

# Module 6 – Sample Size Considerations



**Original Author:** Jonathan Berlowitz, PhD

**PERC Reviewer:** Timothy Lynch, MD

---

## Table of Contents

Table of Contents.....	1
Overview .....	2
Introduction .....	2
Objectives.....	2
Key Concepts.....	2
Activities.....	3
Quick Links .....	3
Task Checklist.....	3
Readings.....	3
Module 6: Sample Size Considerations.....	5
Background .....	5
The First Set of Questions Before “The Question” .....	7
Sample Size Estimation for Descriptive Studies.....	9
Sample Size Estimation for Comparative Studies .....	13
Sample size estimation concerns ensuring enough data so as to keep the probabilities of Type I and Type II errors ( $\alpha$ and $\beta$ ) at suitable levels.....	15
Two-Tailed versus One-Tailed Tests .....	15
Statistical Significance versus Practical / Clinical Significance (or importance).....	16
Sample Size Estimation: Answering the QUESTION? .....	18
Strategies for Minimizing Sample Size and Maximizing Power .....	24
Summary .....	25
Examples .....	27
Assignment.....	29

---

## **Overview**

### ***Introduction***

One of first questions a researcher asks is, “How many subjects do I need?” This is a simple question, with a somewhat complicated answer. The answer depends on the purpose of the research.

Descriptive studies involve consideration of precision or margin of error; comparative studies involve power calculations. The answer also depends on the type of data being collected. Fortunately, the calculations are not onerous and a variety of tools exist to help in the decision. Jacob Cohen, author of the landmark book on power analysis wrote, “Since statistical significance is so earnestly sought and devoutly wished for by behavioural scientists, one would think that the a priori probability of its accomplishment would be routinely determined and well understood. Quite surprisingly, this is not the case.”

### ***Objectives***

You will be able to:

- Understand the logic of statistical inference regarding both margin of error and power
- Identify a primary outcome measure and determine whether it is measurement or categoric in nature
- Determine sample size for descriptive studies when the primary outcome is measurement scale or categoric
- Determine sample size for two-group comparative studies when the primary outcome is measurement scale or categoric
- Determine sample size for more complicated designs
- Understand the limitations of sample size calculations
- Learn how to balance statistical and practical sample size needs

### ***Key Concepts***

- Know whether the sample size is to be based on precision or power
- Learn how many respondents are needed for survey design
- Learn how to compute sample size based on power
- Strategies for minimizing sample size
- Find suitable “sample size calculators” on the web

## **Activities**

- Decide on a primary outcome measure which will be the basis for sample size determination
- Determine sample sizes necessary for various research scenarios
- Be able to apply these calculations to your research project

## **Quick Links**

### **Sample Size Calculators on the Web, in increasing order of complexity:**

- Rollin Brant's simple but effective sample size calculators  
<http://newton.stat.ubc.ca/~rollin/stats/ssize/>
- Deals with epidemiological applications  
[http://hedwig.mgh.harvard.edu/sample\\_size/size.html](http://hedwig.mgh.harvard.edu/sample_size/size.html)
- Nicely designed java-applets  
<http://www.stat.uiowa.edu/~rlenth/Power/index.html>

An excellent index page for many other on-line statistical tools

<http://www.utexas.edu/its/rc/world/stat/on-line.html>

### **Overview of Sample Size Determination**

- <http://www.columbia.edu/~mvp19/RMC/M6/ssfd.pdf>

## **Task Checklist**

1. Compute required sample sizes for various scenarios described in the module
2. Write a sample size justification paragraph suitable for a grant proposal or research manuscript, for any of the scenarios or for your research project.

## **Readings**

### **Main reference (required):**

- Hulley, SB, Cummings, SR, et al. (2001). *Designing Clinical Research*, Second Edition; Lippincott Williams and Wilkins. -- Chapter 6. Estimating the sample size

### **Supplementary references (optional):**

- Cohen, Jacob (1977). *Statistical Power Analysis for the Behavioral Sciences*, Revised Edition; Academic Press. [A newer edition has been published] (This is the granddaddy of books on this subject. If you think this module is detailed, have a look at the book! If you can find a copy, read the discussion on small, medium and large effect sizes. This may help in your understanding of what constitutes clinically important effects.)

- Kraemer HC, & Thiemann S (1987). *How Many Subjects? Statistical Power Analysis in Research*; Sage Publications. (One of the earliest books on the subject, it is only about 100 pages in length. I go back to this one regularly for help in explaining the issues in sample size estimation.)
- Lipsey, Mark (1990). *Design Sensitivity: Statistical Power for Experimental Research*; Sage Publications. (This is quite a readable book, in the Sage style.)

## **Module 6: Sample Size Considerations**

### ***Background***

Questions about sample size are ubiquitous in research. Too small a sample will yield scant information; but ethics, economics, time and other constraints require that a sample size not be too large.

“How many subjects do I need?” Neither 7 nor 30 nor any number is an all-purpose answer. A sample size of 30 is a “large sample” in some textbook discussions of “normal approximation”; yet 30,000 observations still may be too few to assess a rare, but serious teratogenic effect. The best first response to “how many?” may be not a number, but a sequence of further questions. A study’s size and structure should depend on the research context, including the researcher’s objectives and proposed analyses.

If a survey is to be carried out for descriptive purposes such as assessing the prevalence of some characteristic, the sample size is based on the required precision of the prevalence estimate. For example, why are so many national opinion polls based on samples of approximately 1000 responses? And if the poll results are “valid” on a national level, how “valid” are they on a provincial level?

If there is a comparative aspect to the study, the sample size is based on how detailed a comparison is desired. Detecting very small differences requires larger samples than detecting large differences. The appropriate sample size also depends on the precision or variability of the data. Fewer replications are needed if a response variable changes little from one measurement to the next, than if the response varies wildly.

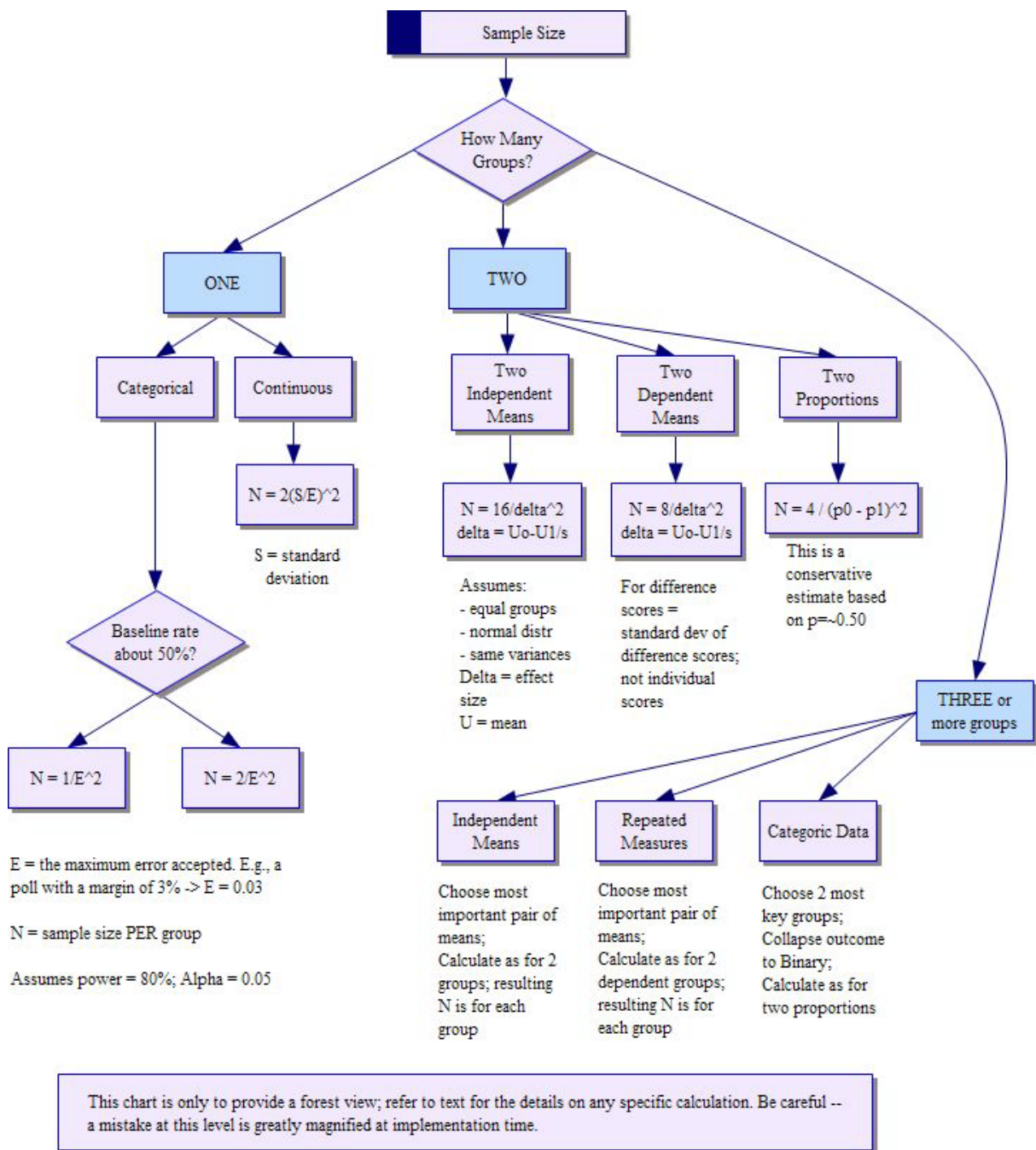
Sample sizes should also be computed with attention to dropout rates. If 100 subjects are enrolled at the beginning of a study, how many can be expected to remain at the end of the study, if a two-year or five-year follow-up is required? This “attrition rate” must be considered.

As well, more sophisticated analytic techniques may require larger samples than simple techniques.

A historical note: A landmark survey paper in the New England Journal of Medicine in 1978, by Freiman et al., brought to people's attention the problem that "negative" findings from clinical trials may be a result of small sample sizes. Since that time, grant proposals and journal articles have included a discussion of the sample size issue.

In this module we will learn the questions to ask, and what to do with the corresponding answers, to provide the answer to THE QUESTION, "How many subjects do I need?"

**One cautionary note:** "How many subjects do I need?" is linked to the companion question, "How many subjects can I afford to get?" The final decision on sample size must pay attention to the various constraints on recruitment. Just as no responsible consumer should go shopping without first setting the limits on how much is available to spend, no decision on sample size should be made without assessing the corresponding affordability.



E = the maximum error accepted. E.g., a poll with a margin of 3% -> E = 0.03

N = sample size PER group

Assumes power = 80%; Alpha = 0.05

This chart is only to provide a forest view; refer to text for the details on any specific calculation. Be careful -- a mistake at this level is greatly magnified at implementation time.

### The First Set of Questions Before “The Question”

“THE QUESTION” “How many subjects do I need?” Two preliminary questions must be asked: Question 1. Is the study descriptive or is the study comparative? Question 2. Is the primary outcome variable a measurement variable (a.k.a. interval or continuous) or is the primary outcome variable a categoric variable? These are discussed below.

#### Question 1. Is the study descriptive or is the study comparative?

Descriptive studies include surveys to assess prevalence, needs assessments, chart reviews, etc. that have as a main aim the estimation of rates, proportions and means in a population with a secondary aim being to examine whether the rates are related to demographic variables (i.e. a correlational analysis). For example, a survey may be undertaken to assess the extent of doughnut consumption in Belltown. The main results to be reported might be the percentage of residents who consume doughnuts on a daily basis, or the mean number of doughnuts consumed by a resident per week. Follow-up analysis might examine whether the consumption rates depend on the sex or age of the resident. Sample size determination for descriptive studies is based on confidence intervals; that is, the level of precision required in providing estimates of the rates, proportions and means.

Comparative studies include case-control designs, randomized clinical trials, etc. where a comparison between two or more groups is the key analysis. The main aim here is to establish whether there are statistically significant differences between groups with respect to some key outcome variable. Sample size determination for comparative studies is based on hypothesis tests and power, that is, the probability of being able to find differences when they do, in fact, exist.

In brief, the first question asks, “Are P-values relevant here?”

Note that descriptive studies often lead to comparative studies; in fact, post hoc analysis often involves informal inference and model-building to examine relationships among variables. The sample size estimation should also take this into account and be sufficiently large for analysis of future questions.

**Question 2. Is the primary outcome variable a measurement variable (a.k.a. interval or continuous) or is the primary outcome variable a categoric variable?**

Choosing a primary outcome variable is analogous to choosing dessert in a restaurant. You may have many favorites, but when the server comes to take your order you have to settle on one, although you can probably procure a taste of everyone else’s dessert. Everyone else’s desserts are analogous to secondary outcome variables. You will be able to assess them too, but your main assessment – what determines whether the research question was answered, or whether the dessert was a success – depends on a single variable.



A measurement variable is one where the characteristic is assessed on a scale with many possible values representing an underlying continuum (e.g. age, height, blood pressure, pain (on a visual analogue scale)). It involved a measuring process and usually requires some sort of “instrumentation” (e.g. ruler, stopwatch, biochemical analysis, psychometric tool). Measurement variables are usually summarized with a mean or median.

A categoric variable involves classification of subjects into one of a number of categories on the basis of a characteristic. There can be two categories only (binary variable), multiple categories where order does not matter (nominal variable) or multiple categories where order does matter (ordinal variable). Categoric variables are usually summarized with proportions.

### ***Sample Size Estimation for Descriptive Studies***

If the answer to Question 1 in the previous section was “descriptive study”, you’re in the right place. No further questions need to be answered. The following section will derive these two formulas.

	Categoric	Measurement
Descriptive Studies (sample size is based on margin of error, E, in confidence intervals)	$N = 1 / E^2, \text{ for } P \text{ near } 0.5$ <p style="text-align: center;">or</p> $N = (2 / E)^2 P(1-P),$ <p style="text-align: center;">for P near 0 or 1.</p>	$N = (2S / E)^2$ <p style="text-align: center;">where S is the standard deviation of the variable</p>

A confidence interval is a range of likely or plausible values of the population characteristic of interest. For example, a sample survey can be used to give a range of values that the true population proportion or population mean is expected to lie within. The intervals can be constructed to provide greater or lesser levels of confidence; however, the usual choice is 95% (with 90% and 99% useful in certain situations). For more information on confidence intervals, see the Modules entitled Analysis I and II.

Confidence intervals usually take the form:

(Point estimate)  $\pm$  (Margin of error)

- The point estimate is a value computed from the sample; for example, the sample mean or sample proportion.

- The margin of error (or “plus or minus number”) is a value computed from a variety of components – the level of confidence (e.g. 95%), the variability in the outcome variable, and the sample size.

Confidence intervals are used to estimate sample sizes as follows.

>>> When interest is in a population mean (i.e. the primary outcome variable is **measurement/continuous**), the total number of subjects required (N) is:

$$N = 4 z_{\alpha}^2 S^2 / W^2$$

where S is the standard deviation of the variable, W is the width of the confidence interval (equal to twice the “margin of error”), and  $z_{\alpha}$  is a value from the normal distribution related to and representing the confidence level (equal to 1.96 for 95% confidence). The table in Appendix 6.D (p 90, Hulley) provides the sample size for common values of W/S and three choice of confidence level.

The formula can be rewritten as:

$$N = (z_{\alpha} S / E)^2$$

where E is the “margin of error” (half the width, W).

As an approximation, for 95% confidence, use the value of 2 for  $z_{\alpha}$  (instead of 1.96) – remember that this is an approximation, after all! Then the formula is a very concise and easily remembered:

$$N = (2S / E)^2$$

That is “twice the standard deviation over the margin of error, all squared”.

Where does the value of S come from? There are a number of sources, including previously published research or a pilot study. When these sources fail, as in the case of brand-new research, with a new instrument or a new population under study, a rough approximation can be made using the six-sigma rule for bell-shaped distributions; the standard deviation is approximately the range (maximum minus minimum) divided by six.

Here is an online calculator that allows you to determine the sample size for a measurement continuous variable in a single sample study: <http://www.surveysystem.com/sscalc.htm>

>>> When interest is in a population **proportion** (i.e. the primary outcome variable is categorical – specifically, binary), the total number of subjects required (N) is:

$$N = 4 z_{\alpha}^2 P(1-P) / W^2$$

where P is the expected proportion who have the characteristic of interest, W is the width of the confidence interval (equal to twice the “margin of error”), and  $z_{\alpha}$  is a value from the normal distribution related to and representing the confidence level (equal to 1.96 for 95% confidence). Note that this formula looks like the one for measurement data except that  $S^2$  has been replaced by  $P(1-P)$ .

The table in Appendix 6.E (p 91, Hulley) provides the sample size for common choices of P and W, and three choices of confidence level.

$$N = (z_{\alpha} / E)^2 P(1-P)$$

where E is the “margin of error” (half the width, W).

As an approximation, for 95% confidence, use the value of 2 for  $z_{\alpha}$  (instead of 1.96) – remember that this is an approximation, after all! Also, use the most conservative value of P, which is 0.5. Then the formula is a very concise and easily remembered:

$$N = 1 / E^2$$

That is, “one over the square of the margin of error”.

This formula can also be easily rearranged to get:  $E = 1 / \sqrt{N}$ ;

That is, the margin of error is one over the square root of the sample size.

For example, if the sample size is 100, the margin of error is  $\pm 10\%$ ; for a sample size of 400 the margin of error is  $\pm 5\%$ ; and for a sample size of 1000, the margin of error is  $\pm 3\%$ .

Note that doubling the sample size from 1000 to 2000 only reduces the margin of error to  $\pm 2\%$ , not much improvement in precision for double the effort. That explains why so many national opinion polls are about 1000 in size.

**Notes:**

If the expected proportion is more than half, then plan the sample size based on the proportion expected NOT to have the characteristic. That is, switch the roles of P and 1-P.

If P or 1-P is very close to 0 or 1 (i.e. the characteristic of interest is rare or happens most of the time), the sample size formula of  $N = 1 / E^2$  is not appropriate. Instead you need to use the fuller version seen earlier:

$$N = (2 / E)^2 P(1-P)$$

[Note for the obsessive-compulsives among you: These formulas assume that the population is “infinite” (i.e. very large) in comparison to the sample. There is a finite population correction factor that will come into play when the final confidence intervals are being constructed. But it can be safely ignored in calculating sample size for a survey.]

A confidence interval for the mean should be based on at least twelve observations. The width of a confidence interval, involving estimate of variability and sample size decreases rapidly until 12 observations are reached and then decreases less rapidly.

Exercise: Go to <http://www.surveysystem.com/sscalc.htm> where there is a calculator used for this type of calculation. Use this calculator to determine the sample size for a survey where you will determine the proportion of people in Belltown who eat doughnuts. You want your estimate of the true proportion to be accurate to  $\pm 6\%$ . (Note that, unfortunately, this calculator calls this the “confidence interval”, a poor choice of terminology – it should be called the margin of error.) The population of Belltown is 65,000. Try out the calculator for other choices of population size; what would be the required sample size if the population of Belltown were only 1500? 15,000? 150,000? 1,500,000? What would you conclude about the role of the population size in these sample size calculations.

## ***Sample Size Estimation for Comparative Studies***

If the answer to Question 2 in the previous section was “comparative study”, you’re in the right place. This section will present:

- A review of hypothesis testing
- Baseline information – Questions to be answered for comparative studies
- Question 1: What is an acceptable significance level (alpha)
- Question 2: How large a power is needed?
- Question 3: How large is the variability in the effect of interest?
- Question 4. What is the smallest detectable effect of interest
- Calculating the sample size for comparative studies

A statistical “test” always challenges some hypothesis. A new treatment is investigated by testing that the given treatment has no effect. A comparative study tests a hypothesis that two groups under different treatment exhibit no differences in responses. We describe the results as “significant” or “positive” when such a challenge has been successful and the tested hypothesis overthrown (i.e. the null hypothesis is rejected).

“Significance” refers to the events and data that were actually observed, but which had small probability (P-value) according to the null hypothesis (so the null hypothesis is rejected as being incompatible with the data).

Before proceeding to sample size estimation we need to review the basic concepts of hypothesis testing.

### **Review of Basic Concepts of Hypothesis Testing**

Hypothesis testing requires, first of all, and not surprisingly, hypotheses! That is, two competing claims about a parameter or parameters (characteristics of a population). In the context of sample size estimation the parameters are usually the mean or proportion of the key outcome variable of interest. The null hypothesis is the status quo hypothesis, the position of no difference, no effect, or no change. The alternative hypothesis is often referred to as the research hypothesis. It represents a difference between groups, a real effect, and an abandonment of the status quo.

A hypothesis test culminates with a conclusion about which of the two hypotheses is supported by the available data. The conclusion can either be correct or incorrect. And statisticians, who have their ignorance better organized than ordinary mortals, have classified the ways in which the conclusion can be correct or incorrect. Errors in the conclusion are imaginatively called either Type I or Type II.

A Type I error occurs when the null hypothesis is rejected, but in fact the null hypothesis is actually true. That is, the conclusion is that there is a significant difference when in fact there really isn't. A Type I error can be thought of as a "false positive".

A Type II error occurs when the null hypothesis is accepted, but in fact the null hypothesis is actually false. That is, the conclusion is that there is no difference when in fact there really is a difference. A Type II error can be thought of as a "false negative".

Next we define alpha ( $\alpha$ ) as the probability of making a Type I error. It is also known as the significance level. Usually  $\alpha$  is set at 0.05 (keeping it consistent with  $1 - \alpha$  or .95 or 95% in the context of confidence intervals).

And we define beta ( $\beta$ ) as the probability of making a Type II error. Although  $\beta$  doesn't have another name,  $1 - \beta$  does. It is known as power.

Power is the probability of correctly rejecting the null hypothesis; for example, concluding that there was a difference when, in fact, there really was one!

Sample size calculations are often called power calculations, which tells you how crucial the concept of power is to the whole exercise.

Aside: A Type III error has been referred to as getting the right answer to the wrong question or to a question nobody asked!

The following two-by-two table summarizes the previous concepts and quantities.

<u>Truth</u>			
		No Difference	Difference
<u>Study</u>	Accept $H_0$	$1 - \alpha$	$\beta$
	Reject $H_0$	$\alpha$	$1 - \beta$

A useful analogy is to our Western legal system. In our system a defendant is “innocent until proven guilty”. The null hypothesis is “not guilty”; the alternative hypothesis is “guilty”. The onus is on the investigator (i.e. the prosecution) to present the evidence to convince the judge or jury to abandon the null hypothesis in favour of the alternative. If the data are convincingly more consistent with the alternative hypothesis, the judge or jury (barring legal technicalities and theatrics) must conclude that the defendant is guilty.

The conclusion, whichever way it goes, may be the right one or the wrong one. Convicting the guilty or acquitting the innocent are correct decisions. However, convicting an innocent person is a Type I error, while acquitting a guilty person is a Type II error. Neither of these errors is desirable (in this case a Type I error is the worse of the two, but there are other situations where a Type II error is the worse).

We would rather not make any errors. Notice however the problem that this presents. In order not to make ANY Type I errors we would have to acquit everyone, which would lead to a high rate of Type II errors. In order not make ANY Type II errors we would convict rather a lot of innocent people along the way. Hypothesis testing, therefore tries to keep both error rates under control, and this is accomplished by collecting more and more evidence (what a non-legal researcher would call data).

Sample size estimation concerns ensuring enough data so as to keep the probabilities of Type I and Type II errors ( $\alpha$  and  $\beta$ ) at suitable levels.

***Two-Tailed versus One-Tailed Tests***

In inference that investigates whether there is a difference between two groups, there are two approaches to formulating the alternative hypothesis. Either you know the direction of the difference in advance of doing the study or you do not. A one-tailed test specifies the direction of the difference in advance. A two-tailed test does not specify the direction of the difference. For sample size estimation stick to two-tailed alternative hypotheses! For example, when comparing a

new therapy to a standard therapy one might be convinced that the new therapy could only be better! But, examples abound where a new therapy under study was actually worse. Think about the case when Coca Cola introduced “New Coke” expecting it to improve sales. The huge negative public outcry was completely unexpected by Coca Cola.

### ***Statistical Significance versus Practical / Clinical Significance (or importance)***

Statistical significance means that a difference is “real” and not just due to sampling variability or chance. That is, the difference would persist if the study were repeated with new random samples.

Clinical significance addresses whether the difference is “important”; i.e. is the magnitude of the difference large enough to be useful clinically so as to warrant a change in operating procedure.

An effect can be statistically significant but not clinically significant. And if an effect is not statistically significant its clinical significance cannot be assessed. That is, statistical significance is a necessary precondition for clinical significance but says nothing about the actual magnitude of the effect.

### **One last concept: the role of variability in power and sample size...**

Streiner and Norman state: “Nearly all statistical tests are based on a signal-to-noise ratio, where the signal is the important relationship and the noise is a measure of individual variation.”

If a measurement scale outcome variable has little variability it will be easier to detect change than if it has a lot of variability. So assessments or estimates of variability (i.e. standard deviation) are an important ingredient in sample size estimation.

We are now ready to tackle sample size estimation for comparative studies.

For Comparative Studies...A Second Set of Questions Before “THE QUESTION”

This time there are FOUR questions that need to be answered before sample size can be calculated. The first two of these additional are easy; the last two are hard.



Question 1: **What is an acceptable significance level (alpha)?** Convention chooses .05 (or, if you like percentages more than proportions, 5%), but some situations dictate a different choice of alpha. Alpha is also known by another name; it is the probability of making a Type I error (discussed earlier in Review of Hypothesis Testing).

Why 5%? Sir Ronald Fisher suggested this as an appropriate threshold level. However, he meant that if the p-value from an initial experiment were less than .05 then the REAL research should begin. This has been corrupted to such an extent that at the first sign of a p-value under .05 the researchers race to publish the result!

Question 2: **How large a power (i.e. probability of detection).** Convention chooses power of .80 or 80%. Note that this assumes that the risk of a Type II error can be four times as great as the risk of a Type I error.

Why 80%? According to Streiner and Norman, this was because “Jacob Cohen [who wrote the landmark textbook on Statistical Power Analysis] surveyed the literature and found that the average power was barely 50%. His hope was that, eventually, both  $\alpha$  and  $\beta$  would be .05 for all studies, so he took  $\beta = .20$  as a compromise and thought that over the years, people would adopt more stringent levels. It never happened.”

Question 3: **How large will be the variability in estimating the effect or difference of interest?** For measurement outcome variables this means estimating the population standard deviation.

Question 4: **What is the smallest effect or non-null difference that the researcher wants to detect?** That is, what is the magnitude of the clinical difference of interest? Low magnification on a microscope may fail to detect something. Too high a magnification may make unimportant details look large. Finding a needle in a haystack is difficult, but finding an elephant in a haystack is comparatively easy!

The answers to Questions 3 and 4 are the key ingredients of the formulas for sample size estimation.

## **Sample Size Estimation: Answering the QUESTION?**

The basic formulas require the four components discussed above:

- an acceptable alpha
- an acceptable power (or beta)
- the population standard deviation of the outcome variables, and
- the magnitude of the clinical difference of interest.

We will consider the following situations:

- For Measurement Outcome Variables
  - Comparison of two independent means
  - Comparison of two dependent means
- For Categorical Outcome variables
  - Comparison of two proportions
- For Complicated Designs
  - Correlation and Regression
  - Multiple Regression
  - Logistic Regression
  - Factor Analysis
  - Rank Tests

### **Comparison of two means (independent):**

We begin with a basic formula for sample size. Start with two groups, a continuous measurement endpoint, a two-sided alternative, normal distributions with the same variances and equal sample sizes. The basic formula is

$$N = 16 / \Delta^2, \text{ where } \Delta = [\mu_0 - \mu_1] / \sigma = \delta / \sigma$$

Note: This is the sample size for EACH group.

$\Delta$  can be thought of as the standardized difference between means, measured in units of the standard deviation. The magnitude of clinical difference of interest and the standard deviation are combined into a single quantity. And this quantity has a famous name – it is known as the Effect Size

(ES). As a guideline, Jacob Cohen classified effect sizes as small, moderate, and large (0.2, 0.5, and 0.8 for two-group comparisons); you can use these as a starting point.

In the one-sample case, the numerator is 8, instead of 16; that is,  $N = 8 / \Delta^2$ . This situation occurs when a single sample is being compared with an external population value (i.e. a target). Note that the sample size for a one-sample case is one-half the sample size for each sample in a two-sample case. But since there are two samples, the total in the two-sample case will therefore be four times that of the one-sample case.

Example: If the standardized treatment difference  $\Delta$  is expected to be 0.5, then  $16/(0.5)^2 = 64$  subjects per treatment will be needed. Hence a total of 128 subjects are required. If the study only requires one group then 32 subjects will be needed; this is one-fourth of the number in the two-sample scenario.

This illustrates the rule that the two-sample scenario requires four times as many observations as the one-sample scenario. The reason is that in the two-sample situations two means have to be estimated, doubling the variance, and, additionally, requires two groups.

Note that the two key ingredients are the difference to be detected ( $\mu_0 - \mu_1$ ) and the inherent variability of the observations indicated by  $\sigma$ .

Note also that the equation can be inverted to allow you to calculate the detectable difference for a given sample size  $N$ .

$$\Delta = 4 / \sqrt{N} \text{ or } (\mu_0 - \mu_1) = 4 \sigma / \sqrt{N}$$

For a one-sample case, replace 4 by 2.

This rule is very robust and useful. Many sample size questions can be formulated so that this rule can be applied.

“Where does the multiplier of “16” come from?” I hear you asking.

The full formula is the following:

$$N = 2 (z_{\alpha} + z_{\beta})^2 / (\delta/\sigma)^2$$

For  $\alpha = .05$ ,  $z_{\alpha} = 1.96$ ; for  $\beta = .20$ ,  $z_{\beta} = 0.84$ . Hence  $2 (z_{\alpha} + z_{\beta})^2 = 2(1.96 + 0.84)^2 = 15.68 \approx 16$

What if you want other values of  $\alpha$  and  $\beta$ ? Here is a small table of the multipliers for various values of  $\beta$  for a two-sided  $\alpha$  of .05.

Power (1 - $\beta$ )	Multiplier	
	One sample	Two sample
0.50	4	8
0.80	8	16
0.90	11	21
0.95	13	26
0.975	16	31

What happens if the sample size is more than you can manage in a year? Double the treatment effect. If the sample size is too small, and you can't justify enough funding, halve the treatment effect. Usually, though it's the smallest difference that you would say is clinically important. Even if a small difference were statistically significant, you wouldn't change your practice because of it.

Don't fool around with the alpha level. But you can pick out beta levels of .05, .10 or .20 (or even .50) if you are really desperate. The usual choices are alpha of .05 and beta of .20.

### Comparison of two means (dependent):

If your design has paired observations (e.g. before-and-after), then what seems to be a two-group test is really just a one-group test of whether the average change is different from zero. (See paired t-test in the Analysis II Module). Another description of this situation is that each subject is serving as his/her own control. In this case the sample size formula is:

$$N = (z_{\alpha} + z_{\beta})^2 / (\delta/\sigma)^2$$

It looks very similar to the two-sample situation, but with two important changes. First, there is no multiplier of “2”. Second, the  $\sigma$  is the standard deviation of the differences within pairs, not the standard deviation of the original measurements. This is almost never known in advance, but as Streiner and Norman say, “On the brighter side, this leaves more room for optimistic forecasts.”

### Comparison of two proportions:

Next, we estimate the sample size required to compare two proportions.

To compare two proportions  $p_0$  and  $p_1$ , use the formula:

$$N = 16 p(1 - p) / (p_0 - p_1)^2 \text{ where } p = (p_0 + p_1)/2.$$

For example, if  $p_0 = .30$  and  $p_1 = .10$ , then  $p = .20$ , so the required sample size per group is 64.

As with the comparison of two means, the multiplier of 16 can be changed. The same values as in the previous table apply here too.

Notes:

An upper limit on the required sample size occurs when  $p = .50$ ; then the formula becomes:  $N = 4 / (p_0 - p_1)^2$ . This is very conservative and works best when the proportions are centered around .50.

When the proportions are less than .05, use:  $N = 4 / (\sqrt{p_0} - \sqrt{p_1})^2$

### More complicated designs:

For situations comparing more than two groups (where analysis of variance will be the technique of analysis), the simplest (and justifiable) approach is to focus on the two groups you are most interested in being able to compare and then use the two-sample formulas above.

For repeated measures designs with more than two measurements per subject, once again, focus on the two measurements of most interest (e.g. baseline and final follow-up) and use the formula for paired measurements.

For categorical data with more than two rows and/or more than two columns, analysis is usually based on chi-squared tests. Sample size procedures are not worked out for these situations. Instead, find the two key groups to be compared, collapse the outcome to binary, and use the procedures for comparing two proportions. Note also that from theoretical considerations, a chi-square test of independence requires at least five observations per cell to give fully valid results, so plan accordingly. That is, find the key crosstabulation, determine the number of rows and columns (i.e. the number of levels of the two variables), compute the number of cells and then multiply by 5; that should be the minimum sample size for this situation.

**For correlation and regression:**

Rarely is a study aiming only to establish a single pairwise correlation between two variables. So the formula for sample size to achieve a significant correlation is not very useful. As well, a test of whether correlation is significant is really only testing whether the correlation is zero or not; it doesn't say anything about the "strength" of the correlation. Correlation is usually just the first step in model-building, especially regression models.

**For multiple regression:**

It is impossible to estimate regression coefficients before doing the research and data collection study so power studies aren't really relevant here. Instead, just ensure that the number of data points (i.e. observations or cases) is considerably more than – 5 to 10 times – the number of variables. (A reference for this rule of thumb is: Kleinbaum, Kupper and Muller (1988).)

**For logistic regression:**

The same rule of thumb applies but I would suggest aiming for a sample size of 10 times the number of variables (rather than 5), because the outcome variable is binary rather than continuous.

**For factor analysis (a useful in tool for instrument development):**

There are no power tables available to say how many subjects to use. Once again we go with a rule of thumb based on conventional wisdom and simulations that says we should have an absolute minimum of five subjects per variables with the added constraint that we have at least 100 subjects in total. This rule will give useful results only in the best of circumstances. To be on the safe side, double everything! Unfortunately, a huge percentage of factor analyses performed each year are in flagrant violation of these!

**For rank tests (e.g. Mann-Whitney, Wilcoxon, Kruskal-Wallis):**

There are also no formulas for sample size calculations. What to do? Streiner and Norman advise, “Determine the sample size from the equivalent parametric test and leave it at that. Recent work has shown that you don’t lose any power with these tests and when the data aren’t normal, they may even be more powerful than their parametric equivalent.” They provide two references for this conclusion: Blair and Higgins (1985), or Hunter and May (1993).

**Other Considerations:**

When unequal samples sizes matter and when they don’t:

In some cases it may be useful to have unequal sample sizes. For example in epidemiological studies it may not be possible to get more cases but more controls are available. Suppose  $N$  subjects are required per group, but only  $N_0$  are available for one of the groups (where  $N_0 < N$ ).

To get the same precision as with  $N$  in each group, take  $kN_0$  subjects in the second group, where  $k = N / (2 N_0 - N)$ .

For example, suppose sample size calculations show that  $N = 16$  cases and controls are needed, but only 12 cases are available. Then  $k = 16 / (2 \times 12 - 16) = 2$  and  $kN_0 = 2 \times 12 = 24$ . That is, use 24 controls to go with the 12 cases.

Rearranging this formula gives  $N_0 = [(k + 1)/2k] \times N$ . With two cases per control,  $k=2$ , then

$$N_0 = 0.75N. \text{ That is only 75\% as many cases are needed.}$$

Dropouts: Sample size estimation determines the number of complete cases which are needed for analysis. But some subjects who enroll in the study may drop out, others may be protocol failures and still others may have incomplete data, especially on the key outcome variables. To deal with this, decide on an “attrition rate” and inflate the sample size by this factor. For example, if you expect to lose about 20% of the sample, then the sample size should be increased by a factor of  $1 / (1 - 0.2)$  or 1.25. That is, enroll 25% more subjects than the sample size calculation called for. This is a reasonable “attrition rate” for many studies – but might it need to be set as high as 33% if elderly or very ill patients are the subjects.

Clustered Samples: Sometimes subjects are sampled by groups. For example, consider a design where 20 physician practices are randomly assigned to the treatment group and 20 to a control group. Then 500 charts are reviewed for each practice. Is the sample size 20 (i.e. 20 practices per group), or 1000 (i.e. the number of charts)? The answer is ... “It depends”! It depends how similar the patients are within a practice and on what the unit of analysis is. Are you interested in comparing practices or patients? For example, each practice might be given a score which is the percentage of patients in the practice who had the characteristic being looked for in the chart review.

Equivalence Studies: To show that two groups are “not different”, that is, they are equivalent, requires setting power higher (say .90 or .95) and the effect size smaller, small enough so as not to be clinically significant. With greater power and smaller effect sizes, equivalence studies therefore require larger sample sizes!

### ***Strategies for Minimizing Sample Size and Maximizing Power***

When the estimated sample size is greater than the number of subjects that can be studied realistically, what can be done? First, check your calculations!

Second, review the ingredients. Is the detectable difference or effect size unreasonably small or the variability unreasonably large? Could alpha or beta or both be increased without harm? Is the confidence level too high or the interval unnecessarily narrow?

If these fail, here are other strategies:

- Use continuous variables instead of dichotomous variables, if this is an option. There is more information in a continuous variable and so you get greater power for a given sample size or a smaller sample size for a given power.
- Use paired measurements – this reduces the between-subject part of the variability of the outcome variable.
- Use more precise variables, perhaps by taking duplicate measurements or refining the measurement tool.
- Use unequal group sizes, if it is easier to recruit in one group than another (e.g. case-control).



Use a more common outcome, that is, one with a frequency closer to 50% than to 0% or 100%.

## **Summary**

If you have jumped ahead to this point in the module, I would encourage you to return to where you left off and go through the material in detail. I have worked hard to present all the various considerations and situations and provide sound advice in how to apply the rules.

But, if you have limited time, a limited attention span, or a limited capacity for quantitative thinking (heaven forbid!), here is an “executive summary”. But, promise me that you WILL estimate the sample size early in the design phase.

First, don't be awed by the formulas and the apparent precision of the numbers arising from the sample size calculations. All the ingredients are really uncertain and crudely estimated. And the choice of 5% significance level and 80% power are somewhat arbitrary definitions for the vague concepts of “small” or “large”.

For the difference between two means, use Lehr's (1992) “Sixteen s-squared over d-squared” rule, a rule which should never be forgotten. In this phrase, “s” is the common standard deviation of the two groups and “d” is the difference between the two means. The rule sounds friendlier since it replaces the Greek letters  $\sigma$  (sigma) and  $\delta$  (delta) by s and d.

Note that if you double the difference (d) you want to detect, the sample size is cut by a factor of four. If you double the standard deviation (s), the sample size goes up by a factor of four. Streiner and Norman observe that plausibly small adjustments in the initial estimate can have big effects on the calculation! Which is why statisticians are so successful at making the calculated sample size exactly equal the number of available patients.

For difference among many means, pick the two means you really care about and then apply Lehr's rule to get the sample size for each group.

For the difference between proportions use  $N = 16 p(1 - p) / (p_0 - p_1)^2$  where  $p = (p_0 + p_1)/2$ .

Hulley, Chapter 6, has some useful sample size estimation tables.

Note: Sample size calculations should be based on the way the data will be analyzed. But, even if more complex methods of analysis will be used ultimately, it is easier and usually sufficient to estimate the sample size assuming a simpler method of analysis, such as the t-test or two means or chi-square test of two proportions.

Warnings:

- Remember to plan for dropouts and for subjects with missing data.
- Make sure you know whether the formulas you used are for equal sample sizes.
- For paired measurement scale (i.e. continuous) data, use the standard deviation of the change scores, not of the variable itself.
- Be aware of clustered data and the unit of analysis.

Finally, remember that approximations for various ingredients in the sample size formulas that are based on educated guesses by the investigator will probably work fine. The process of thinking through the problem and imagining the findings that will result is what sample size planning is all about. Carry on!

Sample Size Calculators on the Web - In increasing order of complexity:

- Rollin Brant's simple but effective sample size calculators  
<http://newton.stat.ubc.ca/~rollin/stats/ssize/>
- Deals with epidemiological applications.  
[http://hedwig.mgh.harvard.edu/sample\\_size/size.html](http://hedwig.mgh.harvard.edu/sample_size/size.html)
- Nicely designed java-applets  
<http://www.stat.uiowa.edu/~rlenth/Power/index.html>
- An excellent index page for many other on-line statistical tools.  
<http://www.utexas.edu/its/rc/world/stat/on-line.html>
- A sample size calculator for chi-square tests  
<http://stattrek.com/Lesson6/SampleSize.aspx>
- Survey-oriented websites  
<http://www.surveysystem.com/sscalc.htm>  
<http://www.researchinfo.com/docs/calculators/index.cfm>

A monstrously large site with every conceivable calculator listed. A “fun” site to browse.

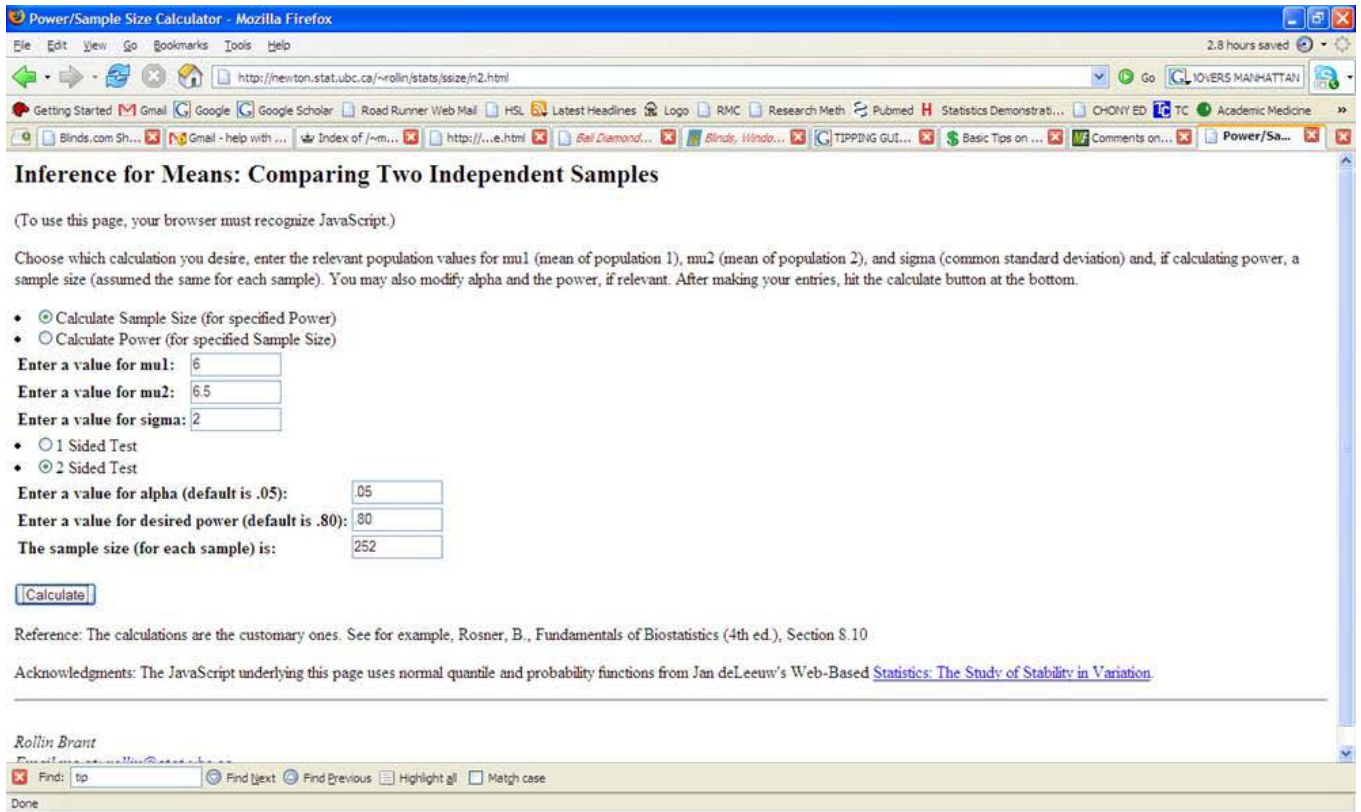
<http://www.martindalecenter.com/Calculators.html>

### ***Examples***

At Kyle and Effie’s school (Belltown Elementary), the school nurse has noticed that the students appear to be on the lower side of population growth charts, especially for height. Anecdotal evidence suggests that the children are not consuming enough milk at home; there doesn’t seem to be enough room in the refrigerators after all the beer cans have been stored. She is considering instituting a milk supplement program to improve growth rates. Her study will be a two-group randomized clinical trial, where half the children (chosen at random) will be given extra milk every day for a year. At this age, children’s height gain in 12 months has a mean of about 6 cm with a standard deviation of 2 cm. An extra increase in height of 0.5 cm in the milk group would be considered an important difference. How large a study should she do?

To have 80% power of detecting this difference (at 5% significance level, two-tailed) she would need, using Lehr’s rule,  $N = 16 \times (2 / 0.5)^2 = 256$  per group. Thus approximately 500 children would be needed for the study.

Calculator example to do this same thing:



In addition to the in-school milk supplement program, the nurse would like to increase the use of daily vitamin supplements for the children by visiting homes and educating about the merits of vitamins. She believes that currently about 50% of families with school-age children give the children a daily megavitamin. She would like to increase this to 70%. She plans a two-group study, where one group serves as a control and the other group receives her visits. How many families should she plan to visit?

To have 80% power of detecting this difference (at 5% significance level, two-tailed) she would need  $N = 16 (0.60)(1 - 0.60) / (0.50 - 0.70)^2 = 96$  per group. That is, she should plan to visit 100 families and have another 100 families in the control group.

Calculator example to do this same thing:

**Inference for Proportions: Comparing Two Independent Samples**

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values (as decimal fractions) for  $p_1$  (proportion in population 1) and  $p_2$  (proportion in population 2) and, if calculating power, a sample size (assumed the same for each sample). You may also modify alpha and the power, if relevant. After making your entries, hit the calculate button at the bottom.

- Calculate Sample Size (for specified Power)
- Calculate Power (for specified Sample Size)

Enter a value for  $p_1$ :

Enter a value for  $p_2$ :

- 1 Sided Test
- 2 Sided Test

Enter a value for alpha (default is .05):

Enter a value for desired power (default is .80):

The sample size (for each sample) is:

Reference: The calculations are the customary ones based on the normal approximation to the binomial distribution. See for example, Rosner, B., *Fundamentals of Biostatistics* (4th ed.), Section 10.7

Acknowledgments: The JavaScript underlying this page uses normal quantile and probability functions from Jan deLeeuw's Web-based [Statistics: The Study of Stability in Variation](#).

*Rollin Brant*  
 Email me at: [rollin@stat.ubc.ca](mailto:rollin@stat.ubc.ca)

Find:  Find Next Find Previous Highlight all Match case

Done

These screen captures are from:

<http://newton.stat.ubc.ca/~rollin/stats/ssize/>

Rollin Brant's simple but effective sample size calculators

## Assignment

1. In this module you will be asked to calculate the sample size for 6 situations. E-mail your answers to these questions to the Course Director.
2. For your research question, write up the Sample Size section of the methods. Be explicit in terms of your assumptions and how you determined your number. Include a paragraph where you determine the effect of doubling the sample size or halving it. How would you compensate for these changes? Would your results still be valid?