Quantitative by Quantitative

An interaction of two quantitative variables is literally represented by their product. For example, consider the model

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Hold x_2 fixed at some particular value, and re-arrange the terms. This yields

$$E[Y] = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1.$$

so that there is a linear relationship between x_1 and E[Y], with both the slope and the intercept depending on the value of x_2. Similarly, for a fixed value of x_1,

$$E[Y] = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2,$$

and the (linear) relationship of x_2 to E[Y] depends on the value of x_1. We always have this kind of symmetry.

Three−way interactions are represented by 3−way products, etc. Its interpretation would be "the 2−way interaction depends ..."

Product terms represent interactions ONLY when all the variables involved and all lower order interactions involving those variables are also included in the model!

Categorical by Categorical

It is no surprise that interactions between categorical independent variables are represented by products. If A and B are categorical variables, IVs representing the A by B interaction are obtained by multiplying each dummy variable for A by each dummy variable for B. If there is a third IV cleverly named C and you want the 3–way interaction, multiply each of the dummy variables for C by each of the products representing the A by B interaction. This rule extends to interactions of any order.

Up till now, we have represented categorical independent variables with indicator dummy variables, coded 0 or 1. If interactions between categorical IVs are to be represented, it is much better to use "effect coding," so that the regression coefficients for the dummy variables correspond to main effects. (In a 2–way design, products of indicator dummy variables still correspond to interaction terms, but if an interaction is present, the interpretation of the coefficients for the indicator dummy variables is not what you might guess.)

Effect coding. There is an intercept. As usual, a categorical independent variable with k categories is represented by k–1 dummy variables. The rule is

Dummy var 1: First value of the IV gets a 1, last gets a minus 1, all others get zero.
Dummy var 2: Second value of the IV gets a 1, last gets a minus 1, all others get zero.

. . .

Dummy var k–1: k–1st value of the IV gets a 1, last gets a minus 1, all others get zero.

Here is a table showing effect coding for Plant from the Greenhouse data.

Country	p1	p2	$E[Y] = \beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

It is clear that $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$, so it's a valid dummy variable coding scheme even though it looks strange.

Country	p1	p2	$E[Y] = \beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Effect coding has these properties, which extend to any number of categories.

- $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$.
- The average population mean (grand mean) is $(\mu_1+\mu_2+\mu_3)/3 = \beta_0$.
- β_1, β_2 and $-(\beta_1+\beta_2)$ are deviations from the grand mean.

The real advantage of effect coding is that the dummy variables behave nicely when multiplied together, so that main effects correspond to collections of dummy variables, and interactions correspond to their products -- in a simple way. This is illustrated for Plant by MCG analysis, using the full greenhouse data set).

```
data nasty;
     set yucky;
     /* Two dummy variables for plant */
        if plant=. then p1=.;
        else if plant=1 then p1=1;
        else if plant=3 then p1=-1;
        else p1=0;
     if plant=. then p2=.;
        else if plant=2 then p2=1;
        else if plant=3 then p2=-1;
```

```
      else p2=0;
   /* Five dummy variables for mcg */
   if mcg=. then f1=.;
      else if mcg=1 then f1=1;
      else if mcg=9 then f1=-1;
      else f1=0;
   if mcg=. then f2=.;
      else if mcg=2 then f2=1;
      else if mcg=9 then f2=-1;
      else f2=0;
   if mcg=. then f3=.;
      else if mcg=3 then f3=1;
      else if mcg=9 then f3=-1;
      else f3=0;
   if mcg=. then f4=.;
      else if mcg=7 then f4=1;
      else if mcg=9 then f4=-1;
      else f4=0;
   if mcg=. then f5=.;
      else if mcg=8 then f5=1;
      else if mcg=9 then f5=-1;
      else f5=0;
   /* Product terms for the interaction */
      p1f1 = p1*f1; p1f2=p1*f2 ; p1f3=p1*f3 ; p1f4=p1*f4; p1f5=p1*f5;
      p2f1 = p2*f1; p2f2=p2*f2 ; p2f3=p2*f3 ; p2f4=p2*f4; p2f5=p2*f5;

proc reg;
   model meanlng = p1 -- p2f5;
   plant:  test p1=p2=0;
   mcg:    test f1=f2=f3=f4=f5=0;
   p_by_f: test p1f1=p1f2=p1f3=p1f4=p1f5=p2f1=p2f2=p2f3=p2f4=p2f5 = 0;
```

Here is the output from the test statement. For comparison, it is followed by `proc glm` output from `model meanlng = plant|mcg`.

```
Dependent Variable: MEANLNG
Test: PLANT    Numerator: 110847.5637  DF:    2   F value: 113.9032
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: MCG      Numerator:  11748.0529  DF:    5   F value:  12.0719
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: P_BY_F   Numerator:   4758.1481  DF:   10   F value:   4.8893
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLANT	2	221695.12747	110847.56373	113.90	0.0001
MCG	5	58740.26456	11748.05291	12.07	0.0001
PLANT*MCG	10	47581.48147	4758.14815	4.89	0.0001

It worked.

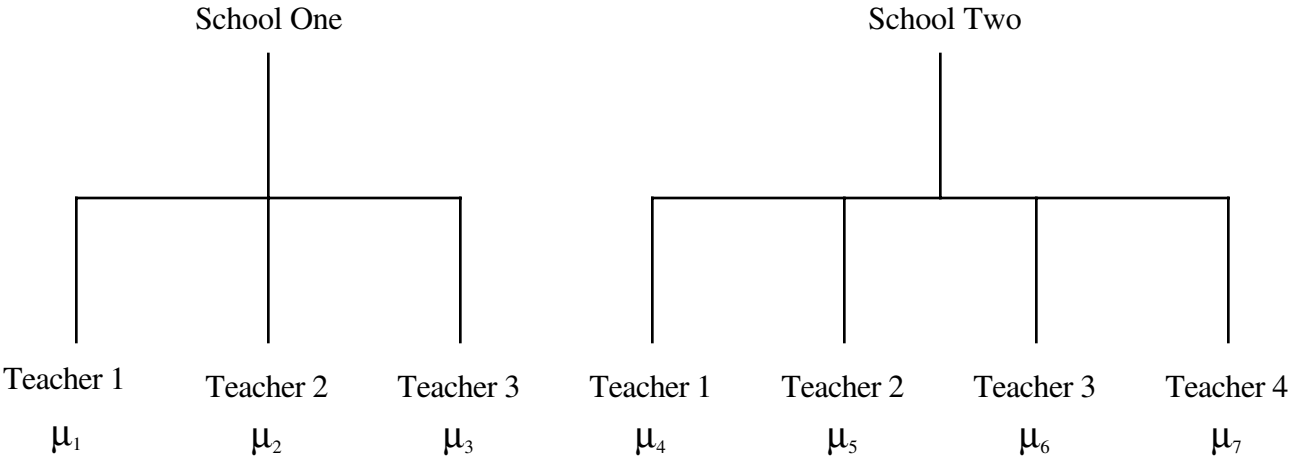
Effect coding works as expected in conjunction with quantitative independent variables. In particular, products of quantitative and indicator variables still represent interactions. In fact, the big advantage of effect coding is that you can use it to test categorical independent variables, and interactions between categorical independent variables -- in a bigger multiple regression context.

Nested and Random Effect models

Nested Designs

Suppose a chain of commercial business colleges is teaching a software certification course. After 6 weeks of instruction, students take a certification exam and receive a score ranging from zero to 100. The owners of the business school chain want to see whether performance is related to which school students attend, or which instructor they have -- or both. They compare two schools; one of the schools has three instructors teaching the course, and the other school has 4 instructors teaching the course. A teacher only works in one school.

There are two independent variables, school and teacher. But it's not a factorial design, because ``Teacher 1" does not mean the same thing in School 1 and School 2; it's a different person. This is called a **nested** design. By the way, it's also **unbalanced**, because there are different numbers of teachers withing each school. We say that *teacher is nested within school.* The diagram below shows what is going on, and give a clue about how to conduct the analysis.



To compare schools, we want to test $\frac{1}{3}(\mu_1+\mu_2+\mu_3) = \frac{1}{4}(\mu_4+\mu_5+\mu_6+\mu_7)$.

To compare instructors within schools, we want to test $\mu_1=\mu_2=\mu_3$ and $\mu_4=\mu_5=\mu_6=\mu_7$ simultaneously.

The first test involves one contrast of μ_1 through μ_7; the second test involves five contrasts. There really is nothing to it.

You can do it with `proc reg` and cell means coding, or you can take advantage of `proc glm`'s syntax for nested models.

```
proc glm;
     class school teacher;
     model score = school teacher(school);
```

The notation `teacher(school)` should be read ``teacher within school."

- It's easy to extend this to more than one level of nesting. You could have climate zones, lakes within climate zones, fishing boats within lakes, ...

- There is no problem with combining nested and factorial structures. You just have to keep track of what's nested within what. Factors that are not nested are sometimes called ``crossed."

Random Effect Models The preceding discussion (and indeed, the entire course to this point) has been limited to ``fixed effects" models. In a **random effects** model, *the values of the categorical independent variables represent a random sample from some population of values*. For example, suppose the business school had 200 branches, and just selected 2 of them at random for the investigation. Also, maybe each school has a lot of teachers, and we randomly sampled teachers within schools. Then, teachers within schools would be a random effects factor too.

It's quite possible to have random effect factors and fixed effect factors in the same design; such designs are called ``mixed." SAS `proc mixed` is built around this, but it does a lot of other things too.

Nested models are often viewed as random effects models, but there is no necessary connection between the two concepts. It depends on how the study was conducted. Were the two schools randomly selected from some population of schools, or did someone just pick those two (maybe because there are just two schools)?

Of course lots of the time, nothing is randomly selected -- but people use random effects models anyway. Why pretend? Well, sometimes they are thinking that in a better world, lakes *would* have been randomly selected. Or sometimes, the scientists are thinking that they really would like to generalize to the entire population of lakes, and therefore should use statistical tools that support such generalization -- even if there was no random sampling. (By the way, no statistical method can compensate for a biased sample.) Or sometimes it's just a tradition in certain sub-areas of research, and everybody expects to see random effects models.

In the traditional analysis of models with random or mixed effects and a normal assumption, F-tests are often possible, but they don't always use Mean Squared Error in the denominator of the F statistic. Often, it's the Mean Square for some interaction term or other. The choice of what error term to use is relatively mechanical for balanced models with equal sample sizes, but even then, sometimes (especially when it's a mixed model) a valid F-test for an effect of interest just doesn't exist.

When the design is unbalanced or has unequal sample sizes, a valid F-test rarely exists. It's a real pain. Sometimes, you can find an error term that produces a valid F-test *assuming* that some interaction (or maybe more than one interaction) is absent. Usually, you can't test for that interaction either. But people do it anyway and hope for the best.

SAS `proc mixed` goes a long way toward solving these problems. It's a great piece of software, based on recent, state-of the-art research as well as more venerable stuff. But we're running out of time. Goodbye, `proc mixed`. Goodbye, random effects.

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 can 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 significant results. Power is a function of 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 α; 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 $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$.

- 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

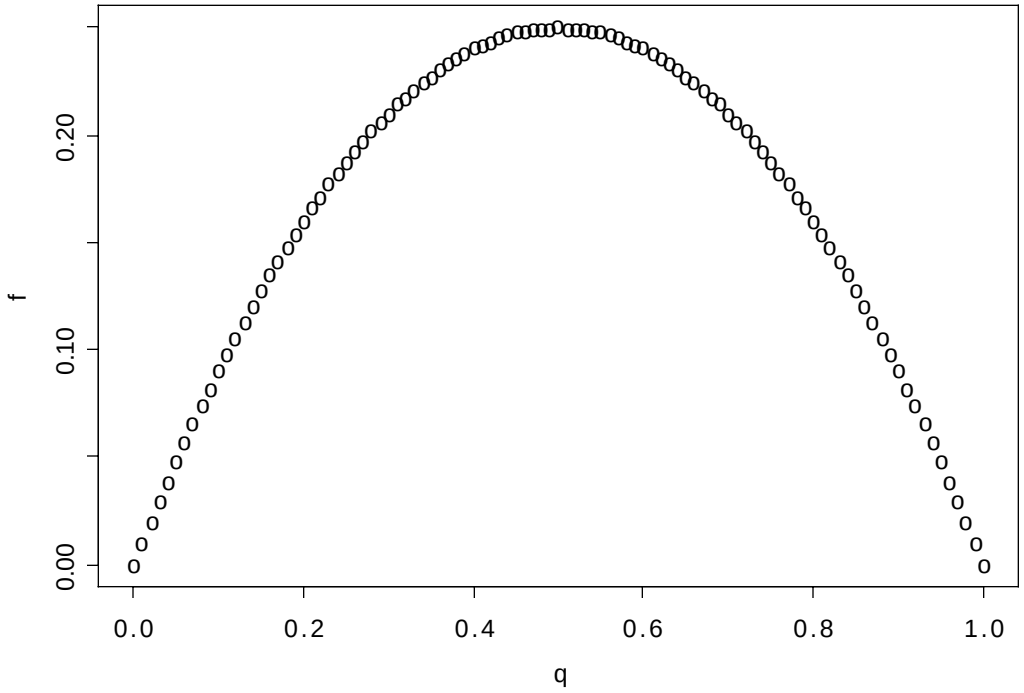
$$\phi = n\ q(1-q)\ \delta^2. \qquad (4.4)$$

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 -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.

The Sample Variation Method

There are at least two main meanings of ``significance." One is statistical significance, and another is *explanatory* significance in the sense of explained variation. Formula (3.4) from Chapter 3 is relevant. It is reproduced here.

$$F = \left(\frac{n - p}{s}\right) \frac{a}{1 - a}, \qquad (3.4)$$

where, after controlling for the effects in a reduced model, a is the proportion of the *remaining* variation that is explained by the full model.

Formula (3.4) tells us that the two meanings of ``significance" need not coincide, since statistical significance can come from either strong results or from a large sample. The sample variation method can be viewed as a way of bringing the two types of significance into agreement. It's not really a power analysis, but it is a rational way to decide on sample size.

In equation (3.4), F is an increasing function of both n and a, so its p-value (the tail area beyond F) is a decreasing function of both n and a. The sample variation method is to choose a value of a that is just large enough to be interesting, and increase n, calculating F and its p-value each time until $p < 0.05$; then stop. The final value of n is the smallest sample size for which an effect explaining that much of the remaining variation will be significant. With that sample size, the effect will be significant if and only if it explains a or more of the remaining variation.

That's all there is to it. You tell me a proportion of remaining variation that you want to be significant, and I'll tell you a sample size. In exchange, you agree not to moan and complain and go fishing for more covariates if your results are almost significant, because they were too weak to be interesting anyway.

There are two questions you might want to ask.

- For a given proportion of the remaining variation, what sample size do I need for statistical significance?

- For a given sample size, what proportion of the remaining variation do I need for statistical significance?

To make things more definite, let us suppose we are contemplating a 2x3x4 analysis of covariance, with two covariates and factors cleverly named A, B and C. We are setting it up as a regression model, with one dummy variable for A, 2 dummy variables for B, and 3 for C. Interactions are represented by product terms, and there are 2 products for the AxB interaction, 3 for AxC, 6 for BxC, and 1*2*3 = 6 for AxBxC. The regression coefficients for these plus two for the covariates and one for the intercept give us p = 26. The null hypothesis is that of no BxC interaction, so s = 6. The "other effects in the model" for which we are "controlling" are represented by 2 covariates and 17 dummy variables and products of dummy variables.

First, let's find out what sample size we need for the interaction to be significant provided it explains at least 10% of the remaining variation after controlling for other effects in the model. This is accomplished by the program `sampvar1.sas`. It is a little unusual in that it uses the SAS `put` statement to write results to the *log* file. It never produces a list file, because there is no `proc` step.

```
/************************  sampvar1.sas ************************/
/*      Finds n needed for significance, for a given proportion of */
/*      remaining variation                                       */
/*****************************************************************/

options linesize = 79 pagesize = 200;
data explvar;     /* Can replace alpha, s, p, and a below.   */
   alpha = 0.05;  /* Significance level.                     */
   s = 6;         /* Numerator df = # IVs being tested.      */
   p = 26;        /* There are p beta parameters.            */
   a = .10  ;     /* Proportion of remaining variation after */
                  /* controlling for all other variables.    */

   /* Initializing ... */  pval = 1; n = p+1;
   do until (pval <= alpha);
      F = (n-p)/s * a/(1-a);
      df2 = n-p;
      pval = 1-probf(F,s,df2);
      n = n+1 ;
   end;
   /* When finished, n is one too many */
   n = n-1; F = (n-p)/s * a/(1-a); df2 = n-p;
   pval = 1-probf(F,s,df2);

   put ' ******************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables controlling for the others,';
   put '  a sample size of ' n 'is needed for significance at the';
   put '  alpha = ' alpha 'level, when the effect explains a = ' a ;
   put '  of the remaining variation after allowing for all other ';
   put '  variables in the model. ';
   put '  F = ' F ',df = (' s ',' df2 '), p = ' pval;
   put ' ';
   put ' ******************************************************';
```

Here is the part of the log file produced by the put statements.

```
 ******************************************************

  For a multiple regression model with 26 betas,
  testing 6  variables controlling for the others,
  a sample size of 144 is needed for significance at the
  alpha = 0.05 level, when the effect explains a = 0.1
  of the remaining variation after allowing for all other
  variables in the model.
  F = 2.1851851852 ,df = (6 ,118 ), p = 0.0491182815

 ******************************************************
```

Suppose you were considering n=120, and you wanted to know what proportion of the remaining variation the interaction must explain in order to be significant. This is accomplished by `sampvar2.sas`.

```
/*************************  sampvar2.sas ***************************/
/*  Finds proportion of remaining variation needed for significance, */
/*  given sample size n                                              */
/*******************************************************************/

options linesize = 79 pagesize = 200;
data explvar;     /* Replace alpha, s, p, and a below.  */
   alpha = 0.05;  /* Significance level.                */
   s = 6;         /* Numerator df = # IVs being tested. */
   p = 26;        /* There are p beta parameters.       */
   n = 120  ;     /* Sample size                        */

   /* Initializing ... */  pval = 1; a = 0; df2 = n-p;
   do until (pval <= alpha);
      F = (n-p)/s * a/(1-a);
      pval = 1-probf(F,s,df2);
      a = a + .001 ;
     end;
  /* When finished, a is .001 too much */
   a = a-.001; F = (n-p)/s * a/(1-a); pval = 1-probf(F,s,df2);

   put ' ******************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables at significance level ';
   put '  alpha = ' alpha ' controlling for the other variables,';
   put '  and a sample size of ' n', the variables need to explain';
   put '  a = ' a ' of the remaining variation to be significant.';
   put '  F = ' F ', df = (' s ',' df2 '), p = ' pval;
   put '   ';
   put ' ******************************************************';
```

And here is the output.

```
*****************************************************

 For a multiple regression model with 26 betas,
 testing 6  variables at significance level
 alpha = 0.05  controlling for the other variables,
 and a sample size of 120 , the variables need to explain
 a = 0.123  of the remaining variation to be significant.
 F = 2.1972633979 , df = (6 ,94 ), p = 0.0499350803

*****************************************************
```

It's worth mentioning that the Sample Variation method is so simple that lots of people must know about it -- but I have never seen it described in print.

The Population Variation Method

This is a method of sample size selection for multiple regression due to Cohen (1988). It combines elements of classical power analysis and the sample variation method. Cohen does not call it the ``Population Variation Method;" he calls it ``Statistical Power Analysis." For most research psychologists, the population variation method *is* statistical power analysis, period.

Cohen's book on power has the curious property that if a statistician and a scientist both read it, the statistician will likely come away more confused than the scientist. I think this happened Cohen has worked extremely hard to translate statistical concepts into language that can be understood by non-statisticians, and in the process has incorporated some very good ideas of his own (and maybe some bad ideas, too), while providing exactly the same flavor of intuitive justification for the standard concepts and the ones he has invented. For the scientist, everything flows and makes sense. For the statistician, it's a lot harder to follow than a more mathematical discussion.

The basic idea is this. Looking closely at the formula for the non-centrality parameter ϕ, Cohen decides that it is based on something interprets as a *population* version of the quantity we are denoting by *a*. That is, one thinks of it as the proportion of remaining variation (Cohen uses the term variance instead of variation) that is explained by the effect in question -- in the population. He calls it ``effect size."

Just a comment: Of course the problem of comparing two means is a special case of multiple regression, but ``effect size" in the population variation method does not reduce to the traditional definition of effect size for the two-sample t-test with equal variances. In fact, effect size in the population variation method mixes the effect together with the design in such a way that they cannot be separated (by the way, this is true of the sample variation method too).

Still, from a so-called ``effect size" and a sample size, it's easy to calculate a non-centrality parameter, and then you can compute power, and increase the sample size until the power is as high as you wish. For most people, most of the time, it's a lot easier to think about proportions of explained variation than to think about collections of non-zero contrasts in units of σ. Plus, it applies to regression models in general, not just factorial ANOVA. To do a classical power analysis with observational data, you need the joint probability distribution of all the observed independent variables (which are presumably independent of any manipulated independent variables). Cohen's method is a lot easier, even if it is a bit murky. Here's a program that does it.

```
/**********************  popvar.sas ****************************/
options linesize = 79 pagesize = 200;
data fpower;        /* Replace alpha, s, p, and wantpow below    */
     alpha = 0.05;  /* Significance level                        */
     s = 6;         /* Numerator df = # IVs being tested         */
     p = 26;        /* There are p beta parameters               */
     a = .10  ;     /* Effect size                               */
     wantpow = .80; /* Find n to yield this power.               */
     power = 0; n = p+1; oneminus = 1-alpha; /* Initializing ... */
     do until (power >= wantpow);
        ncp = (n-p)*a/(1-a);
        df2 = n-p;
        power = 1-probf(finv(oneminus,s,df2),s,df2,ncp);
        n = n+1 ;
     end;
     n = n-1;
     put ' ********************************************************';
     put '   ';
     put '   For a multiple regression model with ' p 'betas, ';
     put '   testing ' s 'independent variables using alpha = ' alpha ',';
     put '   a sample size of ' n 'is needed';
     put '   in order to have probability ' wantpow 'of rejecting H0';
     put '   for an effect of size a = ' a ;
     put '   ';
     put ' ********************************************************';
```

```
*********************************************************

  For a multiple regression model with 26 betas,
  testing 6 independent variables using alpha = 0.05 ,
  a sample size of 155 is needed
  in order to have probability 0.8 of rejecting H0
  for an effect of size a = 0.1

*********************************************************
```

For comparison, when we specified a *sample* proportion of remaining variation equal to 10%, a sample size of 144 was required for significance. Getting into the spirit of the population variation method, we can talk about it like this. If the *population* effect size is 0.10 and n=155, then with 80% probability we'll get a *sample* effect size large enough for significance. How big does the sample effect size have to be? Running `sampvar2.sas`, it turns out that with n=155, you need a sample a=0.092 for significance. So if a=0.10 in the population and n=155, the probability that the sample a > 0.092 = 0.80.

Loosely speaking, that is.