of observations in treatment $j$ by $n_j$, we have

$$\mathbf{X}'\mathbf{X} = \left[ \begin{array}{cccc} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_r \end{array} \right].$$

Let $f_j = \frac{n_j}{n}$ represent the *relative* sample size in treatment $j$. Then, multiplying and dividing by the total sample size $n$, expression (7) for the noncentrality parameter may be re-written as

$$\phi = n \times (\frac{\mathbf{C}\boldsymbol{\beta}-\mathbf{h}}{\sigma})'(\mathbf{C}\left[\begin{array}{cccc} 1/f_1 & 0 & \cdots & 0 \\ 0 & 1/f_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/f_r \end{array}\right]\mathbf{C}')^{-1}(\frac{\mathbf{C}\boldsymbol{\beta}-\mathbf{h}}{\sigma}) \qquad (15)$$

We may observe several things from (15).

- As usual,the noncentrality parameter is sample size times a quantity that we reluctantly call effect size.

- The null hypothesis is $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h}$. The vector $\mathbf{C}\boldsymbol{\beta}-\mathbf{h}$ represents an *effect* —- that is, a particular way in which the null hypothesis is wrong. This difference between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$ is naturally expressed in units of the common within-treatment standard deviation $\sigma$, and in general there is no reasonable way to avoid it.

- Of course, usually $\mathbf{h} = \mathbf{0}$.

- The quantity we are calling effect size is quite literally a squared distance between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$, scaled by $\sigma$, and in a space that is warped and stretched by the relative sample sizes. If the sample sizes are all equal and the contrasts are made to be orthoganal, it would be a multiple of squared Euclidian distance.

- To actually *do* a power analysis in this setting, all you need is a vector of *relative* sample sizes, and a vector of numbers representing the difference between $\mathbf{C}\boldsymbol{\beta}$ and $\mathbf{h}$ in units of $\sigma$ (the effect).

Setting up the matrix $\mathbf{C}$ merits a bit of discussion. Define a *contrast* as a linear combination whose coefficients (not all zero) add up to zero. We will test whether collections of contrasts of treatment means all equal zero. Thus, $\mathbf{h}$ will always be $\mathbf{0}$, and the rows of $\mathbf{C}$ will be the coefficients of $q$ contrasts. Of course $\boldsymbol{\beta}$ is a vector containing the treatment means.

Let's see how the usual tests for main effects and interactions would be represented in terms of contrasts. Look at the cell means in Table (2.2.3), used for the discussion of the Substitution Method. Stringing out the 6 treament combinations into a single factor and using cell means dummy variable coding, we have

$$\boldsymbol{\beta} = \left[\begin{array}{c} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{array}\right].$$

To test the main efects of Factor A, we test whether the three marginal means are equal. The null hypothesis could be written $H_0 : \frac{\mu_{11}+\mu_{12}}{2} = \frac{\mu_{21}+\mu_{22}}{2} = \frac{\mu_{31}+\mu_{32}}{2}$, which is equivalent to the two statements $\mu_{11}+\mu_{12}-\mu_{21}-\mu_{22}=0$ and $\mu_{21}+\mu_{22}-\mu_{31}-\mu_{32}=0$. Therefore, we could use the contrast matrix

$$\mathbf{C}=\left[\begin{array}{rrrrrr} 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 \end{array}\right].$$

Of course infinitely many contrast matrices are correct; and it *can* make a difference which one you use, since the magnitude of the effect can be affected. Since it's more natural to think of differences between marginal means than marginal sums, it might be better to use this contrast matrix for the main effect of factor A:

$$\mathbf{C}=\left[\begin{array}{rrrrrr} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{array}\right].$$

For the main effect of B, the contrast matrix has just one row.

$$\mathbf{C}=\left[\begin{array}{rrrrrr} \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \end{array}\right].$$

An interaction between A and B would mean that the effect of B depends on the value of A. The null hypothesis, then, would be that the effect of B does *not* depend on the level of A. In symbols, it could be written $H_0 : \mu_{11}-\mu_{12}=\mu_{21}-\mu_{22}=\mu_{31}-\mu_{32}$. The contrast matrix has two rows (one for each equals sign):

$$\mathbf{C}=\left[\begin{array}{rrrrrr} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right].$$

It should now be clear that for any ordinary hypothesis test in a factorial ANOVA, it is straightforward to set up the contrast matrix for carrying out the test of $H_o : \mathbf{C}\boldsymbol{\beta}=\mathbf{0}$.

The SAS program `matpow1.sas` uses `proc iml` to implement the matrix approach. Unlike `fpow.sas` and `fpow2.sas`, it writes results on the list file rather than the log file. Here is an example. For A=1 and A=2, the population mean of $Y$ is a quarter of a standard deviation higher for B=2, but if A=3, the population mean of $Y$ is a quarter of a standard deviation higher for B=1.

```
/*********************************  matpow1.sas ******************************/
title 'Power analysis for fixed-effects factorial ANOVA: Matrix approach';
options linesize = 79 pagesize = 100 noovp formdlim='-';
proc iml;
/********* Edit this input: Rows of matrices are separated by commas  ********/
     alpha = 0.05;
     wantpow = .80;
     f = {1,1,1,1};          /* Relative sample sizes       */
     C = { 0  0 -1  1,       /* Contrast matrix             */
           0 -1  1  0,
          -1  1  0  0};
     eff = {.25,.25,.25};    /* In standard deviation units */
/*****************************************************************************/
     r = nrow(f) ; q = nrow(eff); f = f/sum(f);

/*** Echoing input ***/
     print " Significance level: " alpha;
     print " Desired power: " wantpow;
     print " Relative sample sizes:" f;
     print " Contrast matrix "; print C;
     print " Effect matrix (in SD units): " eff;
     print "  ";
/*** Checking input ***/
```

```
     print " Checking input ...  ";
     inerr = 0; /* Input error = no. This may be changed below */
     if nrow(C) ^= q then do;
        print "Error: Length of eff must equal number of rows in C";
        inerr=1;
     end;
     if ncol(C) ^= r then do;
        print "Error: Length of f must equal number of cols in C";
        inerr=1;
     end;
     if eff`*eff = 0 then do;
        print "Error: eff can't be zero; this would cause an infinite loop!";
        inerr=1;
     end;
     aa = min(f##2);
     if aa = 0 then do;
        print "Error: No sample size may be zero.";
        inerr=1;
     end;

/*** Proceed if no input errors ***/
     if inerr = 1 then abort;
        else do;
        print "Input appears to be okay.  ";
             core = inv(C*inv(diag(f))*C`);
             effsize = eff`*core*eff;  print " Effect size = " effsize;
             power = 0; n = r+1; oneminus = 1-alpha;  /* Initializing ...*/
             do until (power >= wantpow);
                n = n+1 ;
                ncp = n * effsize;
                df2 = n-r;
                power = 1-probf(finv(oneminus,q,df2),q,df2,ncp);
             end; /*  End Loop */
             print " ";
             print " Required sample size is " n;
             print " ";
             print " ";
        end; /* End computation (Conditional on no input errors) */
```

Here is the list file.

```
-------------------------------------------------------------------------------

       Power analysis for fixed-effects factorial ANOVA: Matrix approach      1
                                                  14:04 Friday, August 30, 2002

                                                  ALPHA
                         Significance level:       0.05


                                             WANTPOW
                           Desired power:        0.8


                                                       F
                        Relative sample sizes:      0.25
                                                    0.25
                                                    0.25
                                                    0.25

                                Contrast matrix


                            C
                            0         0        -1         1
                            0        -1         1         0
                           -1         1         0         0


                                                        EFF
                    Effect matrix (in SD units):       0.25
```

```
                                        0.25
                                        0.25



             Checking input ...

         Input appears to be okay.


                        EFFSIZE
           Effect size =   0.078125





                                         N
       Required sample size is        144
```

We will presently see that the matrix approach is quite useful for larger designs and more obscure hypotheses, where it might be quite difficult to locate or derive a special purpose formula. But even in the one-way case, the martix approach has some advantages. Suppose you wanted to play with relative sample sizes, in the case of four equally spaced population means. As mentioned earlier, the optimal allocation is to split all the observations equally between treatments One and Four. But what if one went part of the way there – say, by giving half the sample to those two treatments? The program `matpow1b.sas` is identical to `matpow1.sas`, except for the line specifying relative sample sizes: beginverbatim f = 2,1,1,2; /* Relative sample sizes */ endverbatim Here is the output.

```
Power analysis for fixed-effects factorial ANOVA: Matrix approach      1
                                     14:44 Friday, August 30, 2002

                                     ALPHA
             Significance level:       0.05


                        WANTPOW
           Desired power:        0.8


                                          F
            Relative sample sizes: 0.3333333
                                   0.1666667
                                   0.1666667
                                   0.3333333

                 Contrast matrix


              C
              0          0        -1          1
              0         -1         1          0
             -1          1         0          0


                                         EFF
        Effect matrix (in SD units):    0.25
                                        0.25
                                        0.25



             Checking input ...

         Input appears to be okay.
```

```
                                    EFFSIZE
                        Effect size =  0.0989583



                                                     N
                 Required sample size is           115
```

Thus, with these unequal sample sizes, we can obtain the same power with 26 fewer subjects, or 18%. Of course, this assumes you can guess which means will be highest and lowest.

### 2.2.4 The General Case: One or more independent variables random

When there are more than two treatments, 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 $\mu_1$ with the average of $\mu_2$ and $\mu_3$. Further, suppose that $\mu_2 = \mu_3 \neq \mu_1$ ($sigma^2$ 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. This seems silly, but it's rational given the (silly) assuption that we have *certain knowledge* of the true parameter values.

Effect size is

$$\Delta^2 = \frac{(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})'(\mathbf{C}(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h})}{\sigma^2}, \tag{16}$$

Exercises?

1. Carry out the details necessary to derive (8) from (7)
2. Cell means coding for 2-sample
3. Effect coding for 2-sample
4. critical value is a function of n too – why do you still know that power is an increasing function of n? Answer - critical value is decreeasing.
5. 
6. 
7. 
8. Do the 6-group with SAS.
9. 
10. 
11.

# 3 The Sample Variation Method

# 4 The Population Variation Method

# 5 The Dummy Population Method

## References

[1] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd Edition), Hilsdale, New Jersey: Erlbaum.

[2] Lohr, S. (1999) *Sampling: Design and analysis.* Pacific Grove, California: Brooks/Cole.

[3] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.

[4] Scheffé, H. (1959) *The analysis of variance.* New York: Wiley.