

# **Power and Sample Size for Research Studies**

**Presented by**

**Scott J. Richter**  
**Statistical Consulting Center**  
**Dept. of Mathematics and Statistics**  
**UNCG**

## 1. Introduction

Statistical techniques are used for purposes such as estimating population parameters using either point estimates or interval estimates, developing models, and testing hypotheses. For each of these uses, a sample must be obtained from the population of interest. The immediate question is then

“How large should the sample be?”

If sampling is very inexpensive in a particular application, we might be tempted to obtain a very large sample, but settle for a small sample in applications where sampling is expensive. The cliché “bigger is better” can cause problems that users of statistical methods might not anticipate, however.

## 2. Case 1/Motivation—Estimating the mean of a population; Review of power, relation to sample size, standard deviation, Type I error rate.

Suppose a supplier provides laboratory mice with an advertised mean weight of 100 g, with standard deviation 8 g. A researcher wishes to test if a batch of mice recently received has a higher average weight. She will weigh a random sample of mice from the batch. The null hypothesis is that a population mean,  $\mu$ , is equal to 100 and we want to have a probability of 0.90 of rejecting that hypothesized value if the true value is actually 105. The value 0.90 is the selected *power* for the study:

**Power--the probability of rejecting the null hypothesis in favor of the alternative hypothesis for a specific assumed true value of the parameter (in this case, 105)**

Assume further that the chosen significance level is 0.05 and that the population standard deviation reported by the supplier is assumed to be true.

**Significance level—the probability of rejecting the null hypothesis in favor of the alternative hypothesis even though the null hypothesis is exactly true (also known as the Type I error probability)**

This will be a *one-sided* test since we are interested only in detecting a value greater than 100—that is, we have good a priori reason to believe the mean weight is greater than 100.

Given the above information, and assuming the population is normally distributed, the test statistic for testing  $H_0 : \mu_0 = 100$  versus  $H_a : \mu > 100$  is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \quad (1)$$

These inputs can be entered into software (MINITAB 17, in this case), to obtain the necessary sample size to achieve the stated goals.

Necessary information:

1. Null hypothesis:  $\mu_0 = 100$  ; Alternative hypothesis:  $\mu > 100$ ; Further assume the population of response values is normally distributed.
2. Significance level:  $\alpha = 0.05 = P(\text{conclude } \mu > 100 \text{ when } \mu = 100)$  ;
3. Difference of actual mean from hypothesized mean\*:  $105 - 100 = 5$ ;
4. Population standard deviation,  $\sigma = 8$ ;
5. Power:  $1 - \beta = 0.90 = P(\text{conclude } \mu > 100 \text{ when } \mu = 105)$ ;

(\*I will refer to this as the *hypothesized effect size*, not to be confused with Cohen's standardized effect size—more on this later)

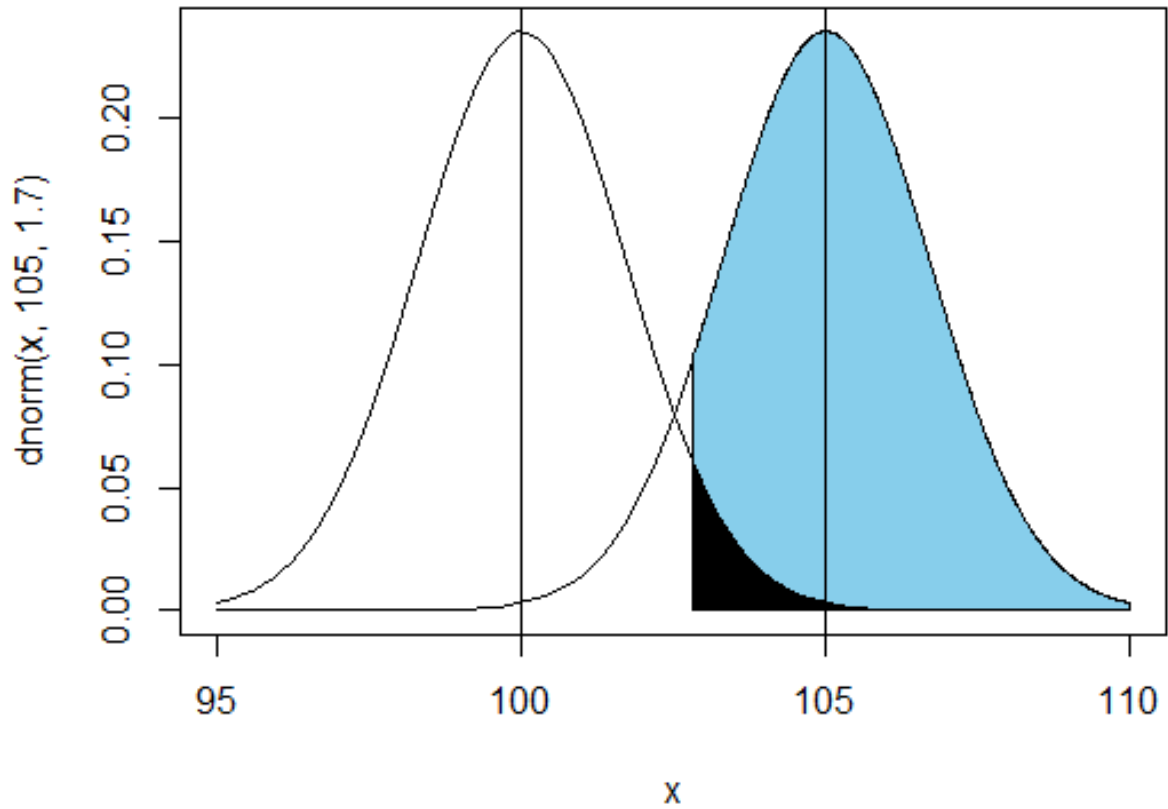
The required sample size is given by

$$n = \left[ \frac{(Z_\alpha + Z_\beta)\sigma}{\mu - \mu_0} \right]^2 \quad (2)$$

where:

$Z_\alpha$  is the critical value of the standard normal distribution under the null hypothesis, whose value is determined by the choice of significance level;  
 $Z_\beta$  is the critical value of the standard normal distribution under the alternative, whose value is determined by the choice of significance level and power;  
 $\mu - \mu_0$  is the difference of the actual mean from the hypothesized mean.

$$n = \left[ \frac{(1.645 + 1.282)8}{5} \right]^2 = 21.93 \Rightarrow n = 22$$



Using software (MINITAB):

The screenshot shows the Minitab software interface. The main window displays the 'Power and Sample Size' dialog box for a 1-Sample Z Test. The session window shows the following text:

```

6/19/2012 11:19:45 AM
Welcome to Minitab, press F1 for help.
Power and Sample Size
1-Sample Z Test
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 8

```

The dialog box 'Power and Sample Size for 1-Sample Z' is open, showing the following settings:

- Specify values for any two of the following:
  - Sample sizes: [ ]
  - Differences: 5
  - Power values: .9
- Standard deviation: 8

The dialog box also includes buttons for 'Options...', 'Graph...', 'Help', 'OK', and 'Cancel'. The background shows a worksheet grid with columns C1 through C21 and rows 1 and 2.

## Power and Sample Size

### 1-Sample Z Test

Testing mean = null (versus  $>$  null)

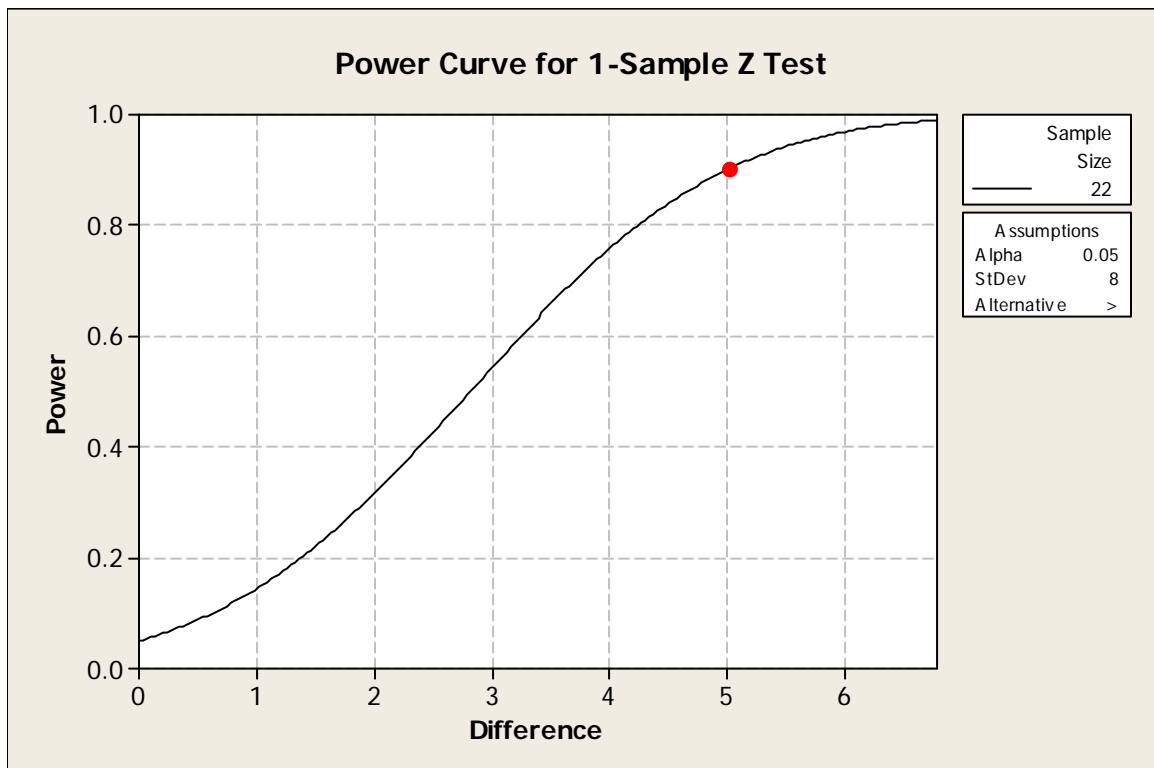
Calculating power for mean = null + difference

Alpha = 0.05 Assumed standard deviation = 8

Sample Target

Difference Size Power Actual Power

5 22 0.9 0.900893

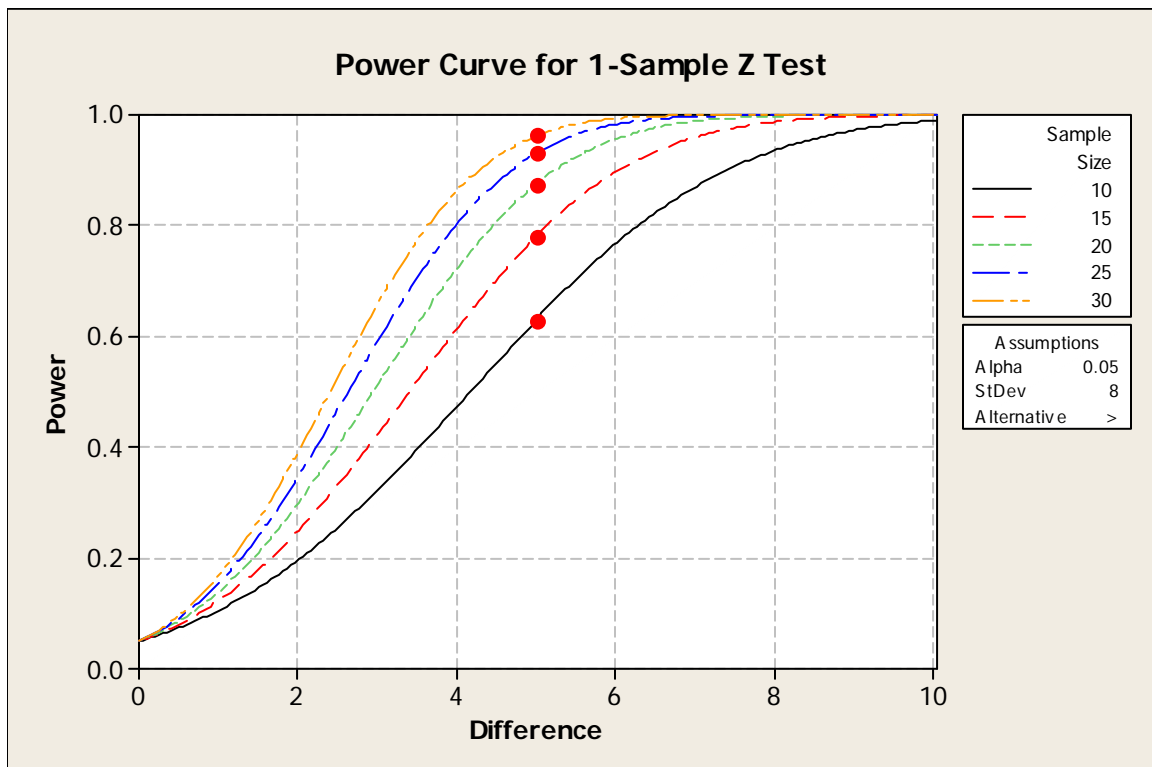


Factors that affect power/sample size. Note: All other things being fixed, greater power requires a larger sample size, and vice-versa.

If everything else is held fixed:

1. *If significance level decreases (e.g., to account for multiple testing), power will decrease (and thus required sample size increases).*
2. *If hypothesized effect size decreases power will decrease (and thus required sample size increases).*
3. *If the estimate of the standard deviation decreases, power will increase (and thus required sample size decreases).*

The plot below illustrates the effect of sample size on power. Deciding upon sample size often involves a trade-off among sample size, power and difference from hypothesized value.



In general, any sample size/power analysis involves the following elements:

1. Specify a hypothesis test on some parameter,  $\theta$  (e.g., population mean), along with the underlying probability model for the data (e.g., normal, lognormal);
2. Specify the significance level of the test;
3. Specify a value of the parameter,  $\tilde{\theta}$ , that reflects an alternative of scientific interest;
4. Obtain estimates of other parameters needed to compute the power function of the test;
5. Specify the desired power of the test when  $\theta = \tilde{\theta}$ .



How do we determine inputs?

I. Standard deviation (and possibly) other parameters

An accurate estimate of the standard deviation is crucial to an accurate sample size estimate. Three different estimates for the previous example are given below. Note that substantially under or overestimation of the true standard deviation can yield vastly different sample size estimates!

1-Sample Z Test

Testing mean = null (versus > null)  
Calculating power for mean = null + difference  
Alpha = 0.05 **Assumed standard deviation = 4**

| Sample Difference | Target Size | Power      | Actual Power    |
|-------------------|-------------|------------|-----------------|
| 5                 | <b>6</b>    | <b>0.9</b> | <b>0.921760</b> |

1-Sample Z Test

Testing mean = null (versus > null)  
Calculating power for mean = null + difference  
Alpha = 0.05 **Assumed standard deviation = 8**

| Sample Difference | Target Size | Power      | Actual Power    |
|-------------------|-------------|------------|-----------------|
| 5                 | <b>22</b>   | <b>0.9</b> | <b>0.900893</b> |

1-Sample Z Test

Testing mean = null (versus > null)  
Calculating power for mean = null + difference  
Alpha = 0.05 **Assumed standard deviation = 16**

| Sample Difference | Target Size | Power      | Actual Power    |
|-------------------|-------------|------------|-----------------|
| 5                 | <b>88</b>   | <b>0.9</b> | <b>0.900893</b> |

Of course we generally don't know the value of the population standard deviation.

**So what do ¥ people do in practice to get around the problem of unknown standard deviation?**

\*Lenth (2007): “If you have no idea of  $\sigma$ , then you are not ready to do a definitive study and should first do a pilot study to estimate it.”

Ways to estimate standard deviation

1) Pilot study

1) External—Subjects used will not be part of the full study. Could be data collected specifically for the current study, or from a previous study using the same outcome.

- a) Often underestimates the true variance (or is less than the eventual estimate from the full study).
- b) Can be unrepresentative of the population under study.
- c) Vickers (2003)—Found 80% of clinical trials examined used a smaller standard deviation estimate to compute sample size than was eventually found in the full studies.

2) Internal—Subjects used will be part of the full study. Usually works as follows:

- a) Sample size for the full study is estimated;
- b) At some point the standard deviation is re-estimated using the data collected up to that point;
- c) If the re-estimate is no bigger than the original estimate, then use the original sample size estimate;
- d) Otherwise, revise the sample size based on the new (larger) standard deviation estimate.

Issues:

- a) At what point in the data collection should the re-estimate occur?
- b) If the study is stopped too soon, there will be a large variance associated with the estimate of the population standard deviation, so the latter might be poorly estimated, which in turn would produce a sample size that is much too large;

1. Could use an upper confidence bound for the standard deviation
    - a. at least 80% confidence
    - b. generally not implemented in software
  2. Note these issues also apply to external pilot studies
- c) Type I error rate can be inflated, since the pilot and main studies are not independent

2) Based on expected range of outcome

If outcome can be assumed approximately normally distributed, then

$\hat{\sigma} = \frac{\text{range}}{6}$  is a conservative estimate. Based on the fact that

$(\mu + 3\sigma) - (\mu - 3\sigma)$  should include virtually all of the distribution.

## II. Determining difference from hypothesized mean (effect size)

- Define “clinically meaningful” amount
- Think about what you “expect/hope” to see
  - Treat as an upper bound on the effect
  - Establishes minimum sample size
- Sometimes helpful to think in terms of relative differences instead of absolute (e.g., is a 10% decrease in systolic blood pressure of practical importance?)
  - van Bell & Martin (1993), van Bell (2008) discuss using the coefficient of variation (CV) in this context
- Put yourself in the subject/patient’s shoes—e.g., would the benefits of reducing systolic BP by 15 points outweigh the cost, inconvenience and potential side effects of the treatment?
- Software—Can help show different sample size/detectable effect scenarios
- Should be specified in actual measurement units (See Lenth, 2001, 2007)
- Should be reasonable, in the sense that an effect of this magnitude could be anticipated in the given field of research
- Avoid standardized effect sizes

### III. Power/significance level

The choice of the power value and the significance level for the test should be based on what would seem appropriate for each specific application, rather than based on how they affect sample size. Ideally, costs associated with incorrect decisions should be estimated and factored into the determination (e.g., incorrectly concluding that a treatment or intervention is effective, when in reality it is not.)

Mudge et al (2012) argue that “in most studies (and perhaps all) the significance level should be set with the objective of either minimizing the combined probabilities of making Type I or Type II errors at a critical effect size, or minimizing the overall cost associated with Type I and Type II errors given their respective probabilities.” The steps of their proposed method are as follows:

1. Determine the critical effect size;
2. Choose whether to minimize the average probability of Type I and Type II errors, or the relative cost of errors;
3. Calculate the optimal significance level;
4. Report all of the following for the test:
  1. Sample size;
  2. Critical effect size;
  3. Chosen relative cost of Type I to Type II error;
  4. Optimal Type I error rate;
  5. Optimal Type II error rate;
  6. Average of Type I and Type II error rates.

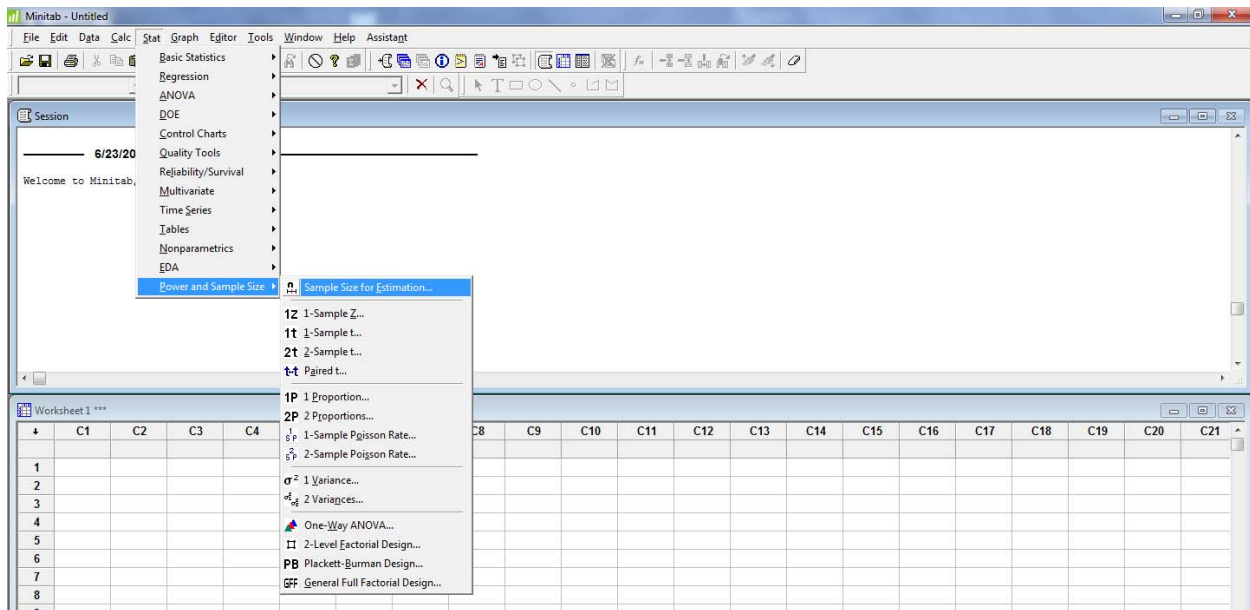
Their suggestion involves simply averaging the probabilities of Type I and Type II errors, possibly weighting each based on cost or prior probabilities.

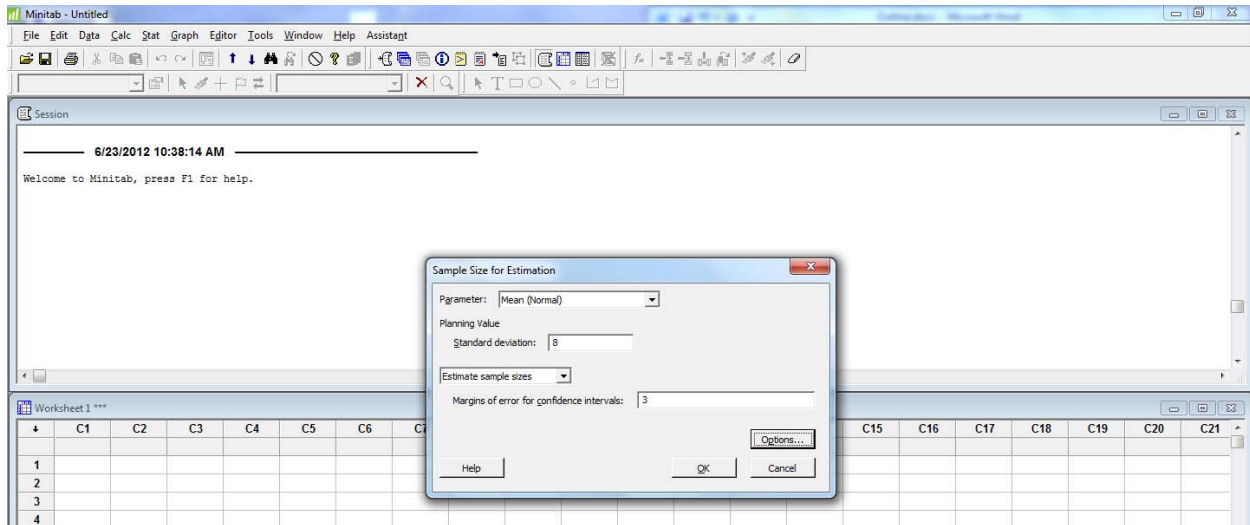
While acknowledging that their method is not perfect, they claim it is a substantial improvement over the traditional method of basing the interpretation of the test on the  $p$ -value and its juxtaposition to 0.05.

## What about confidence intervals?

- In many (perhaps most) situations, confidence intervals yield at least as much information as  $p$ -values
- Samples should be large enough so that point and interval estimates can be obtained with high precision
- Revised sample size elements:
  1. No hypotheses regarding the value of the parameter,  $\theta$ .
  2. Specify the confidence level of the estimate;
  3. Specify precision, that is, answer the question, “how close to the true value of the parameter does the estimate need to be?”;
  4. Obtain estimates of other parameters needed to compute the precision;
  5. No concept of power, since there is no test.

For the previous example, suppose it is desired to estimate the true mean with 95% confidence, and so that the estimate is at most 3 units from the true mean.





\*Under options, choose “Assume population standard deviation known”

## Sample Size for Estimation

### Method

|                     |                      |
|---------------------|----------------------|
| Parameter           | Mean                 |
| Distribution        | Normal               |
| Standard deviation  | 8 (population value) |
| Confidence level    | 95%                  |
| Confidence interval | Two-sided            |

### Results

|                 |             |
|-----------------|-------------|
| Margin of Error | Sample Size |
| 3               | 28          |

**The required sample size is  $n = 28$ .**

## t-methods

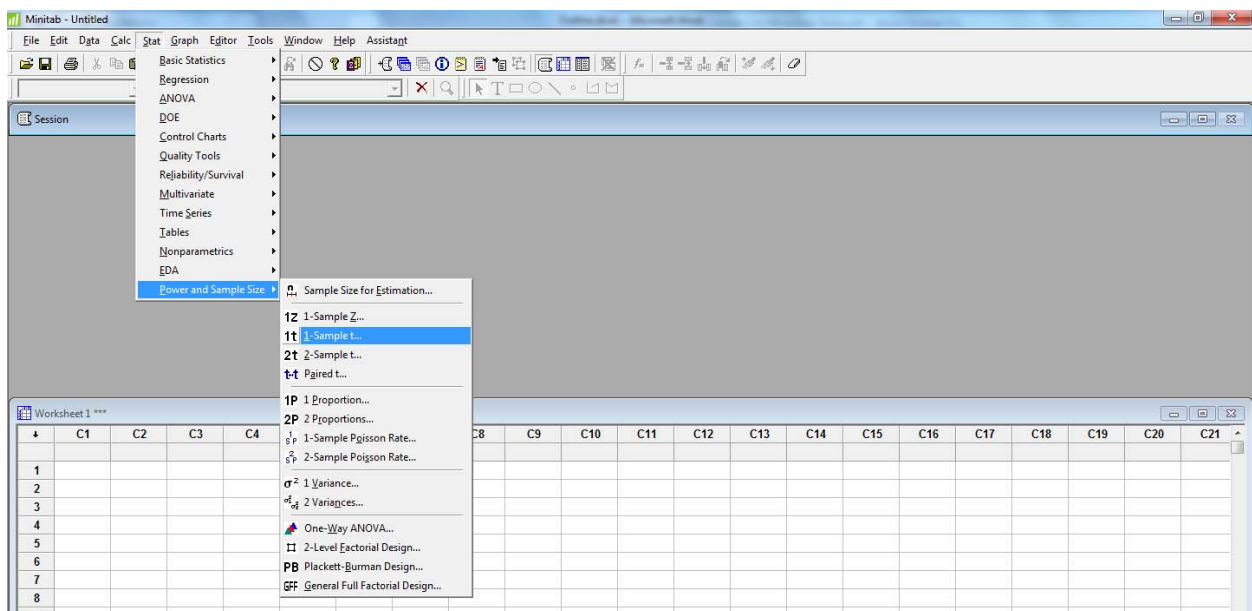
In practice, the sample standard deviation will usually be used in the analysis (thus a  $t$ -test or interval will be calculated instead of a  $z$ -test or interval). In that case, there is greater uncertainty in the estimate, which must be incorporated into the sample size estimate. For the case of testing the mean of a normal population, the formula is now

$$n = \left[ \frac{(t_{\alpha, n-1} + t_{\beta, n-1})s}{\mu - \mu_0} \right]^2 \quad (3)$$

Note that this looks very much like the formula when  $\sigma$  is assumed known, except with critical values from a Student's  $t$ -distribution replacing those of the standard normal distribution. This introduces two complications:

1.  $t_{\beta, n-1}$  has a noncentral  $t$ -distribution, for which tables are generally not available;
2.  $n - 1$ , the degrees of freedom, must be specified in order to determine  $t_{\alpha, n-1}$  and  $t_{\beta, n-1}$ .

Thus, the calculation must be done by trial and error, and is best performed using software. Using Minitab to redo the sample size calculation using the above formula results in the slightly larger  $n = 24$ .



## 1-Sample t Test

Testing mean = null (versus > null)

Calculating power for mean = null + difference

Alpha = 0.05 Assumed standard deviation = 8

| Difference | Sample Size | Target Power | Actual Power |
|------------|-------------|--------------|--------------|
| 5          | 24          | 0.9          | 0.907420     |



### 3. Comparing two means

#### 3.1. Case 1--Independent samples

The same issues are present as before, plus a few more.

Necessary information:

1. Specify hypotheses--Null hypothesis is usually  $\mu_1 - \mu_2 = 0$  ; Alternative hypothesis:  $\mu_1 - \mu_2 > 0$  ; Assume both populations of response values are normally distributed;
2. Specify significance level;
3. Specify difference between population means to be detected;
4. Specify population standard deviations—several options
  - a. population values known—can use  $z$ -test for analysis
  - b. population values unknown but assumed equal—can use pooled  $t$ -test for analysis
  - c. population values unknown and not assumed equal—will use Welch/Satterthwaite tests for analysis
5. Specify power;
6. Additionally, sample sizes may be computed to be equal or unequal.

Example.

Newcombe (2001) referred to a study by Heasman et al (1998) in which they found that the mean toothbrushing force at baseline was 220 grams for boys and 181 grams for girls. Motivated somewhat by this study, Newcombe (2001) stated “Suppose we decided that in a new study, we want 80 percent power to detect a difference of 30 grams as statistically significant at the 5 percent level. Based on the existing study, it seems reasonable to assume a SD of 130g. Suppose that, as in the published study, **we expect to recruit twice as many girls as boys ...**”

Most software will not accommodate unequal sample sizes per group. However, SAS Power and Sample Size will. First, the calculation for equal sample sizes yields  $n_i = 296$  per group. The second calculates  $n_1 = 444$  girls and  $n_2 = 222$  boys.

SAS Power and Sample Size

File Tools Help

Two-sample t test

Edit Properties View Results

Summary Table  
Graph  
Narratives  
SAS Log  
SAS Code

**Narratives**

For a two-sample pooled t test of a normal mean difference with a two-sided significance level of 0.05, assuming a common standard deviation of 130, a sample size of 296 per group is required to obtain a power of at least 0.8 to detect a mean difference of 30. The actual power is 0.8.

Create Narratives

Select one or more scenarios

| Select                              | Index | Sides | Alpha | MeanDiff | StdDev | NominalPo... | NullDiff | Power         | NPerGroup |
|-------------------------------------|-------|-------|-------|----------|--------|--------------|----------|---------------|-----------|
| <input checked="" type="checkbox"/> | 1     | 2     | 0.05  | 30       | 130    | 0.8          | 0        | 0.80035661... | 296       |

SAS Power and Sample Size

File Tools Help

Two-sample t test

Edit Properties View Results

Summary Table  
Graph  
Narratives  
SAS Log  
SAS Code

**Narratives**

For a two-sample pooled t test of a normal mean difference with a two-sided significance level of 0.05, assuming a common standard deviation of 130, a total sample size of 666 assuming an allocation ratio of 2 to 1 is required to obtain a power of at least 0.8 to detect a mean difference of 30. The actual power is 0.8.

Create Narratives

Select one or more scenarios

| Select                              | Index | Sides | Alpha | MeanDiff | StdDev | Weight1 | Weight2 | NominalPo... | NullDiff | Power         | NTotal |
|-------------------------------------|-------|-------|-------|----------|--------|---------|---------|--------------|----------|---------------|--------|
| <input checked="" type="checkbox"/> | 1     | 2     | 0.05  | 30       | 130    | 2       | 1       | 0.8          | 0        | 0.80049917... | 666    |

### 3.2. Case 2--Dependent samples

Example (Cohen, 1988, p.51): In a study to appraise the efficacy of prescribing a program of diet and exercises to a group of overweight male students, subjects will be weighed, prescribed to the program, and then weighed again 60 days later.

Necessary information:

1.  $H_0 : \mu_{after} - \mu_{before} = \mu_d = 0$  ;  $H_a : \mu_{after} - \mu_{before} = \mu_d > 0$ ;
2.  $\alpha = 0.05$ ;
3. Detect a mean loss of 4 lb.;
4. Desired power = 0.90;
5.  $\sigma = ?$

You may recall that computationally this test is identical to that described in Section II, treating the differences as a sample from a normal population. Thus, it is important to remember that while the standard deviation needed here is that of the difference score (before – after), the standard deviation of the difference is a function of both sample variances as well as the correlation,  $\rho$ , between measurements:

$$\sigma_{y_1 - y_2} = \sqrt{\sigma_{y_1}^2 + \sigma_{y_2}^2 - 2\rho\sigma_{y_1}\sigma_{y_2}} \quad (4)$$

\*Stronger positive correlation -> smaller SD -> smaller sample size required

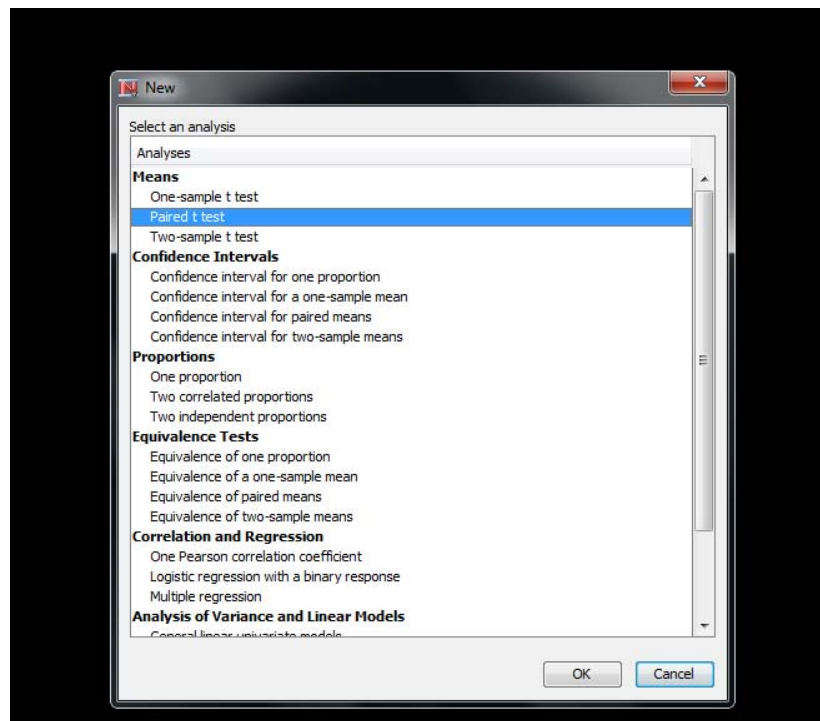
\*Stronger negative correlation -> larger SD -> larger sample size required

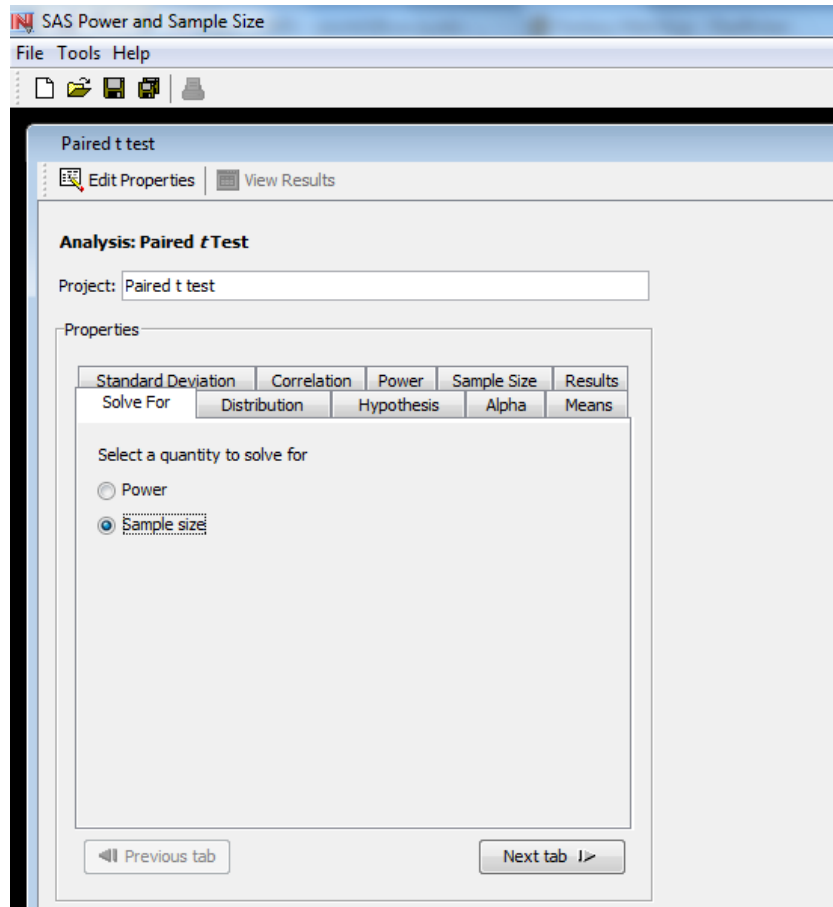
\*Correlation = 0 -> SD is the same as for independent samples

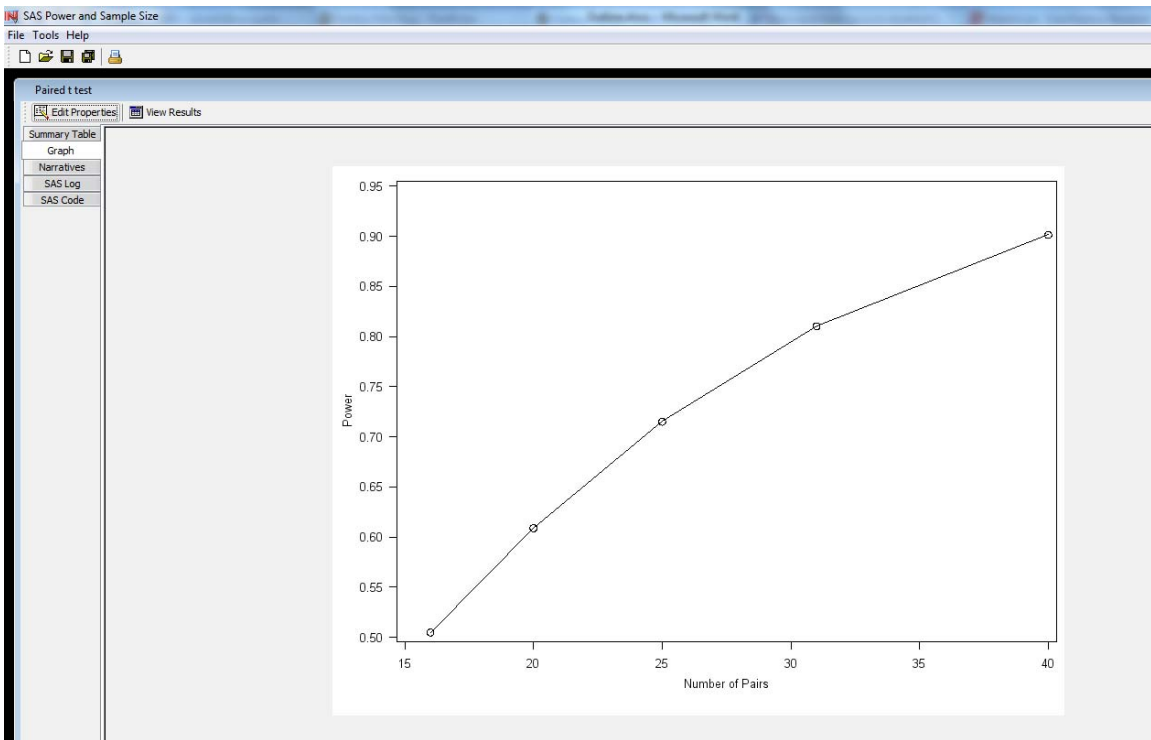
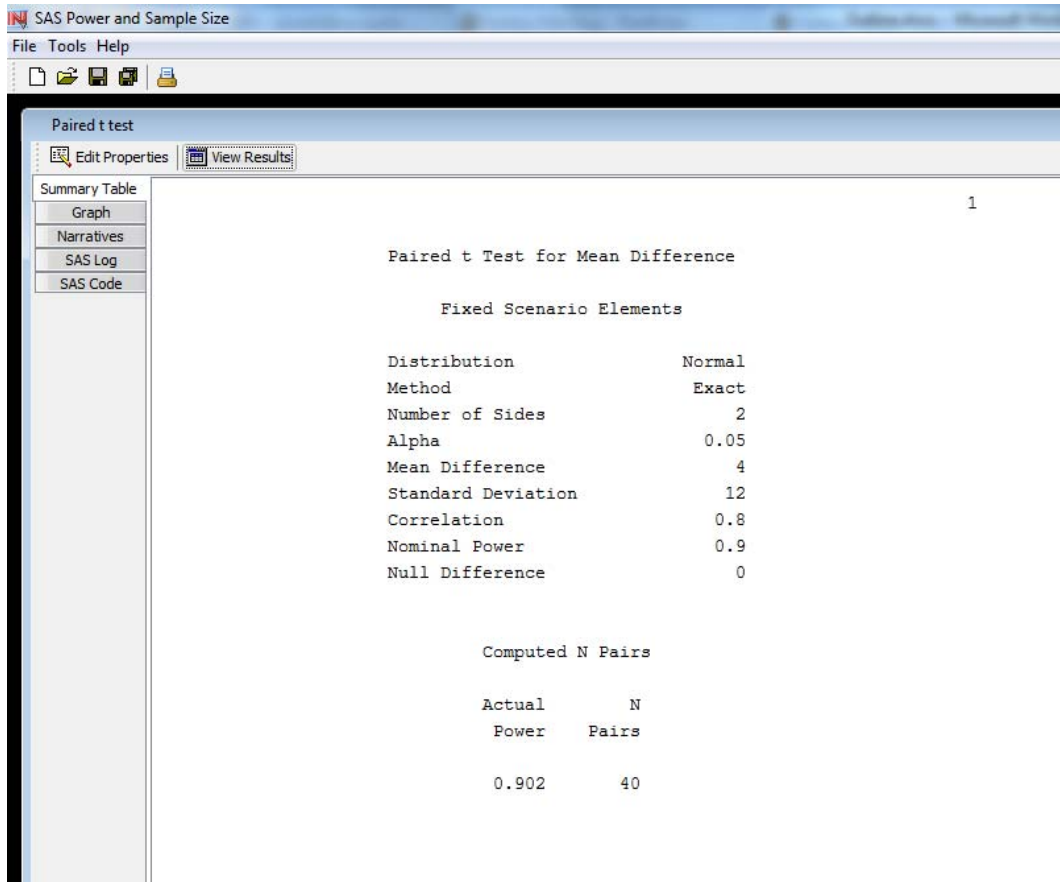
A pilot study could be used to estimate this standard deviation. However, if pilot data are unavailable, 3 parameters must be estimated to determine the power/sample size.

Suppose it is assumed that  $\sigma = 12$  for both time points, and that  $\rho = 0.80$ . Then  $\hat{\sigma}_{y_1 - y_2} = 7.6$ .

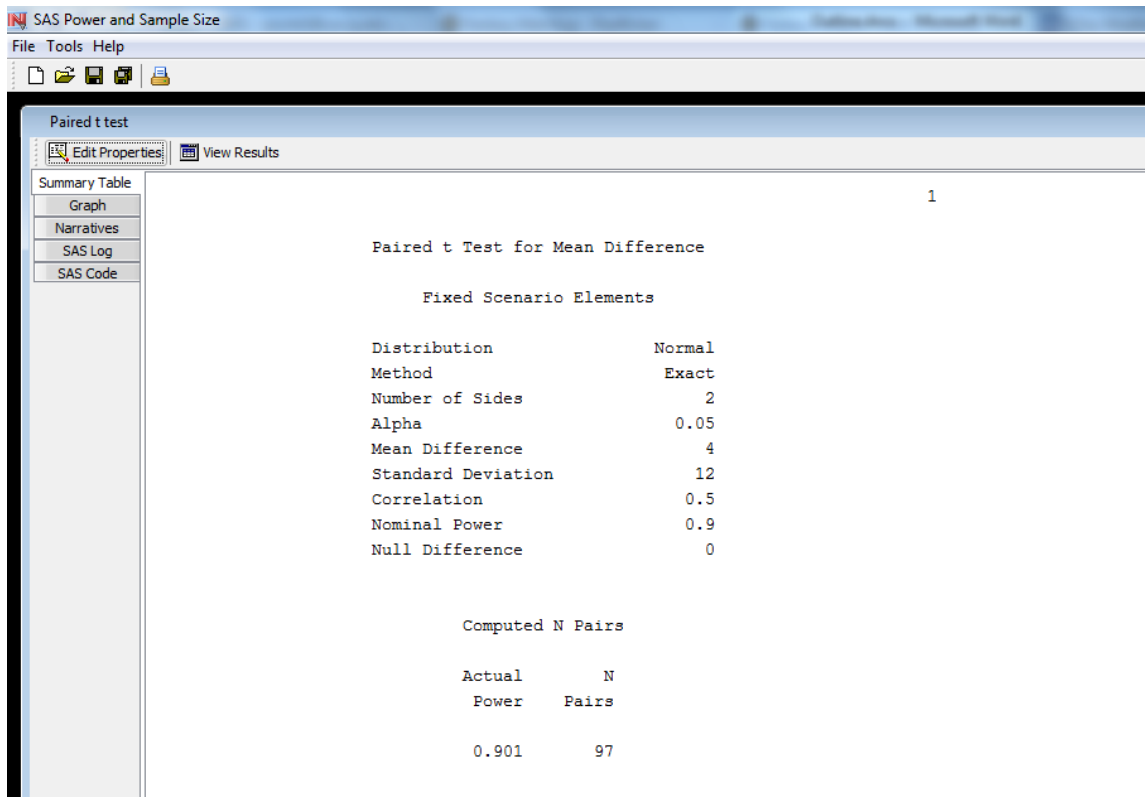
Many software programs require that the value  $\hat{\sigma}_{y_1 - y_2}$  be supplied. A few can take as inputs the separate standard deviations and correlation, for instance, SAS Power and Sample Size:





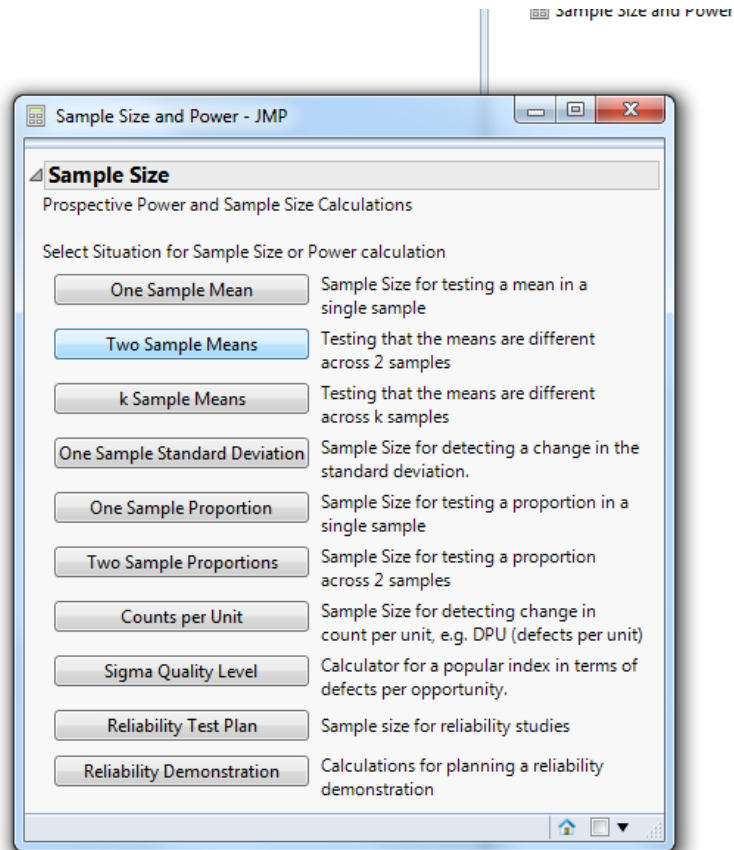


What is the effect of correlation on sample size? Below is the calculation if  $\rho = 0.50$ :

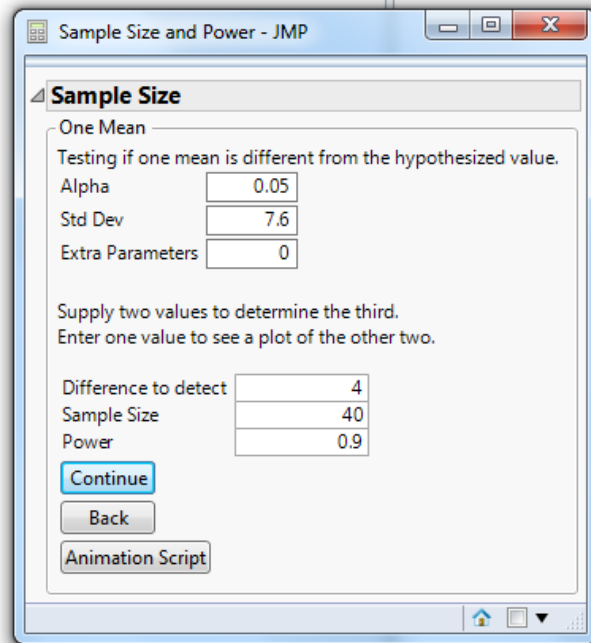


\*As expected, a larger sample size is required for a weaker correlation.

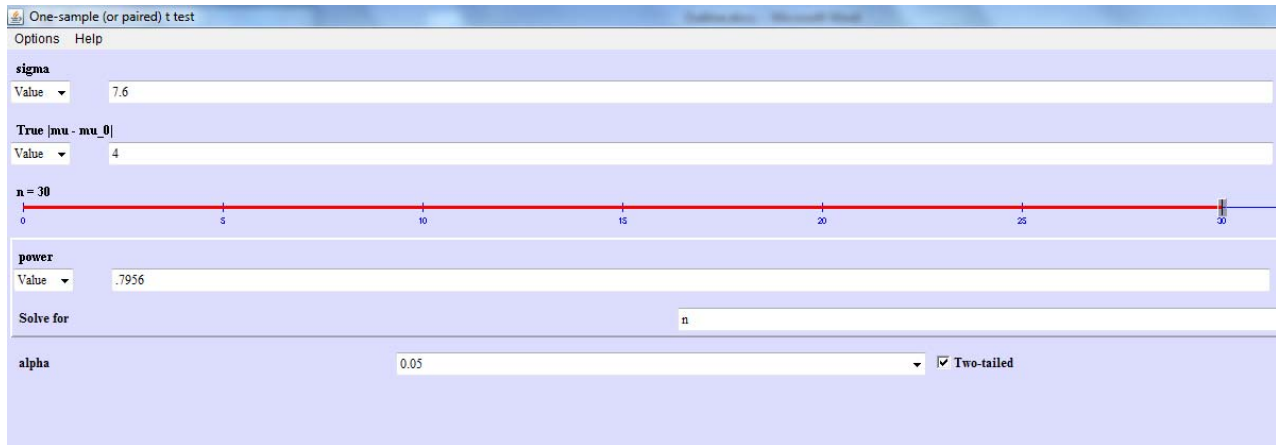
JMP requires that  $\hat{\sigma}_{y_1 - y_2} = 7.6$  be input:







Lenth's calculator also requires that  $\hat{\sigma}_{y_1 - y_2} = 7.6$  be input:



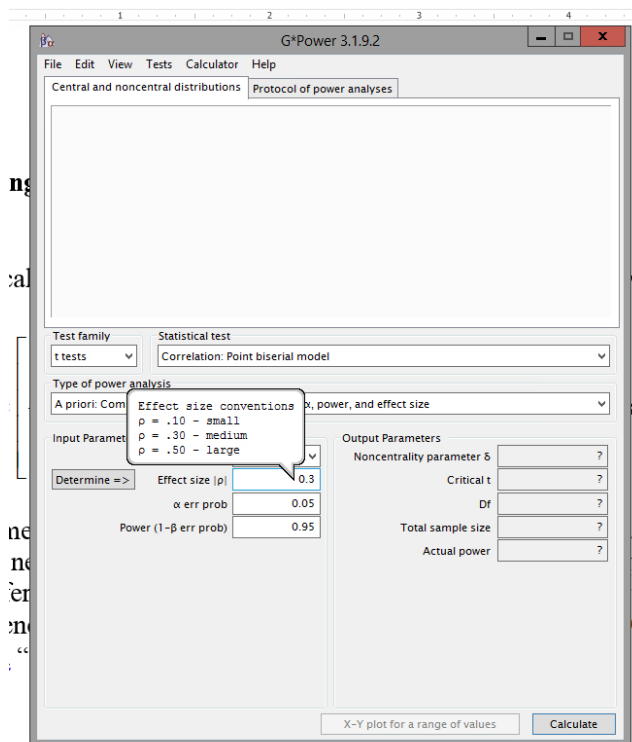
### 3.3 Using standardized effect sizes

Recall for the one-sample case we found  $n = \left[ \frac{(Z_\alpha + Z_\beta)\sigma}{\mu - \mu_0} \right]^2$  which is equivalent to

$$n = \left[ \frac{(Z_\alpha + Z_\beta)}{\left( \frac{\mu - \mu_0}{\sigma} \right)} \right]^2 = \left[ \frac{(Z_\alpha + Z_\beta)}{d} \right]^2. \text{ The quantity } \frac{\mu - \mu_0}{\sigma} = d \text{ is a } \textit{standardized effect}$$

*size*.

Some software (such as GPower) will allow input of a standardized effect size, without the need to provide an estimate of the standard deviation or hypothesized mean difference. Cohen (1988), based on a survey of several hundred studies in the social sciences, divided the effect sizes observed into “small”, “medium” and “large”.



Lenth (2007) criticized them as *T-shirt effect sizes*: “This is an elaborate way to arrive at the same sample size that has been used in past social science studies of large, medium, and small size (respectively). The method uses a standardized effect size as the goal. Think about it: for a ‘medium’ effect size, you’ll choose the same  $n$  regardless of the accuracy or reliability of your instrument, or the narrowness or diversity of your subjects. Clearly, important considerations are being ignored here. ‘Medium’ is definitely not the message!”

## Illustration

Suppose we plan to use BMI as the primary outcome variable for a study, and suppose we have two different methods to measure BMI. From past experience, the standard deviation of Method 1 is estimated to be 5%BF, while Method 2 is 10%BF. We further decide we want to be able to detect a 5%BF difference in the means of our treatment and control groups.

The effect size using Method 2 is  $d = \frac{5}{10} = 0.5$ , which Cohen would classify as “medium”. The required total sample size for power 0.9 and  $\alpha = 0.05$  is  $N = 172$ , or 86 per group.

Using Method 1, however,  $d = \frac{5}{5} = 1$  (“large”), and the required sample size is  $N = 46$ , or 23 per group.

***Depending on the method used, the same desired effect size to detect could be classified as either “medium” or “large” using the standardized effect size.***

A recent U.S. Dept. of Education sponsored report stated, “*The widespread indiscriminate use of Cohen’s generic small, medium, and large effect size values to characterize effect sizes in domains to which his normative values do not apply is thus likewise inappropriate and misleading.*”

## 4. Testing/estimating a population proportion

We will first consider the case of a single proportion, with  $p_0$  denoting the value under the null hypothesis, and  $p$  denoting the true value. For a one-sided test, the usual textbook test statistic (but not necessarily the best approach under all conditions) which is based on the assumed adequacy of the normal approximation to the binomial distribution is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (5)$$

This statistic is appropriate when  $n$  is large and both  $p_0$  and  $\hat{p}$  are not too far from 0.5.

Note that here we are in luck, since the standard deviation for a binary outcome is  $\sigma = p(1-p)$ , and thus can be calculated once  $p_0$  and  $p$  have been specified.

Then the expression for  $n$  is

$$n = \left[ \frac{Z_\alpha \sqrt{p_0(1-p_0)} + Z_\beta \sqrt{p(1-p)}}{p - p_0} \right]^2 \quad (6)$$

Example.

Suppose we have:

1. Null hypothesis:  $p = 0.50$  ; Alternative hypothesis:  $p > 0.50$  ;
2. Significance level:  $\alpha = 0.05$  ;
3. Expected proportion under the alternative:  $p = 0.60$  ;
4. Power:  $1 - \beta = 0.80$ .

Then the required sample size is

$$n = \left[ \frac{1.645\sqrt{0.5(1-0.5)} + 0.8416\sqrt{0.6(1-0.6)}}{0.6 - 0.5} \right]^2 = 152.5 \Rightarrow n = 153.$$

Most software will compute this. Below Lenth's calculator is used.



## Other approaches

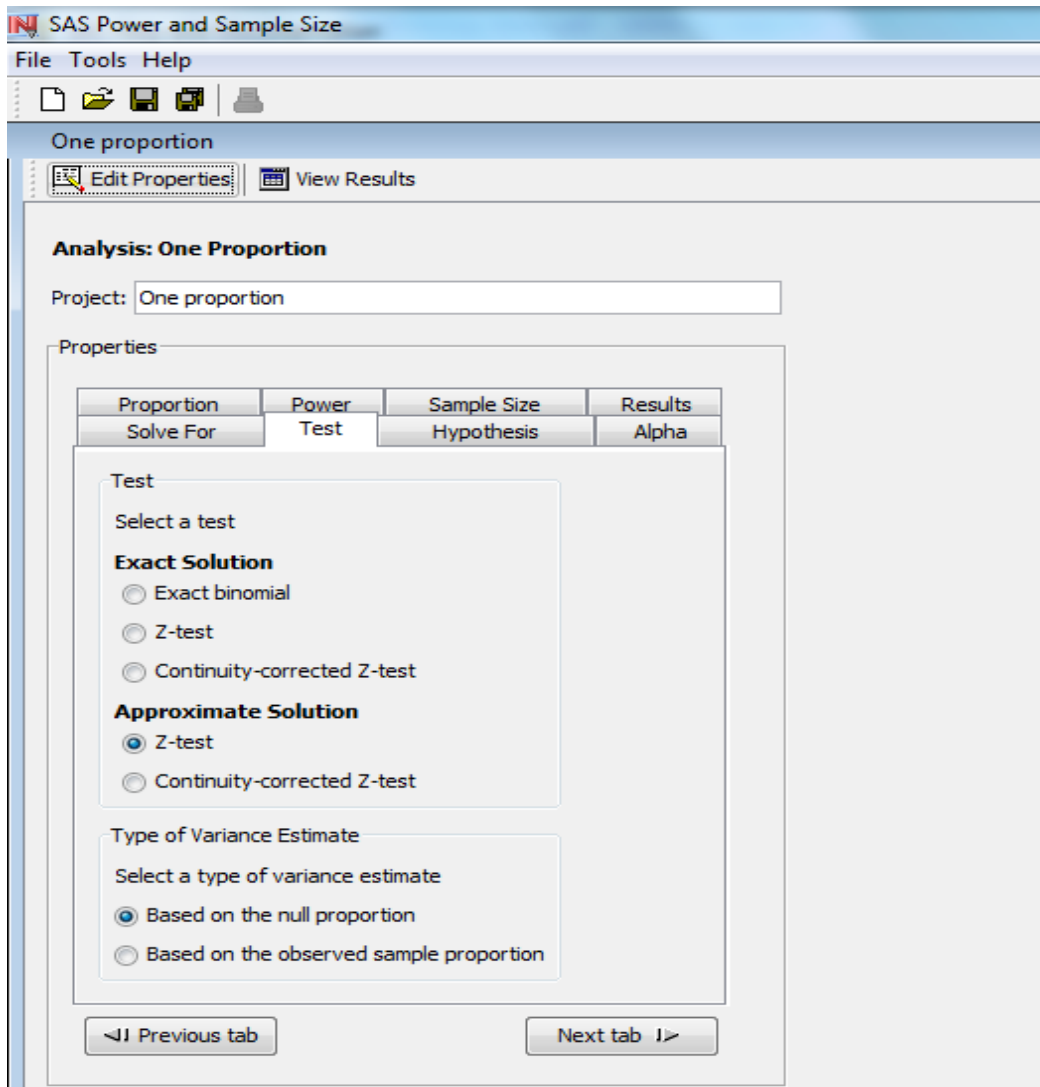
If the normal approximation is not expected to be adequate, due to sample size expectation and/or the true proportion near 0 or 1, then the above formula is not appropriate. There are several other approaches that may be available.

### Lenth's calculator

*Exact:* In the exact test, the significance level alpha is taken as an upper bound on the size of the test (its power under the null hypothesis). Since  $X$  has a discrete distribution, the size cannot be controlled exactly and is often much lower than the specified alpha.

*Exact (Wald CV):* This is like the exact method, except the critical values are calculated based on an adjusted Wald statistic (Agresti and Coull, 1998). This does NOT guarantee that the size of the test is less than alpha.

## SAS Power and Sample Size:



## 5. Regression

The general form of the linear regression model may be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon$$

with  $Y$  the dependent variable,  $X_1, X_2, \dots, X_m$  the independent variables (predictors, regressors, covariates), and  $\varepsilon$  is the error term, assumed to be independent and have a normal distribution with mean 0 and standard deviation  $\sigma_\varepsilon$ .

**Draper and Smith (1998) suggested a simple rule of thumb to use at least 10 observations per predictor in the model.** However, their suggestion is not based on statistical theory, and bypasses the serious consideration of detecting results of practical significance with specified power.

### 5.1. Case 1--Simple regression ( $m = 1$ )

If we try to take an analytical approach to sample size determination, we run into problems, even when there is only a single predictor (Ryan, 2012).

Consider the dialog help for Lenth's calculator:

“This is a simple interface for studying power and sample-size problems for simple or multiple linear regression models. It is designed to study the power of testing one predictor,  $x[j]$ , in the presence of other predictors. The power of the  $t$  test of a regression coefficient depends on the error SD, the SD of the predictor itself, and the multiple correlation between that predictor and other predictors in the model. The latter is related to the variance inflation factor. It is assumed that the intercept is included in the model.

The components in the dialog are as follows:

- 1) No. of predictors: Enter the total number of predictors (independent variables) in the regression model.
- 2) SD of  $x[j]$ : Enter the standard deviation of the values of the predictor of interest.

3) Alpha: The desired significance level of the test.

4) Two-tailed: Check or uncheck this box depending on whether you plan to use a two-tailed or a one-tailed test. If it is one-tailed, it is assumed right-tailed. If a left-tailed test is to be studied, reverse the signs and think in terms of a right-tailed test.

5) Error SD: The SD of the errors from the regression model. You likely need pilot data or some experience using the same measurement instrument.

6) Detectable beta[j]: The clinically meaningful value of the regression coefficient that you want to be able to detect.

7) Power: The power of the  $t$  test, at the current settings of the parameter values.”

Both the standard deviation of the dependent variable (response) and the standard deviation of the predictor are required. If the predictor is fixed (known values) then estimating its standard deviation is not an issue. However, if the predictor is random (the more likely scenario), then pilot data may be needed to estimate both its standard deviation as well as that of the response. Other methods, discussed earlier (such as using the expected range) could also be employed.



Ryan (2009) suggested that it is not sufficient to simply reject  $H_0 : \beta_1 = 0$ , since the model may still have poor predictive power. He suggests **a rule of thumb (based upon the work of Wetz (1964)) to only conclude the model is useful if the  $t$ -statistic for testing  $H_0 : \beta_1 = 0$  is at least twice the critical value** (Ryan, 2009, p.20).

The  $t$ -statistic for testing  $H_0 : \beta_1 = 0$  is  $t = \frac{\hat{\beta}_1}{s_e / \sqrt{S_{xx}}}$ , where  $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$ .

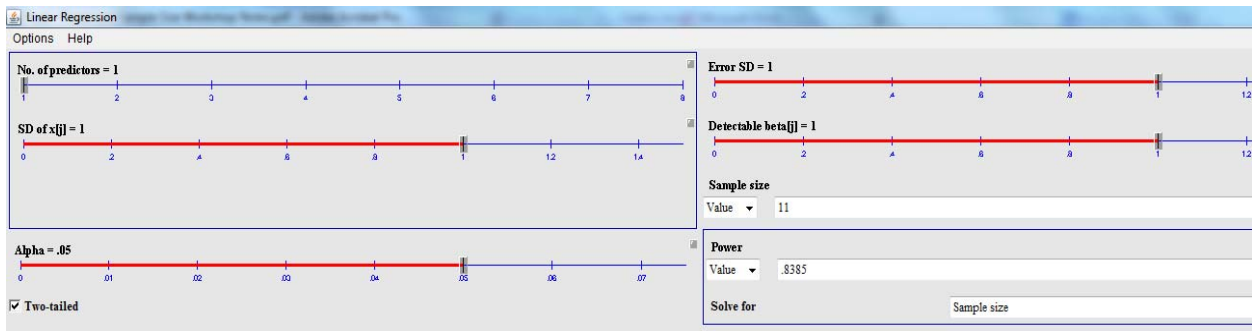
Since  $\sqrt{S_{xx}} = s_x \sqrt{n-1}$ , where  $s_x$  is the standard deviation of the predictor, it

follows the  $t$ -statistic can be written as  $t = \frac{\hat{\beta}_1}{\frac{s_e}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1}{\frac{s_e}{s_x \sqrt{n-1}}} = \left( \frac{s_x}{s_e} \right) \hat{\beta}_1 \sqrt{n-1}$ . Then

using the rule of thumb, we want

$$t = \left( \frac{s_x}{s_e} \right) \hat{\beta}_1 \sqrt{n-1} \geq 2t_{\alpha/2, n-2} \quad (7)$$

For a two-sided test. Ryan (2009) suggests as a reasonable starting point to set  $s_x = s_e = \hat{\beta}_1 = 1$ . Using Lenth's calculator with the inputs suggested above, indicates a sample size of  $n = 11$  would have power of 0.8385 for detecting  $\hat{\beta}_1 = 1$  when  $s_x = s_e = 1$ .



However, the  $t$ -statistic would be

$$t = \left( \frac{s_x}{s_e} \right) \hat{\beta}_1 \sqrt{n-1} = \left( \frac{1}{1} \right) 1 \sqrt{10} = 3.16,$$

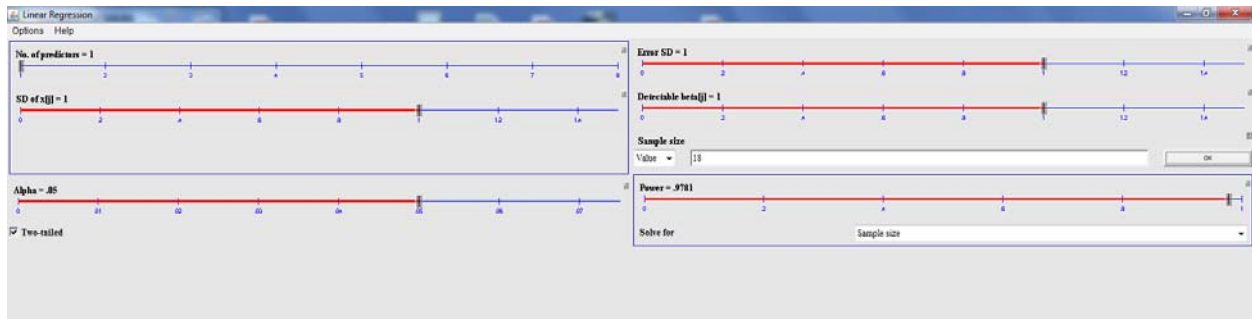
and two times the test statistic value is

$$2 * t_{0.025,9} = 2 * (2.262) = 4.524.$$

Thus, the rule of thumb of Ryan/Wetz suggests a sample size of at least  $n = 18$ , since

$$\sqrt{18} = 4.243 > 2 * 2.120 = 4.240$$

in order to achieve the desired goal of detecting  $\hat{\beta}_1 = 1$  with power at least 0.80, while at the same time hopefully ensuring that the model will have good predictive power. Note that this results in much higher power (0.9781).



## 5.2. Case 2--Multiple regression

*“The multiple linear regression case is far more difficult and in some ways is practically intractable because the values of multiple linear regression parameter estimates generally do not have any practical meaning when the predictor values are random.” (Ryan, 2009, p. 168).*

*“If the parameter estimates do not have any meaning, then determining sample size so as to be able to detect regression parameters with specific values also seems highly questionable.” (Ryan, 2012)*

Adcock (1997) stated *“What to do in the multiple-regression cased remains a topic for further study...”*

Let's begin with the simplest case: Testing one regression coefficient while adjusting for one or more others. Thus, we wish to test  $H_0 : \beta_j = 0$ , in a model containing two or more predictors.

It can be shown that the standard error of  $\hat{\beta}_j$ , the estimator of  $\beta_j$ , is now multiplied by the *variance inflation factor (VIF)*, where

$$VIF = \frac{1}{1 - R_x^2}$$

and  $R_x^2$  comes from the model  $X_j = \beta_0 + \beta_1 X_1 + \dots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \beta_p X_p$ .

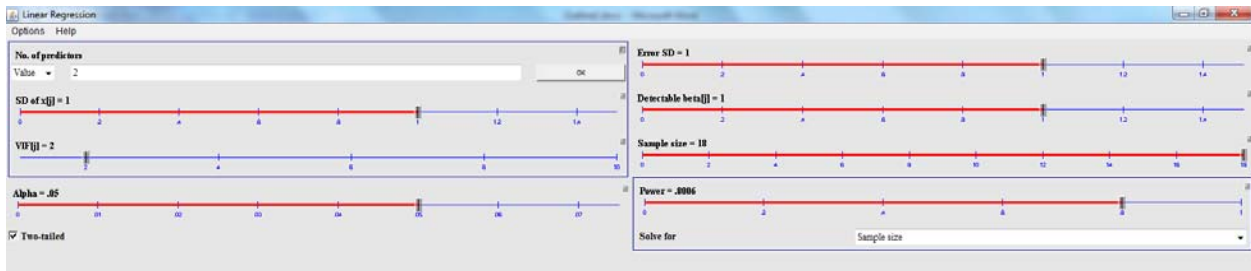
\*The more highly correlated predictor  $X_j$  is with the other predictors, the larger the standard error, and thus the larger the sample size required.

When entering 2 or more predictors using Lenth's calculator, an estimate of the VIF is required:



Notice that when  $VIF = 1$  is entered, along with the previous inputs, a sample size of 11 still achieves power of at least 0.80.

However, if  $VIF = 2$ , power drops to 0.54 when  $n = 11$ , and now  $n = 18$  would be required to have power at least 0.80.



Below is from Lenth's Help menu for the Regression dialog:

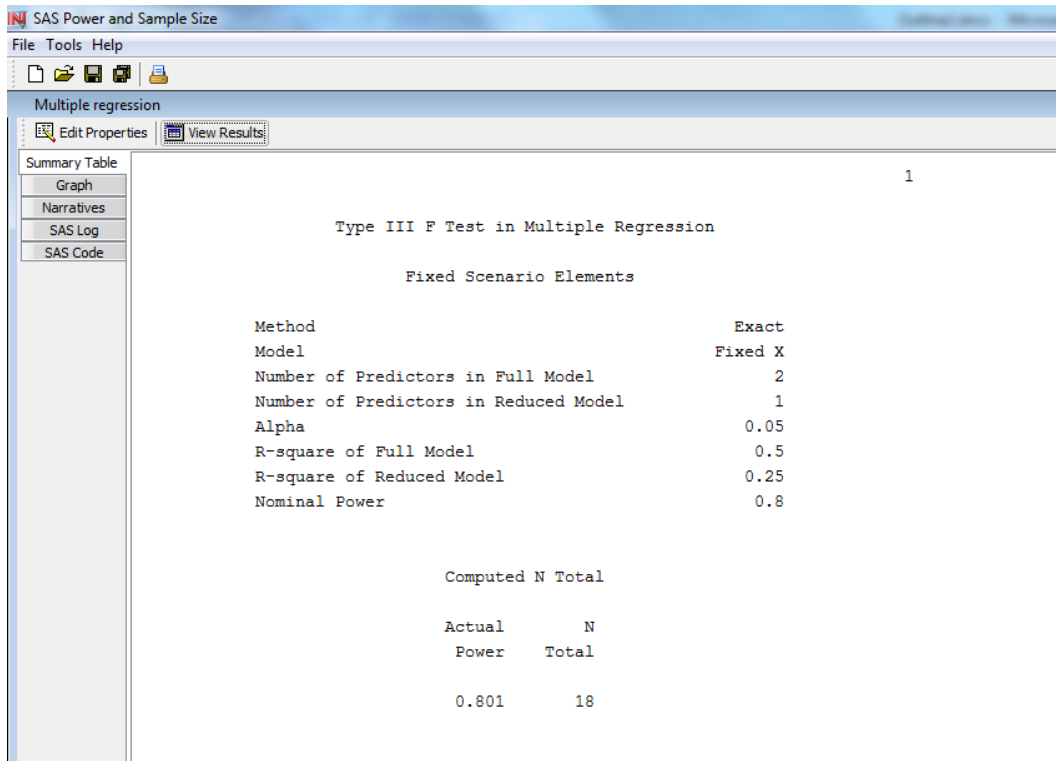
3) VIF[j]: (This slider appears only when there is more than one predictor.) Enter the variance-inflation factor for  $x[j]$ . *In an experiment where you can actually control the  $x$  values, you probably should use an orthogonal design where all of the predictors are mutually uncorrelated -- in which case all the VIFs are 1. Otherwise, you need some kind of pilot data to understand how the predictors are correlated, and you can estimate the VIFs from an analysis of those data.*

## Using $R^2$ as an effect size

Some software allows/requires input in terms of change of variance explained by the model. In the simple regression example, the test of  $H_0 : \beta_1 = 0$  is equivalent to  $H_0 : \rho = 0$ , which implies  $R^2 = 0$ . Using SAS-PSS, the same sample size as before ( $n = 11$ ) is required to detect an increase in variance explained of 0.50. Note, that  $R^2 = 0.50$  is not necessarily indicative of a model with good predictive ability, so the usefulness of the model may be questionable.

For the multiple regression case, SAS-PSS does not allow input of the detectable  $\beta_j$  and  $VIF$ . Instead, the input required is the proportion of variance ( $R^2$ ) under the full (2 predictors) and reduced (1 predictor) models. Thus, we now require an estimate of  $R^2$  under each of the full and reduced models.

Suppose we assume the reduced model will account for 25% of the variation and the addition of  $X_j$  to the model will account for an additional 25% of variance explained, then we arrive at the same sample size as using Lenth's calculator with the previous inputs.



The screenshot shows the SAS Power and Sample Size software interface. The main window displays the results of a Type III F Test in Multiple Regression. The results are summarized in a table with the following data:

| Fixed Scenario Elements               |         |
|---------------------------------------|---------|
| Method                                | Exact   |
| Model                                 | Fixed X |
| Number of Predictors in Full Model    | 2       |
| Number of Predictors in Reduced Model | 1       |
| Alpha                                 | 0.05    |
| R-square of Full Model                | 0.5     |
| R-square of Reduced Model             | 0.25    |
| Nominal Power                         | 0.8     |

| Computed N Total |   |       |
|------------------|---|-------|
| Actual Power     | N | Total |
| 0.801            |   | 18    |

If the inputs are changed to “R-square of Full Model” = 0.75 and “R-square of Reduced Model” = 0.50, only a sample size of  $n = 11$  is required. If you try some more examples, you will find that the sample size required to detect an increase of 25% in variance explained decreases as the amount of variance explained in the reduced model increases.

\*Thus, a “safer” estimate for the reduced model variance explained would be a lower bound.

### Recap

Using either the inputs required for Lenth’s calculator ( $s_x, s_e, \beta, VIF$ ), or the inputs required by SAS-PSS (R-square of Full Model, R-square of Reduced Model) probably require pilot data to obtain meaningful estimates.

[The ad-hoc method of Ryan/Wetz could also be adapted to the multiple predictor case.]

## 6. Ethical and cost considerations

Bacchetti et al. (2005) discuss ways in which ethical considerations should influence sample size.

Sampling costs are also generally not incorporated into software, so it is up to practitioners to assume the initiative. As Simon (2008) stated, “No one does it this way, but they should”

Bacchetti et al (2008) have presented alternative ways to choose sample size, focusing on justification of sample size based on cost efficiency—the ratio of a project’s scientific value relative to its cost.

1. Choose the sample size that minimizes the average cost per subject;
2. Choose sample size to minimize total cost divided by the square root of sample size.

They argue that these methods should be regarded as acceptable alternatives to conventional approaches.

## 7. Further Discussion

### 7.1. More on confidence intervals

Ramsey and Schafer (2002, Ch.23) argue for the use of confidence intervals for power/sample size analysis.

Possible outcomes, using confidence intervals:

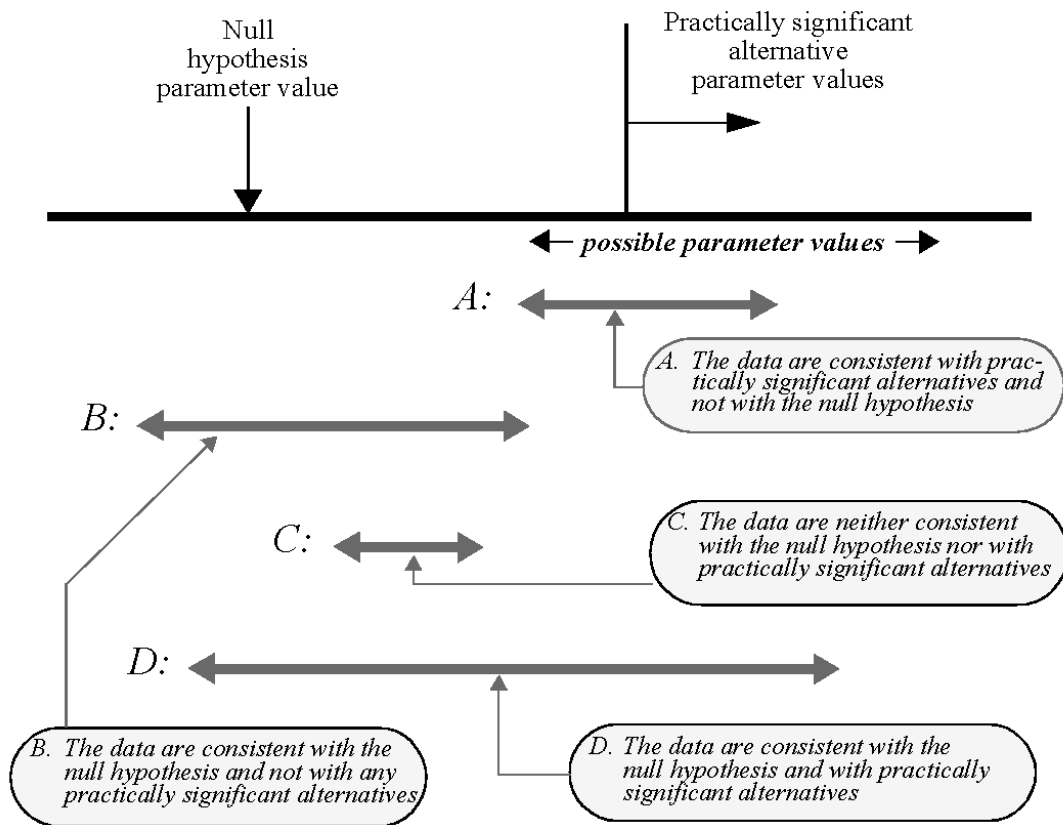
Display 23.1

p. 675

---

#### Four possible outcomes to a confidence interval procedure

---



They present sample size formulas for comparing means and proportions, as well as for estimating a regression coefficient, designed to help avoid outcome (D):



Comparing two means, confidence interval for  $\mu_1 - \mu_2$  :

$$n = \frac{8\sigma^2 (t_{\alpha/2, df})^2}{(\text{practically significant difference})^2}$$

Simple linear regression, confidence interval for slope:

$$n = \frac{4\sigma^2 (t_{\alpha/2, df})^2}{(\text{practically significant difference})^2 \sigma_x^2} + 1$$

Multiple linear regression, confidence interval for the coefficient of a single predictor:

$$n = \frac{4\sigma^2 (t_{\alpha/2, df})^2 (VIF)}{(\text{practically significant difference})^2 \sigma_x^2} + 1$$

## 7.2. Parting Thoughts

\*Essential ingredients for power analysis (Lenth, 2007):

- 1) Put science before statistics—involves serious discussion of study goals and effects of clinical importance, on the actual scale of measurement.
- 2) Pilot study—for estimating standard deviation and make sure the planned procedures actually work.

\*Practices to avoid:

1. Post-hoc power analysis (See Hoenig & Heise, 2001; Lenth, 2001, 2007)
  - a. Avoid the claim that, “the test result is not significant but the power is high, which suggests evidence to support the null hypothesis”—instead do a **test of equivalence** if that claim is desired.
  - b. Avoid “chasing significance”—computing a new sample size based on the observed effect of a study.
2. Use of “canned” or “t-shirt” effect sizes (Cohen, 1988).
  - a. Lenth (2007)—“*If only a standardized effect is sought without regard for how this relates to an absolute effect, the sample size calculation is just a pretense.*”
  - b. Lenth would also argue against using  $R^2$  in the regression setting, without considering the separate contributions of absolute effect size, variance and experimental design.

## References

- 1) Adcock, C. J. (1997). Sample size determination: A review. *The Statistician*, 46(2), 261-283.
- 2) Ahn, C. and S.-H. Jung (2005). Effect of dropouts on sample size estimates for test on trends across repeated measurements. *Journal of Biopharmaceutical Statistics*, 15, 33-41.
- 3) Agresti, A. and B. A. Coull (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119-126.
- 4) Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, 8, 17. Rejoinder: Good intentions versus CONSORT's actual effect.
- 5) Bacchetti, P., C. E. McCulloch, and M. R. Segal (2008). Simple, defensible sample sizes based on cost efficiency. *Biometrics*, 64, 577-585.
- 6) Bacchetti, P., L. E. Wolf, M. R. Segal, and C. E. McCulloch (2005). Ethics and sample size. *American Journal of Epidemiology*, 161(2), 105-110. Discussion: 111-113.
- 7) Birkett, M. A. and S. J. Day (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13, 2455-2463.
- 8) Coffey, C. S. and K. E. Muller (2001). Controlling test size while gaining the benefits of an internal pilot study. *Biometrics*, 57, 625-631.
- 9) Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2<sup>nd</sup> ed. Academic Press, New York.
- 10) Day, S. (2000). Operational difficulties with internal pilots studies to update sample size. *Drug Information Journal*, 34, 461-468.
- 11) Denne, J. S. and C. Jennison (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine*, 18(13), 1575-1585.
- 12) Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*, 3rd ed. New York: Wiley.
- 13) Heasman, P. A., I. D. M. MacGregor, Z. Wilson, and P. J. Kelly (1998). Toothbrushing forces in children with fixed orthodontic appliances. *British Journal of Orthodontics*, 27, 270-272.
- 14) Hoening, J. M. and Heise, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations in data analysis. *The American Statistician*, 55, 19-24.
- 15) Kieser, M. and T. Friede (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19, 901-911.
- 16) Kraemer, H. C., J. Mintz, A. Noda, J. Tinklenberg, and Y. A. Yesavage (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, 63, 484-489.
- 17) **Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3), 187-193. Available at <http://www.stat.uiowa.edu/techrep/tr303.pdf>**
- 18) Lenth, R. V. (2006-2009). Java applets for power and sample size (computer software). Available at <http://www.cs.uiowa.edu/~rlenth/Power>.
- 19) Lenth, R. V. (2007). Statistical power considerations. *Journal of Animal Science*, 85 (13), E24-E29.
- 20) Lipsey, M.W.; et al. (2012). [Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms](#). *United States: U.S. Dept of Education, National Center for Special Education Research, Institute of Education Sciences, NCSE 2013-3000*.
- 21) Minitab 16.0 (2010). Minitab, Inc., State College, PA.
- 22) Mudge, J. F., Baker, L. F., Edge, C. B. & Houlahan, J. E. (2012). Setting an optimal  $\alpha$  that minimizes errors in null hypothesis significance tests. *PLoS ONE* 7(2): e32734. doi:10.1371/journal.pone.0032734.
- 23) Newcombe, R. G. (2001). Statistical applications in orthodontics, part III: How large a study is needed? *Journal of Orthodontics*, 28(2), 169-172.

- 24) Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury, Pacific Grove, CA.
- 25) Ryan, T. P. (2009). *Modern Regression Methods*, 2nd ed. Hoboken, NJ: Wiley.
- 26) SAS Power and Sample Size 3.1 (2007). SAS Institute, Inc., Cary, NC.
- 27) van Belle, G. (2008). *Statistical Rules of Thumb*, 2nd edition. Hoboken: Wiley.
- 28) van Belle, G. and D. C. Martin (1993). Sample size as a function of coefficient of variation and ratio of means. *The American Statistician*, 47(3), 165-167.
- 29) Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*, 3, 1-22.
- 30) Wetz, J. M. (1964). Criteria for judging adequacy of estimation by an approximating response function. PhD thesis. Department of Statistics, University of Wisconsin.
- 31) Wittes and Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9, 65-72.
- 32) Wittes, J., O. Schabenberger, D. Zucker, E. Brittain, and M. Proschan (1999). Internal pilot studies I: Type 1 error rate of the naive t-test. *Statistics in Medicine*, 18(24), 3481-3491.
- 33) Yi, Q. and T. Panzarella (2002). Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. *Controlled Clinical Trials*, 23(5), 481-496.
- 34) Zucker, D. M. and J. Denne (2002). Sample size redetermination for repeated measures studies. *Biometrics*, 58(3), 548-559. Discussion: 60, 284-285.
- 35) Zucker, D. M., J. T. Wittes, O. Schabenberger, and E. Brittain (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, 18, 3493-3509.

References given by Mudge et al (2012) for determining critical effect size:

- Munkittrick KR, Arens CJ, Lowell RB, Kaminski GP (2009). A review of potential methods of determining critical effect size for designing environmental monitoring programs. *Environ Toxicol Chem* 28: 1361–1371.
- Nakagawa S, Cuthill IC (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev* 82: 591–605.
- Huberty CJ (2002). A history of effect size indices. *Educ Psychol Meas* 62:227–240.

#### **Additional reading** (Courtesy of T. P. Ryan)

- 1) Bartlett, J. E., II, J. W. Kotrik, and C.C. Higgins (2001). Organizing research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19(1), 43-50.
- 2) Basaga~na, X. and D. Spiegelman (2010). Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure. *Statistics in Medicine*, 29, 181-182.
- 3) Basaga~na, X., X. Liao, and D. Spiegelman (2011). Power and sample size calculations for longitudinal studies estimating a main effect of a time-varying response. *Statistical Methods in Medical Research*, 20, 471-487.
- 4) Baskerville, N. B., W. Hogg, and J. Lemelin (2001). The effect of cluster randomization on sample size in prevention research. *Journal of Family Practice*, 50, 241-246.
- 5) Bassiakos, Y. and P. Katerelos (2006). Sample size calculation for the therapeutic equivalence problem. *Communications in Statistics: Simulation and Computation*, 35, 1019-1026.
- 6) Bochmann, F., Z. Johnson, and A. Azuara-Blanco (2007). Sample size in studies on diagnostic accuracy in ophthalmology: A literature survey. *British Journal of Ophthalmology*, 91, 898-900.

- 7) Branscum, A. J., W. O. Johnson, and I. A. Gardner (2007). Sample size calculations for studies designed to evaluate diagnostic test accuracy. *Journal of Agricultural, Biological, and Environmental Statistics*, **12**(1), 112-127.
- 8) Buderer, N. M. (1996). Statistical methodology: Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine*, **3**(9), 895-900.
- 9) Cai, G., X. Lin, and K. Lee (2010). Sample size determination with false discovery rate adjustment for experiments with high-dimensional data. *Statistics in Biopharmaceutical Research*, **2**(2), 165-174.
- 10) Carley, S., S. Dosman, S. R. Jones, M. Harrison (2005). Simple nomograms to calculate sample size in diagnostic studies. *Emergency Medical Journal*, **22**, 180-181.
- 11) Cheng, D., A. J. Branscum, and J. D. Stamey (2010). Accounting for response misclassification and covariate measurement error improves power and reduces bias in epidemiologic studies. *Annals of Epidemiology*, **20**(7), 562-567.
- 12) Cohen, M. P. (2005). Sample size considerations for multilevel surveys. *International Statistical Review*, **73**(3), 279-287.
- 13) D'Amico, E. J., T. B. Neilands, and R. Zambarano (2001). Power analysis for multivariate and repeated measures designs: A flexible approach using the SPSS MANOVA procedure. *Behavior Research Methods, Instruments, and Computers*, **33**(4), 479-484.
- 14) Dobbin, K. and R. Simon (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**(1), 27-38.
- 15) Dupont, W. (1988). Power calculations for matched case-control studies. *Biometrics*, **44**, 1157-1168.  
Edwards, B. J., C. Haynes, M. A. Levenstein, S. J. Finch, and D. Gordon (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genetics*, **6**(18).
- 16) Efrid, J. T. and K. Aliminetti (2005). Computing exact power for multivariate repeated measurements design. *WUSS Proceedings*, SAS Institute, San Jose, CA.
- 17) Fang, H.-B., G.-L. Tian, W. Li, and M. Tang (2009). Design and sample size for evaluating combinations of drugs of linear and loglinear dose-response curves. *Journal of Biopharmaceutical Statistics*, **19**(4), 625-640.
- 18) Ferreira, J. A. and A. H. Zwinderman (2006). Approximate sample size calculations with microarray data: An illustration. *Statistical Applications in Genetics and Molecular Biology*, **5**(1), article 25.
- 19) Fitzpatrick, B. M. (2009). Power and sample size for nested analysis of molecular variance. *Molecular Ecology*, **18**, 3961-3966.
- 20) Foppa, I. and D. Spiegelman (1997). Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology*, **146**(7), 596-604.
- 21) Freedman, K. B., S. Back, and J. Bernstein (2001). Sample size and statistical power of randomised, controlled trials in orthopaedics. *The Journal of Bone and Joint Surgery*, **83-B**(3), 397-402.
- 22) Gadbury, G. L., G. P. Page, J. Edwards, T. Kayo, T. A. Prolla, R. Weindruch, P.A. Permana, J. D. Mountz, and D. B. Allison (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, **13**(4), 325-338.
- 23) Gauderman, W. J. (2002). Sample size requirements for matched-case control studies of gene-environment interaction. *Statistics in Medicine*, **21**, 35-50.
- 24) Gupta, P. L. and R. D. Gupta (1987). Sample size determination in estimating a covariance matrix. *Computational Statistics & Data Analysis*, **5**, 185-192.
- 25) Gustafson, P. (2006). Sample size implications when biases are modeled rather than ignored. *Journal of the Royal Statistical Society, Series A*, **169**, 865-881.

- 26) Hancock, G. R. and M. J. Freeman (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, **61**(5), 741-758.
- 27) Hanley, J. A., I. Csizmadi, and J.-P. Collet (2005). Two-stage case-control studies: Precision of parameter estimates and consideration in selecting sample size. *American Journal of Epidemiology*, **162**, 1225-1234.
- 28) Hedeker, D., R. D. Gibbons, and C. Waternaux (1999). Sample size estimation for longitudinal designs with attrition: time-related concepts between two groups. *Journal of Educational and Behavioral Statistics*, **24**(1), 70-93.
- 29) Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in metaanalysis. *Psychological Methods*, **6**(3), 203-217.
- 30) Hedges, L. V. & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, **9**(4), 424-445.
- 31) Hwang, S.-J., T. H. Beaty, K.-Y. Liang, J. Coresh, and M. J. Khoury (1994). Minimum sample size estimation to detect gene-environment interaction in case control designs. *American Journal of Epidemiology*, **140**, 1029-1037.
- 32) Johnson, W. O., C.-L. Su, I. A. Gardner, and R. Christensen (2004). Sample size calculations for surveys to substantiate freedom of populations from infectious agents. *Biometrics*, **60**, 165-171.
- 33) Jung, S.-H. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, **6**, 157-169.
- 34) Jung, S. H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, **21**(14), 3097-3104.
- 35) Jung, S. H. and C. Ahn (2003). Sample size estimation for GEE method comparing slopes in repeated measurements data. *Statistics in Medicine*, **22**(8), 1305-1315.
- 36) Jung, S. H., H. Bang, and S. S. Young (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, **6**, 157-159.
- 37) Kosinski, A. S., Y. Chen, and R. H. Lyles (2010). Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. *Statistics in Medicine*, **29**, 1572-1579.
- 38) Kozlitina, J., C. Xing, A. Pertsemliadis, and W. S. Schucany (2010). Power of genetic association studies with fixed and random genotype frequencies. *Annals of Human Genetics*, **74**(5), 429-438.
- 39) Kumar, R. and A. Indrayan (2002). A nomogram for single-stage cluster-sample surveys in a community for estimation of a prevalence rate. *International Journal of Epidemiology*, **31**(2), 463-467.
- 40) Lee, M. L. and G. A. Whitmore (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, **21**(23), 3543-3570.
- 41) Li, J. and J. Fine (2004). On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. *Statistics in Medicine*, **23**, 2537-2550.
- 42) Lin, W.-J., H.-M. Hsueh, and J. J. Chen (2010). Power and sample size estimation in microarray studies. *BMC Bioinformatics*, **11** (Supplement 1), S52.
- 43) Littell, R. C., J. Pendergast, and R. Natarajan (2000). Modeling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793-1819.
- 44) Liu, J.-P., H. Hsueh, and J. J. Chen (2002). Sample size requirements for evaluation of bridging evidence. *Biometrical Journal*, **44**, 969-981.
- 45) Liu, P. and J. T. G. Hwang (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, **23**(6), 739-746.
- 46) Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics*, **28**, 231-248.
- 47) Lu, K., D. V. Mehrotra and G. Liu (2009). Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine*, **28**, 679-699.

- 48) Luan, J. A., M. Y. Wong, N. E. Day, and N. J. Wareham (2001). Sample size determination for studies of gene-environment interaction. *International Journal of Epidemiology*, **30**(5), 1035-1040.
- 49) Lubin, J. H., M. H. Gail, and A. G. Ershow (1987). Sample size and power for case-control studies when exposures are continuous. *Statistics in Medicine*, **7**(3), 363-376.
- 50) Maccallum, R. C., M. W. Browne, and H. M. Sugawara (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, **1**(2), 130-149.
- 51) Moerbeek, N. and W. K. Wong (2008). Sample size formula for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine*, **27**, 2850-2864.
- 52) Moon, H., J. J. Lee, H. Ahn, and R. G. Nikolova (2002). A web-based simulator for sample size and power estimation in animal carcinogenicity studies. *Journal of Statistical Software*, **7**, 1-36.
- 53) Mukherjee, S. P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov (2004). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, **10**(2), 119-142.
- 52) Müller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, **99**(468), 990-1001.
- 53) Muller, K. E., L. LaVange, S. L. Ramey, and C. T. Ramey (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, **87**, 1209-1226.
- 54) Musonda, P., C. P. Farrington, and H. J. Whitaker (2006). Sample sizes for self-controlled case series studies. *Statistics in Medicine*, **25**, 2618-2631.
- 55) Naing, L., T. Winn, and B. N. Rusli (2006). Practical issues in calculating the sample size for prevalence studies. *Archives of Orafacial Sciences*, **1**, 9-14.
- 56) Nam, J.-M. (1992). Sample size determination for case-control studies and the comparison of stratified and unstratified analyses. *Biometrics*, **48**, 389-395.
- 57) Noble, R. B., J. A. Bailer, S. R. Kunkel, and J. K. Straker (2006). Sample size requirements for studying small populations in gerontology research. *Health Services & Outcomes Research Methodology*, **6**, 59-67.
- 58) Obuchowski, N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, **7**, 371-392.
- 59) Ogungbenro, K and L. Aarons (2010). Sample-size calculations for multi-group comparison in population pharmacokinetic experiments. *Pharmaceutical Statistics*, **9**, 255-268.
- 60) Ogungbenro, K and L. Aarons (2010). Sample size/power calculations for population pharmacodynamic experiments involving repeated measurements. *Journal of Biopharmaceutical Statistics*, **20**(5), 1026-1042.
- 61) Ogungbenro, K, L. Aarons, and G. Graham (2006). Sample size calculations based on generalized estimating equations for population pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics*, **16**, 135-150.
- 62) Orr, M. and P. Liu (2009). Sample size estimation while controlling false discovery rate for microarray experiments using `ssize.fdr` package. *The R Journal*, **1**(1), 47-53.
- 63) Overall, J. E. (1996). How many repeated measurements are useful? *Journal of Clinical Psychology*, **52**(3), 243-252.
- 64) Overall, J. E. and S. R. Doyle (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials*, **15**(2), 100-123.
- 65) Pennington, M. and J. H. Volstad (1991). Optimum size of sampling unit for estimating the density of marine populations. *Biometrics*, **47**, 717-723.
- 66) Roy, A., D. K. Bhaumik, S. Aryal, and R. D. Gibbons (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, **63**(3), 699-707.

- 67) Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, **53**, 1253-1261.
- 68) Sethuraman, V. S., S. Leonov, L. Squassante, T. R. Mitchell, and M. D. Hale (2007). Sample size calculation for the Power Model for dose proportionality studies. *Pharmaceutical Statistics*, **6**, 35-41.
- 69) Shao, Y. and C. Tseng (2007). Sample size calculation with dependence adjustment for FDR-control in microarray studies. *Statistics in Medicine*, **26**, 4219-4237.
- 70) Shun, Z., Y. He, Y. Feng, and M. Roessner (2009). A unified approach to flexible sample size design with realistic constraints. *Statistics in Biopharmaceutical Research*, **1**(4), 388-398.
- 71) Siqueira, A. L., A. Whitehead, S. Todd, and M. M. Lucini (2005). Comparison of sample size formulae for 2 x 2 cross-over designs applied to bioequivalence studies. *Pharmaceutical Statistics*, **4**(4), 233-243.
- 72) Sparks, T. H., J. O. Mountford, S. J. Manchester, P. Rothery, and J. R. Treweek (1997). Sample size for estimating species lists in vegetation surveys. *The Statistician*, **46**(2), 253-260.
- 73) Su, G. (2005). Sample size and power analysis for endometrial safety studies. *Journal of Biopharmaceutical Statistics*, **15**, 491-499.
- 74) Tan, M. Y., H.-B. Fang, and G.-L. Tian (2009). Dose and sample size determination for multi-drug combination studies. *Statistics in Biopharmaceutical Research*, **1**, 301-316.
- 75) Tibshirani, R (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, **7**, 106.
- 76) Tsai, C.-A., S.-J. Wang, D.-T. Chen, and J. J. Chen (2005). Sample size for gene expression microarray experiments. *Bioinformatics*, **21**(8), 1502-1508.
- 77) Tseng, C.-H. and Y. Shao (2010). Sample size analysis for pharmacogenetic studies. *Statistics in Biopharmaceutical Research*, **2**(3), 319-328.
- 78) Tu, X. M., J. Kowalski, J. Zhang, K. G. Lynch, and P. Crits-Christoph (2004). Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine*, **23**(18), 2799-2815.
- 79) Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics*, **31**, 643-649.
- 80) van Iterson, M., P. A. C.'t Hoen, P. Pedotti, G. J. E. J. Hooiveld, J. T. den Dunnen, G. J. B. van Ommen, J. M. Boer, R. X. Menezes (2009). Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, **10**, 449.
- 81) Walters, S. J. (2004). Sample size and power estimation for studies with health related quality of life outcomes: A comparison of four methods using the SF-36., *Health and Quality of Life Outcomes* **2** 26. (Open access; available at [www.hqlo.com/content/2/1/26](http://www.hqlo.com/content/2/1/26) and [www.hqlo.com/content/pdf/1477-7525-2-26.pdf](http://www.hqlo.com/content/pdf/1477-7525-2-26.pdf))
- 82) Wang, S. and H. Zhao (2003). Sample size needed to detect gene-gene interactions using association designs. *American Journal of Epidemiology*, **158**, 899-914.
- 83) Wei, C., J. Li, and R. E. Bumgartner (2004). Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, **5**(1), 87.
- 84) Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, **9**(6), 476-487.
- 85) Xie, J., T. T. Cai, and H. Li (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, **98**(2), 273-290.