

# **An Introduction to Statistical Power Calculations for Linear Models with SAS 9.1**

Robin High, University of Oregon

## **Abstract**

In Version 9.1 SAS® introduced two new procedures that compute power under a variety of statistical designs which are very effective study planning tools. In particular, this paper focuses on how SAS computes power for continuous response data. PROC POWER computes power for T-tests, one-way ANOVAs, correlations, and regression models. PROC GLMPower computes power for ANOVA and ANCOVA models with one or more between-subject factors and continuous covariates. PROC MIXED is shown to be yet another very effective study planning tool as data inputs can be manipulated to generate power calculations for comparisons of means from linear models of greater complexity, such as repeated measurements. The essential points of prospective power analysis will also be reviewed and contrasted with some of the fallacies that underlie retrospective power analysis.

## **Essential Concepts of Power for Review**

When planning a study or an experiment, for each research question submitted to a statistical test a null hypothesis of no difference (HO) and an alternative hypothesis that a “meaningful” difference exists (HA) are presented. For example, under the design of two independent groups with subjects randomly assigned to each group, to test the equality of the two population means for a continuous response variable with a T-test the relevant hypotheses are:

HO: the two population means are equal

HA: the two population means are different

The “ideal conditions” for the two-sample T-test are independence of observations, equal variances across the two groups, and normality of the residuals. Power is defined as “the probability that the significance test will reject the null hypothesis for a specified value of the alternative hypothesis.”

The components which go into a power analysis begin with a statement of the study’s objectives, followed by specification of a statistical model, from which an appropriate test statistic is computed. The inputs for power analyses then include the following essential components:

1. Significance level, the probability of a Type I error:  $\alpha = \text{PROB}(\text{Reject HO} \mid \text{HO is true})$ .

Common, yet arbitrary, choices are  $\alpha = .05$  or  $.01$ .

2. Power to detect a difference is expressed as  $(1 - \beta)$  where  $\beta$  is the probability of a Type II error) [that is,  $\beta = \text{Prob}(\text{Fail to reject HO} \mid \text{HA is true})$ :  $\text{power} = \text{PROB}(\text{Reject HO} \mid \text{HA is true})$ . Power of 0.80 and 0.90 are common, yet arbitrary, choices.

3. The effect size is the minimum difference of interest between the two means adjusted by the standard deviation. This difference represents the “smallest” effect of clinical or substantive significance, and the magnitude of its value depends on the study’s objectives.
4. The sample size is the number of subjects to be studied either computed for each group (NperGroup) or total number to be studied summed across groups (Ntotal).

These four components of a power analysis are not independent: in fact, the specification of any three of them automatically determines the fourth. The usual objective of a power analysis is to calculate the sample size (4) required to satisfy values given for (1)-(3). It can also be utilized in studies with limited resources where the maximum total sample size (4) is known. In this situation power analysis becomes a helpful tool to determine if sufficient power exists (2) or what effect size (3) can be detected for specified values of the other components. As a result, the researcher can evaluate whether the study is worth pursuing.

Since every experimental situation is different, the values of  $\alpha$  and power= $1-\beta$  should be set regarding the consequences of the Type-I and Type-II errors. A Type-I error,  $\alpha$ , is usually considered to be a more serious problem, so it is usually set smaller than the Type-II error,  $\beta$ . However, if the study is exploratory in nature, then  $\alpha$  can be set to a larger value (e.g.,  $\alpha=.10$ ); if major changes would result because  $H_0$  is rejected, especially regarding any impact on public safety, then  $\alpha$  should be smaller (e.g.,  $\alpha=.01$ ). Power can be increased to 0.90 (i.e.,  $\beta=0.10$ ) if the need to detect a meaningful difference, should it exist, is very important. Of course, you can continue to protect yourself by decreasing  $\alpha$  and increasing power, yet there comes a point where the required sample size will quickly overwhelm the situation making results very expensive or time consuming to achieve.

Effect size (3) depends on the statistical design and parameters of the model; the most common ones are summarized in Cohen (1988, 1992) and Cortina (2000): they are functions of means and variances for t-tests and ANOVAs, R-Square values for linear regression models, or transformations of proportions and correlations. The chosen values of these parameters may be determined from the researcher’s experience or by utilizing data from existing studies. What will generally be avoided in this discussion (and the way SAS computes power in general) are canned effect sizes: small, medium, or large. To really understand what an effect size means, you need to interpret it in the framework of the actual values of population means and variances of the data under consideration and what differences are of interest in them for the proposed study.

### **Power Calculations in SAS Version 9.1**

Suppose you want to compute the total number of subjects required to test the equality of population means from two independent groups (A and B) and you plan to utilize the two-sample T-test. For the purposes of this article, the statistical design will assume equal group sizes (they can also be unequal). The response variable,  $y$ , is normally distributed within each group which have means  $\mu_A$  and  $\mu_B$  respectively, and both groups have a common standard deviation. You plan to make a two-sided test with the following hypotheses for the population means:

$$H_0: \mu_A - \mu_B = 0$$

$$H_A: \mu_A - \mu_B \neq 0$$

The alternative hypothesis,  $H_A$ , is specified to be a two-sided test when deviations in either direction from 0 would be important to determine. For a power analyses you can also specify one-sided tests where not equals  $\neq$  is replaced with a 1-sided comparison symbol, less than,  $<$ , or greater than,  $>$ .

The typical objective is to determine the sample size required for each group such that the probability of obtaining a test statistic equal to or larger than a critical value is  $\alpha=.05$  under  $H_0$  and power=0.9 for a specified effect size (i.e., a value which belong to  $H_A$ ).

Calculating the total sample size is relatively simple when assuming normality of residuals and you want to compare two treatment averages in an independent (completely randomized) groups design. Assuming a common variance of an observation in both groups:  $\sigma^2$ . The standard deviation of the difference of two independent observations is the square root of the sum of the two variances:

$$\sigma_{dif} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{2 * \sigma^2} = \sqrt{2} * \sigma$$

The formula for the required sample size for each group utilizes the standard deviation of the difference which is expressed as:

$$\begin{aligned} NperGroup &= (\sqrt{2} * \sigma / \Delta)^2 * (z_{a2} + z_b)^2 \\ &= (\sigma_{dif} / \Delta)^2 * (z_{a2} + z_b)^2 \end{aligned}$$

where  $\sigma$  is the standard deviation of a single observation and  $\Delta$  is the difference in the two population means to be detected ( $\Delta = \mu_A - \mu_B$ ). For a two-sided test,  $z_{a2}$  is the critical value of  $z$  corresponding to  $\alpha/2$ , that is, with  $\alpha=.05$ ,  $z=1.96$  so that 0.025 area is assigned to each tail. The value for  $z_b$  is set according to the Type-II error,  $\beta$ . The sample size for each group can then be calculated in a SAS DATA step:

```
DATA nperGroup;
sigma=4;          * the magnitude of random variation (std. deviation);
delta=2;          * the minimum treatment difference of interest;
alpha=.05;        * the probability of a type I error;
beta=.0968;       * probability of a type II error chosen to produce integer NperGroup;
power=1-beta;     * power to detect the specified difference;
sides=2;          * 1 or 2 sided test;
z_a=ABS(PROBIT(alpha/sides));
z_b=ABS(PROBIT((1-beta)));
sigma_dif = sigma*SQRT(2); * standard deviation of a difference;
NperGroup = CEIL( ((sigma_dif/delta)**2) * ((z_a + z_b)**2));
NTotal=2*NperGroup;
RUN;
```

```
PROC PRINT NOobs; RUN;
```

| sigma | delta | alpha | beta   | power  | sides | z_a   | z_b   | Nper<br>Group | NTotal |
|-------|-------|-------|--------|--------|-------|-------|-------|---------------|--------|
| 4     | 2     | 0.05  | 0.0968 | 0.9032 | 2     | 1.960 | 1.300 | 86            | 172    |

Computing technology for power analyses is now available with two new procedures in SAS Version 9.1. This paper will first introduce how power for independent group ANOVA designs such as the one just demonstrated can easily be computed with PROC POWER and PROC GLMPower. Power computed from more unusual designs can be addressed with PROC MIXED. These procedures are designed to provide power results to assist you with study planning.

### Power Computations with PROC POWER

PROC POWER calculates power for many statistical designs including one- and two-sample t-tests, correlations, proportions, regression models, and one-way ANOVAs, among others. For example, PROC POWER will easily compute power for the design just illustrated. It consists of one introductory statement (no input dataset is required) followed by a statement with a keyword which specifies the statistical test. Included on it are your choices of the relevant inputs.

```
PROC POWER;
TWOSAMPLEMEANS
  TEST = diff          /* Detect a difference in means */
  ALPHA = .05          /* Significance level */
  SIDES = 2            /* two-sided test */
  MeanDiff = 2         /* muA - muB : difference in population means */
  STDDEV = 4           /* Standard deviation of an observation */
  NperGroup = .        /* NTotal = 2 * NperGroup */
  POWER = .9           /* desired power */
;
RUN;
```

If you want to compute power for unequal group sizes, the GROUPWeights option can be entered. You also need to replace NperGroup= with the Ntotal= :

```
GROUPWeights = (2 1) /* Group A is twice as large as Group B */
NTotal = .           /* Compute NTotal = n_grp_A + n_grp_B */
```

The effect size for the difference in two population means can be inferred from the values placed on the POWER statement (MeanDiff = 2) divided by the standard deviation of a single response (STDDEV = 4):

$$\text{Effect Size} = (\mu_A - \mu_B) / \sigma = (12 - 10) / 4 = 0.5$$

The effect size of 0.5 is the difference in the two group means divided by the common standard deviation of an observation. Although this value has been defined as a medium effect size (known as Cohen's d,

Cohen, 1988), throughout this paper its value is determined by the desired difference in means and the population standard deviation. With regards to computing power, SAS assumes the effect size is expressed in terms of the parameters of interest, not merely canned values of small, medium, and large. The difference in means is the actual difference in population means, whatever their individual magnitudes, e.g., a difference of  $12-10=2$  will produce the same result as  $25-23=2$  or  $2-0=2$ .

The required components of a power analysis are included among the options for the TWOSAMPLEMEANS statement with one of them set to “missing” (in SAS the period is the usual missing value entered for numbers). By replacing any one of the numbers specified above with a period and entering relevant values for the other options, you can solve for the missing item. In the example given above, the missing item is the group sample size (NperGroup = . ) with the following sample size:

```
Computed N Total

Actual
Power      NperGroup

0.903      86
```

Thus, the total sample size required to meet the specified inputs is  $N_{Total}=172$ , since 86 subjects are computed for each group. The actual power printed on the output is 0.903 which is slightly higher than the specified or nominal value of  $power=.90$  on the TWOSAMPLEMEANS statement, since an integer for the number of subjects in each group is required and these two numbers add to 172. The procedure rounds the computed sample size to the next larger integer so  $N_{perGroup}=86$  increases the actual power of the study slightly (the actual total sample size to achieve  $power=.90$  is  $N_{perGroup} = 85.03$ ).

Typically, you will solve for either power (as demonstrated above) or by entering the group sample size as an actual number (e.g.,  $N_{perGroup} = 86$ ) and placing the period for  $POWER = .$  in these statements. In this latter scenario, power is computed as 0.903.

PROC POWER allows you to enter multiple values of each parameter for each option on the keyword statement. For example, you can enter two values for the option

```
ALPHA = .05 .01
```

which will compute results for these two levels of alpha. You can also enter multiple values of NperGroup with multiple numbers (50 60 70 80 90) or notation which will allow you to quickly enter many sample sizes across a range (e.g., 50 to 90 by 10) to display how power changes with sample size.

Although PROC POWER has the capability to produce plots, a flexible approach to produce graphs is to place the power calculations into a SAS dataset with the Output Delivery System (ODS). These results can then be plotted with PROC GPLOT to produce a smooth curve for each level of alpha for varying sample sizes. The specific sample sizes where power reaches 0.8 and 0.9 can easily be determined through a visual inspection of the plot by adding  $vref=(.8 .9)$  to the GPLOT PLOT statement options.

## Power Computations with PROC GLMPower

Power computations to compare the means from two (or more) groups with one or more independent variables can be illustrated with PROC GLMPower. It computes power for multifactor designs which include main effects and interactions.

The first task is to write a SAS dataset with the group means of interest for each combination of levels of the independent variables. The next example shows that the computation of power is determined from inputs to an “exemplary” dataset, in this case the actual means of interest for the two groups:

```
DATA anv;
INPUT group $ mean;
CARDS;
A 10
B 12
;

PROC GLMPower DATA=anv;
CLASS group;
MODEL mean = group;
CONTRAST 'group diff' group 1 -1;
POWER Alpha = .05
      StdDev = 4
      Ntotal = .
      Power = .9 ;
RUN;
```

The syntax for PROC GLMPower looks much like a combination of the commands from PROC GLM and PROC Power. The common group standard deviation (=4) and equal group sizes are assumed. The option NperGroup is not available here, so to compute it, divide Ntotal by the number of groups.

| Computed N Total |          |            |            |             |                 |            |
|------------------|----------|------------|------------|-------------|-----------------|------------|
| Index            | Type     | Source     | Test<br>DF | Error<br>DF | Actual<br>Power | N<br>Total |
| 1                | Effect   | group      | 1          | 170         | 0.903           | 172        |
| 2                | Contrast | group diff | 1          | 170         | 0.903           | 172        |

As illustrated, GLMPower also computes power for contrasts of interest which you specify coefficients among the levels of the factors.

More complicated mean structures can be entered in a SAS dataset and then MODEL or CONTRAST statements entered in GLMPower just like you would enter with PROCs GLM or MIXED. Since effect size computations for ANOVAs depend on the actual values of the means across all levels of the factors, entering them in a SAS dataset is more efficient than entering them directly into the procedure.

## Power Computations with PROC MIXED

Studies frequently utilize more complicated designs than T-tests or independent groups ANOVAs. Researchers may compute power for studies with these simpler designs and then run the experiment with a more complex one, assuming the power results from the former act as an “upper bound” since efficiencies are produced by the chosen design. That is, if 30 subjects are deemed necessary to meet study objectives under an independent groups design (e.g., a study will have .80 power to detect a given effect), power will likely increase if a repeated measures design is chosen.

To illustrate the flexibility of PROC MIXED, its ability to compute power for the independent groups design will be shown first. The general process is described in Chapter 12 of “SAS for Mixed Models (Littell, et. al.). With manipulation of the inputs to PROC MIXED and specifying the variance, results from the table of “Type3 Tests of Fixed Effects” will produce the power calculation for a two-sample t-test:

```
DATA anv;
INPUT group $ mean count @@; * assume equal group sizes;
CARDS;
A 10 86
B 12 86
;

ODS OUTPUT tests3=tst3;

PROC MIXED DATA=anv NOPRINT; * add the NOPRINT option;
CLASS group;
WEIGHT count;
MODEL mean = group / ddf=170; * add the error degrees of freedom;
PARMS (16) / NOITER ; * enter the assumed population variance;
RUN;

DATA pwr;
SET tst3; alpha = .05;
NonCen = NumDF*Fvalue;
Fcrit = FINV(1-alpha, numdf, dendf, 0);
Power = 1- PROBF(Fcrit, NumDF, DenDF, NonCen);

PROC PRINT DATA=pwr NOOBS;
FORMAT power 6.3; RUN;
```

|        | Num | Den |        |          |        |       |         |       |
|--------|-----|-----|--------|----------|--------|-------|---------|-------|
| Effect | DF  | DF  | FValue | ProbF    | NonCen | alpha | Fcrit   | Power |
| group  | 1   | 170 | 10.75  | 0.001265 | 10.75  | 0.05  | 3.89674 | 0.903 |

This simple example demonstrates the following five steps required to compute power with PROC MIXED for any design as presented in Chapter 12 of Littell, et. al. (2006) and will be employed in subsequent examples.

1. Make an exemplary dataset.
2. Run PROC MIXED on the dataset specifying covariance parameters and holding them fixed with a PARMS statement. Save the F-statistics from the Table of Type 3 tests into an output dataset with ODS.

Note: step 2 is crucial for the process to work. For more complicated designs, be sure you know how the PARMS statement assumes the variances are ordered which you can check through the analysis of a real or simulated dataset.

3. In a DATA step, compute the approximate noncentrality parameter:  
NonCen = Fvalue \* Degrees\_of\_Freedom.
4. Compute the critical F value (Fcrit) by entering a specified alpha and degrees of freedom into the FINV function.
5. With the inputs to cumulative FPROB function, compute power.

### Power for More Uncommon Design Situations

Power calculations for more complicated designs such as repeated measures (e.g., crossover) or multi-level models are not yet available as supported features in SAS and are available with limited capabilities in other software. However, this next example demonstrates how the outputs of PROC MIXED entered into a DATA step can compute power to detect a given effect for a specified sample size. For example, in a repeated measures design two or more data values are collected from each subject which implies they will be correlated. Thus, the standard deviation of the difference of the paired observations is the square root of the sum of the variances, adjusted for the correlation between them:

$$\sigma_{dif} = \sqrt{\sigma_1^2 + \sigma_2^2 - 2 * \rho * \sigma_1 * \sigma_2} = \sqrt{2 * \sigma^2 - 2 * \rho * \sigma^2} = \sqrt{2 * (1 - \rho) * \sigma^2}$$

For example, in a 2x2 crossover design each subject receives the two treatments, where one-half of a randomly selected subgroup of the subjects is assigned the sequence AB and the other one-half is assigned the sequence BA. In each subgroup the first treatment given is assumed to not affect the result of the second (i.e., no carryover).

The sample size for the number of subjects (i.e., pairs of observations) for a paired T-Test is computed with the same formula as presented for the two-sample case, now with a presumably smaller standard deviation for the difference:

$$\begin{aligned} \text{NSubjects} &= (\sqrt{2 * (1 - \rho) * \sigma} / \Delta)^2 * (z_{a2} + z_b)^2 \\ &= (\sigma_{dif} / \Delta)^2 * (z_{a2} + z_b)^2 \end{aligned}$$

The inputs to this equation are the same as described earlier for the two-sample T-test. What has changed is that two observations are collected from each subject under different treatment conditions. A positive correlation between these two observations will reduce the size of  $\sigma_{dif}$ , thus requiring fewer subjects to achieve a given power. For example, in a 2x2 crossover design, the objective is to detect a specified difference in the two means ( $\Delta=2$ ) of the two treatments. We will again assume the same standard deviation in each group ( $\sigma = 4$ ),  $\alpha=.05$ , power=.9032, and that pairs of observations have a correlation of  $crr=0.25$ .

```
DATA prd_t;
sigma=4;          * standard deviation of an observation =  $\sigma$  ;
delta=2;          * difference in means to be detected =  $\Delta$  ;
crr=.25;          * within subject correlation =  $\rho$  ;
alpha=.05;        * Type I error =  $\alpha$  ;
power=.9032;      * power of the test =  $1-\beta$  ;
beta=1-power;     * Type II error =  $\beta$  ;
sides=2;          * 1 or 2 sided;

* compute standard deviation of a difference of correlated values;
sigma_dif = sigma*SQRT(2*(1-crr));

z_a=ABS(PROBIT(alpha/sides)); z_b=ABS(PROBIT((1-beta)));
NSubjects = CEIL(((z_a + z_b)**2) * ((sigma_dif/delta)**2));
RUN;

PROC PRINT NOobs; run;
```

| sigma | delta | crr  | alpha | power  | beta   | sides | z_a   | z_b   | sigma_<br>dif | NSubjects |
|-------|-------|------|-------|--------|--------|-------|-------|-------|---------------|-----------|
| 4     | 2     | 0.25 | 0.05  | 0.9032 | 0.0968 | 2     | 1.960 | 1.300 | 4.89898       | 64        |

That is, 64 subjects are given the two treatments (for a total of 128 measurements) in a randomized order as described earlier. This model achieves the same power to detect a mean difference of 2 that would have required 86 subjects for each group, or a total of 172 measurements. Nearly the same results for this design can be achieved with the PairedMeans statement in PROC POWER:

```
PROC POWER;
PairedMeans
  Test = diff
  ALPHA = .05          /* Significance level */
  SIDES = 2            /* two-sided test */
  PairedMeans = (10 12) /* enter the two population means */
  PairedStddevs = (4 4) /* Enter a standard deviation in each group*/
  Corr = .25           /* correlation of paired observations */
  Npairs= .            /* Number of subjects*/
  POWER = .90          /* desired power */
;
RUN;
```

Computed N Pairs

| Actual<br>Power | N<br>Pairs |
|-----------------|------------|
| 0.90            | 65         |

### PROC MIXED to Compute Power for a Paired T-test

To compute power with PROC MIXED with more complex designs, understanding the within-subject covariance matrix R for repeated measures or the V matrix for random effects models is essential. For a design with two observations collected from each subject the structure is:

$$R = \begin{pmatrix} \sigma_{Sub+Res}^2 & \sigma_{Sub}^2 \\ \sigma_{Sub}^2 & \sigma_{Sub+Res}^2 \end{pmatrix}$$

That is, the variance of a single observation is now the sum of two variance components:

$$\text{totvar} = \sigma_{Sub+Res}^2$$

Since subjects are ‘different’, the correlation of two observations collected from each is usually positive. It is a function of the subject and residual variance components. The correlation of the pairs is found as the off diagonal entry of the correlation matrix computed as:

$$\text{crr} = \sigma_{Sub}^2 / \sigma_{Sub+Res}^2$$

In the next example the total variance of a single observation will be the same value as specified for the independent groups and repeated measures ANOVA. One may be able to determine the correlation between the two observations than knowing the variance between subjects. It is then simple to solve for the subject variance with a DATA step:

```
%LET crr=.25;      * enter the within-subject correlation;
%LET totvar=16;    * enter the total variance of a single observation;

* solve for the within-subject and residual variances;

DATA compnts;
subj = &totvar. * &crr. ;
resid = &resid - subj;
```

```
PROC PRINT DATA=compnts NOobs;
VAR subj resid;
RUN;
```

```
  subj      resid

    4         12
```

The process to compute power for a 2x2 crossover design with PROC MIXED is to first make an exemplary dataset containing the actual means of interest for each level of the within-subject factor for sample sizes of interest. That is, make two records for every subject, one for each treatment; this input dataset will assist in entering the correct degrees of freedom into the statistical calculations.

```
DATA mns;
DO subj = 1 to 65;
  trt='A'; mean = 10; OUTPUT;
  trt='B'; mean = 12; OUTPUT;
END;
RUN;
```

This design can be specified as a repeated measures analysis within PROC MIXED:

```
ODS OUTPUT tests3=tst3;

PROC MIXED DATA=mns NOPROFILE ;
CLASS subj trt;
MODEL mean = trt ;
PARMS (12 4) / NOITER ;
*   resid subj;
REPEATED / SUBJECT=SUBJ TYPE=CS R CORR R ;
RUN;
```

```
DATA pwr; SET tst3; alpha = .05;
```

```
Noncen = NumDF*Fvalue;
Fcrit = FINV(1-alpha, numdf, dendf, 0);
power = 1- PROBF(Fcrit, NumDF, DenDF, Noncen);
```

```
PROC PRINT DATA=pwr NOobs; FORMAT power 6.3; RUN;
```

| Effect | Num<br>DF | Den<br>DF | FValue | ProbF   | alpha | Noncen  | Fcrit   | power |
|--------|-----------|-----------|--------|---------|-------|---------|---------|-------|
| Trt    | 1         | 64        | 10.83  | 0.00162 | 0.05  | 10.8333 | 3.99092 | 0.900 |

The same power computation can also be achieved with a RANDOM effects model (whenever the covariance term in a repeated measures design is non-negative):

```

ODS OUTPUT tests3=tst3 V=v;
ODS EXCLUDE v ;

PROC MIXED DATA=mns NOitPrint NOProfile ;
CLASS subj trt;
MODEL mean = trt ;
PARMS (4 12) / NOITER ; * enter the subject and the residual variance;
*   subj resid;   * Note the reversed order from the REPEATED statement;
RANDOM subj / v ; * produces the V matrix of variance components;
RUN;

PROC PRINT DATA=v(obs=2) NOobs;
VAR row coll col2;
RUN;

```

| Row | Coll  | Col2  |
|-----|-------|-------|
| 1   | 16.00 | 4.00  |
| 2   | 4.00  | 16.00 |

The v option on the RANDOM statement provides the subject covariance matrix. The power computed from this random effects model is also 0.90.

### Power for the 3 Treatment Crossover Design

The crossover design with 3 treatments is one possibility for which power calculations with existing software may not be as readily available, esp. when carryover effects may be present or to specify alternative covariance structures. The design for analyzing the differences among 3 treatments collected from 3 periods and 3 sequences is:

| Period | Sequence |   |   |
|--------|----------|---|---|
|        | 1        | 2 | 3 |
| 1      | A        | B | C |
| 2      | B        | C | A |
| 3      | C        | A | B |

In this example each sequence will consist of n subjects, an equal number for each group.

An on-line power calculator provided by R. V. Lenth (2006) will calculate power for a variety of experimental situations. In particular, for a crossover design, go to the website and choose “Balanced ANOVA – any model” and then select “Crossover” from the options available. For a 3x3 crossover with 10 subjects in each sequence on the line marked “Levels” enter:

Subject 10 / sequence=period=treatment 3

Choose “Differences” and in the following window, enter a standard deviation of  $2=\sqrt{4}$  for SD of a subject and a standard deviation of  $3.464=\sqrt{12}$  for the SD {Residual}. In the right-hand side enter a 2 for the difference in treatment means to be detected. The result for n=10 subjects per sequence with the t-method is the power to detect a maximum difference of 2 among the three treatments is 0.594. Can this power result also be replicated in SAS with PROC MIXED?

First, make an exemplary dataset which reflects the 3x3 crossover design with 10 subjects, including 3 periods and 3 treatments:

```
%LET nsub = 10; * enter number of subjects in each sequence;

DATA mns;
ARRAY prd{3} _temporary_ (0 .1 .2);    * enter minimal period and;
ARRAY seq{3} _temporary_ (.1 .2 -.1);  * sequence effects;
sequence=1;
DO subject=1 to &nsub.;
  period=1; trt='A'; t=1; mn=10 + prd{period} + seq{1} ; OUTPUT;
  period=2; trt='B'; t=2; mn=10 + prd{period} + seq{1} ; OUTPUT;
  period=3; trt='C'; t=3; mn=12 + prd{period} + seq{1} ; OUTPUT;
END;

sequence=2;
DO subject=1 to &nsub. ;
  period=1; trt='B'; t=2; mn=10 + prd{period} + seq{2} ; OUTPUT;
  period=2; trt='C'; t=3; mn=12 + prd{period} + seq{2} ; OUTPUT;
  period=3; trt='A'; t=1; mn=10 + prd{period} + seq{2} ; OUTPUT;
END;

sequence=3;
DO subject=1 to &nsub.;
  period=1; trt='C'; t=3; mn=12 + prd{period} + seq{3} ; OUTPUT;
  period=2; trt='A'; t=1; mn=10 + prd{period} + seq{3} ; OUTPUT;
  period=3; trt='B'; t=2; mn=10 + prd{period} + seq{3} ; OUTPUT;
END;
RUN;

PROC TABULATE data=mns NOseps;
CLASS sequence period trt; VAR t mn;
TABLE period, sequence*(t*(mean=' '*f=3.0 n=' '*f=3.0) mn*mean=' '*f=4.1) /
rts=11;
RUN;

ODS OUTPUT tests3=tsts;
ODS OUTPUT V=v; ODS EXCLUDE v;
```

```

PROC MIXED data=mns NOitPrint NOProfile;
CLASS trt period sequence subject;
MODEL mn = sequence trt period / htype=3 ddf=27 56 56;

* run these two statements for RANDOM effects analysis;
RANDOM subject(sequence);* / v;
PARMS (4 12) / NOITER ; * enter the subject and residual variances;
*      subj res;

* run these statements (choose only 1 PARMS) for REPEATED measures approach;
*REPEATED trt / subject=subject(sequence) type=un r rcorr;
*PARMS (16) (4) (16) (4) (4) (16) / NOITER; * same covariance as RANDOM effects;
*PARMS (16) (4) (16) (6) (6) (20) / NOITER; * enter unstructured covariance;
RUN;

PROC PRINT DATA=v(obs=3) NOobs;
VAR row coll-col3;
run;

DATA pwr; SET tst; alpha = .05;
Noncen = NumDF*Fvalue;
Fcrit = FINV(1-alpha, numdf, dendf, 0);
power = 1-PROBF(Fcrit, NumDF, DenDF, Noncen);
RUN;

PROC PRINT DATA=pwr NOobs; FORMAT power 5.3;
WHERE effect = 'trt';
RUN;

```

| Effect | Num<br>DF | Den<br>DF | FValue | ProbF   | alpha | Noncen  | Fcrit   | power |
|--------|-----------|-----------|--------|---------|-------|---------|---------|-------|
| trt    | 2         | 56        | 3.33   | 0.04288 | 0.05  | 6.66667 | 3.16186 | 0.608 |

The computed power of 0.608 is reasonably close to the value computed on the indicated website. Note that the degrees of freedom were set equal to 56 in this design for consistency across the various analyses presented below. (Typically a RANDOM statement applies the “contain” option and the REPEATED statement applies “between-within” which will give slightly different results.) Also, note that since all treatment comparisons are computed within subjects the random subject effect in this model does not affect the power computations. Rather than entering a subject variance of  $\sigma_{sub}^2=4$ , it can be small (e.g.,  $\sigma_{sub}^2=1$ ) or large ( $\sigma_{sub}^2=20$ ) and the power to detect a difference in treatment means of 2 remains 0.608. It is the size of the residual variance ( $\sigma_{Res}^2=12$ ) that affects power.

PROC MIXED allows evaluations of this design to examine what happens when carryover of Treatment C exists and/or the variance covariance matrix is unstructured, that is Treatments A and B are expected to produce the same variances as before, whereas Treatment C is new and assumed to be more variable or perhaps will exert a carryover effect. How does power change under these conditions?

The covariance matrix of the RANDOM effects model demonstrated earlier is found in the V matrix. (It is the matrix entered into the first PARMS statement for the REPEATED measures approach):

| Row | Col1 | Col2 | Col3 |
|-----|------|------|------|
| 1   | 16.0 | 4.0  | 4.0  |
| 2   | 4.0  | 16.0 | 4.0  |
| 3   | 4.0  | 4.0  | 16.0 |

Assume the revised hypothetical covariance matrix is unstructured, allowing the variance due to Treatment C to be more variable than Treatments A or B (and likewise the covariances of C with A and B):

| Row | Col1 | Col2 | Col3 |
|-----|------|------|------|
| 1   | 16.0 | 4.0  | 6.0  |
| 2   | 4.0  | 16.0 | 6.0  |
| 3   | 6.0  | 6.0  | 24.0 |

This revised hypothetical covariance matrix in a REPEATED measures analysis is entered into the second PARMS statement. The expected order of the variances enters the lower triangle of this matrix, beginning with col1 of Row 1 and then moving down through Rows 2 and 3 up to the diagonal. Replace the ddfm=contain with ddf=27 56 56 (to ensure the same degrees of freedom as the REPEATED approach does not compute DenDf correctly with the PARMS statement) and replace the RANDOM statement with the REPEATED statement specifying an unstructured R matrix. That is, enter the revised variance/covariance into the PARMS statement as follows:

```
REPEATED trt / subject=subject(period) type=un r rcorr;
PARMS (16) (4) (16) (6) (6) (24) / NOITER; * enter unstructured covariance;
```

Under this new scenario, the power to detect a difference of 2 among the 3 treatment means is reduced to 0.518, since the variance of a difference with Treatment C compared to A or B increases.

The ability to examine what happens in the presence of carryover effects can also be included in the modeling process though is not demonstrated here. In all situations described so far, adjustments should be made for multiple comparisons of the treatments, so these power calculations should be regarded as a liberal upper bound.

In summary, although several programs (including SAS) and websites exist to allow one to compute power for the standard or common designs, the approach demonstrated in this paper shows how PROC MIXED can be employed for the ‘other’ situations when power calculations aren’t as readily available. In addition to serving as a power data analysis tool, it can also be very helpful in study planning.

### **What about Post-Hoc Power?** (also known as observed power or retrospective power)

You have collected the data, ran an appropriate statistical analysis, and did not observe statistical significance as indicated by a relatively “large” p-value. So you decide to compute post-hoc power to see how powerful the test was, which, by itself is essentially an empty, meaningless result. Of course the statistical test wasn't powerful enough -- that's why the p-value isn't significant. Post-hoc power is merely a one-to-one transformation of the p-value (based on the F-statistic and degrees of freedom as illustrated above). In this situation power was computed based only on what this particular sample data showed: the observed difference in means, the computed standard error, and the actual sample sizes of the groups all contributed to the observed “power” exactly as they did to the p-value.

Post hoc power also assumes the observed results establish the minimum effect size that you would like to detect; that is, the minimum observed difference in means is now dictated by the data and is not based on your knowledge of the subject matter as to what difference would be meaningful in relation to the objectives of the study. Observed results may help you interpret the sources of variability better, but if you now compute power with different group sizes or if you want to detect a different minimum effect size, the question immediately becomes prospective. What were formerly sample statistics are elevated to the status of population parameters. So, power calculations can only be considered as a prospective or an “a priori” concept. Power calculations should be directed towards planning a study, not an after-the-experiment review of the results.

Intuitively, it would seem that a small value of  $\beta$  (i.e., Type II error) should exist in order for one to declare a difference of no effect. However, since “observed” power equals  $(1-\beta)$  and increases as the p-value decreases, small values of  $\beta$  are linked to small values of  $\alpha$  which provide evidence against  $H_0$ . So if you want to show an effect does not exist, what is perhaps needed is an equivalence test; see Hoenig and Heisey, 2001.

Variations in the parameters of interest to compute power under different scenarios should be explored *before* data are collected. Concerns about sample size after a study is done can generally be refocused more directly on whether the article properly presents and interprets the uncertainty of the observed results.

None of the SAS statistical procedures (e.g., PROCs REG, TTEST, GLM, or MIXED and others) provide retrospective (post hoc) power calculations. (However, through saving results from PROC MIXED with the ODS and following through with a few basic SAS functions as illustrated in this paper, it is quite simple to compute them in a DATA step or with the inputs to PROC POWER or PROC GLMPower.) SAS developers know these computations produce misleading and biased results and thus won't automatically do it for you (although they are commonly found in the output from other statistical procedures and all-too-often are requested by some journals and their reviewers). See Hoenig and Heisey, 2001, for reasons behind this fallacious thinking.

## References

- Cohen, Jacob. (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Hillsdale, N.J., L. Erlbaum Associates.
- Cohen, Jacob, (1992) A Power Primer, Psychological Bulletin, Vol. 112, No. 1, 155-159.
- Cortina, Jose and Nouri, Hossein (2000). Effect size for ANOVA Designs. Sage University Papers Series on Quantitative Applications in the Social Sciences, -7-129. Thousand Oaks, CA: Sage.
- Hoenig, John M. and Heisey, Dennis M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," The American Statistician, 55, 19-24.
- Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," The American Statistician, 55, 187-193
- Lenth, R. V. (2006) Java Applets for Power and Sample Size (Computer Software). Retrieved 08/15/2007 from <http://www.math.uiowa.edu/~rlenth/Power/>
- Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. SAS@ for Mixed Models, Second Edition. Cary, NC: SAS Institute.

## CONTACT INFORMATION

Your comments and questions are encouraged. Please contact the author at:

Robin High  
Statistical Consultant  
University of Oregon  
110 Gerlinger Hall  
Eugene, OR 97403-1246

Email: [robinh@uoregon.edu](mailto:robinh@uoregon.edu)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.