

SAMPLE SIZE CALCULATIONS

Author: Rahul M. Dodhia

Posted: April 28, 2007

Modified: July 31, 2007

For the latest version of this article, please go to www.RavenAnalytics.com/articles.php

WHY PERFORM A SAMPLE SIZE ANALYSIS?

Every analyst has at some point been surprised that his or her results are not statistically significant, or that the significant results cannot be replicated later. Much consternation later, a post-hoc sample size analysis is likely to show that they had not collected enough data. In many instances, this is a very costly design flaw that could have been avoided. Imagine designing a study in which you spend thousands of dollars, only to find out in the end you hadn't planned for enough data points. If the data collection process was logistically and financially complex, simply adding new data points is not an easy option, not to say statistically questionable.

A sample size analysis after the data has already been collected is not really of much help other than determining what went wrong with the study. A sample size analysis done before the study is planned may well save time and money.

Take as an example a manufacturer of designer handbags who is concerned about counterfeits. They want to determine within a specified margin of error, and with a certain level of confidence, how many handbags are counterfeit. They want to be able to make a statement like the following:

9% ($\pm 1.2\%$) of handbags are counterfeit, with 95% confidence.

The reasons for making such a statement are numerous: they may need to present this evidence on court when suing for damages, or they may need this information when in a pricing model of their own genuine product. The data collection procedure may involve buying handbags in 30 markets in multiple countries. After spending many tens of thousands of dollars in a data collection process, they may end up with estimates with such wide margins of error or low confidence that the estimate may be unusable.

A sample size analysis enable you to decide a priori what the correct number of data points are needed. You don't want too few, and although too many doesn't hurt from a statistical point of view, you want to be careful with your research dollars. For example, in a clinical trial comparing treatment and control groups, a large difference between the two groups would still not be evidence if each group had only 5 subjects. At the other extreme, it would be great evidence if both groups had 10,000 subjects. But if each subject costs about \$8,000 to run, even a pharmaceutical company would begin to consider these costs excessive. The trick is to find the proper balance between too few and too many, and we call this trick sample size analysis.

WHEN IS A SAMPLE SIZE ANALYSIS NECESSARY?

If we want to make any statements of these types, we need to know how many data points we need;

- We are 95% confident that the number of people who believe in issue x is $43\% \pm 3\%$
- We are 99% confident that the average increase in price per unit will be $85\text{¢} \pm 2\text{¢}$.
- To determine whether the A group is better than the B group, we need 75 participants in each group.
- This test can detect a difference of 0.4cm between the two groups.

The benefits of performing a sample size analysis before running the tests are:

- A statistically valid procedure
- Helps set rigorous standards for subsequent analysis
- Required in order for your analysis to stand up under scrutiny, such as in a legal court, federal oversight body, and many companies that are now strong on quantitative decision-making.

Behind each of these statements is a lot of statistical number crunching and also a few assumptions. Most of the times, these assumptions have been validated by

conventional wisdom; for example, 95% confidence is often the benchmark for statistical significance in most academic disciplines and industrial practice. The next section lays out which inputs are needed to generate a number for the required sample size.

DETERMINANTS OF SAMPLE SIZE

How does one determine how many patients, subjects or data points are needed? We need certain inputs which will be crunched through an equation, and the output will be the sample size. We'll illustrate this by returning to the pacemaker example.

Is it important to understand the equation by which the sample size is calculated? It certainly helps to know what is happening to your inputs, but more important is to know what the inputs mean, and how they affect the sample size. A lot of online sample size calculators will crunch the numbers for you, keeping the equation hidden, but it's important not to plug in numbers blindly, and understanding the inputs will ensure you're not using the calculators incorrectly.

There isn't a single sample size equation. They vary depending on the type of data you will be using and the type of statistical test you plan on performing.

To determine what type of sample size analysis to do, first ask yourself what kind of statement you want to make with the results of your study. The answer you are trying to get to will require different sample sizes.

EXAMPLE 1

A biotech company wants to be able to make the following statement:

The number of epileptic events in the treatment group is statistically significantly less than in the control group.

In other words, the treatment significantly reduced the number of epileptic events. The statement suggests a comparison between the two groups along the lines of a t-test. To determine the correct sample size for a t-test, the required inputs are

Input	Change in input value	Change in sample size
Minimum difference between the two groups that can be detected	↑	↓

Confidence level	↑	↑
Power of the test	↑	↑
Expected variance in the data	↑	↑

Minimum difference between two groups. How much of a difference between the two groups needs to be detectable? For example, if the treatment works really well, there should be a large difference between the treatment and control groups, and you will not need very many patients to see that the treatment works. If the treatment is not that great, but still better than no treatment, then the difference will be small. To make sure that this small difference is not just a fluke, you'll need more patients in both groups. This hints at a trick of statistics that is sometimes exploited shamelessly – just by increasing the number of data points, you can make almost any difference between two groups statistically significant.

Confidence level. You can think of confidence as the likelihood of the result repeating itself if someone asked you to do the study again. So a 95% confidence would mean that if you were to repeat the experiment 100 times, it would show about same difference about 95 times. This is fairly strong evidence that a difference between the two groups exist, given all the variability that exists in gathering data. So you'd like confidence to be as high as possible, but usually 95% is an accepted value.

Power. Power can be thought of as the likelihood that your statistical test will find a difference if there actually is a difference. So let's say the drug actually helped the patients decrease their symptoms – how likely is the statistical test to capture that reality? So you want your power to be as high as possible, and 80% is usually acceptable.

Expected variance. The variance in the data is the hardest one to estimate. And often all you can do is give your best guess for it. You can use similar studies to guide you, or be conservative and give a high value for the variance. Note that some calculators ask for the standard deviation, which is the square root of the variance.

After putting in these inputs, the sample size calculator will give you a number, say 84. This means that to detect a difference of at least 0.1, with 95% confidence and a power of 80% (and assuming whatever variance in the data), then you need at least 84 patients in each group. Once you actually do the study and have real data, you will see that your power and other inputs will be difference, but hopefully not by much.

EXAMPLE 2

Imagine you want to make these types of statement

9% ± 1.2% of handbags are counterfeit

43% ± 3% of respondents polled believe that North Korea is an immediate threat to global peace

This situation usually has the data follow a Bernoulli distribution, so to be able to figure out the correct number of handbags to sample, we need inputs for a binary decision test, i.e., is the handbags fake or not. A number of assumptions are also made here – a certain confidence level is assumed, usually 95%, but it could also be 90% or 99%. This is reflected in the inputs needed for this sample size analysis.

Input	Change in input value	Change in sample size
Margin of Error	↑	↓
Confidence level	↑	↑
Population size	↑	↑

Margin of error. The margin of error is how precise you want your estimate to be. The more precise you want, i.e., the more accurately you want to estimate the proportion of fakes in the population, the closer your sample size has to be to the population sample size. If your sample size equaled the population size, your margin of error would be 0.

Confidence level. See definition on page 2.

Population size. You may know the exact population size (there are 8000 unique investors in this company) or you may have a ballpark figure (about 100,000 handbags were produced last year). After a certain value, it makes very little difference what the population size is. For example, 2000 is not very difference from 20000. But 200 maybe very different from 2000.

So, to calculate sample size correctly, you need to know the type of statistical test you will be conducting and the probability distributions that the test assumes. You also need to know what the inputs for the sample size calculation mean. Otherwise, you might end up using the wrong calculator and get wrong results.

There are a lot of sample size calculators online that allow you to quickly estimate sample sizes. But a lot of these calculators are specialized and are strictly correct only for the specific test or experimental design that the author of the test meant them for. The best general online calculators are found at

<http://www.stat.uiowa.edu/~rlenth/Power/>. But you do

need to know what the inputs mean. StatsConsult provides similar calculators in Excel spreadsheets.

Rahul Dodhia is a founding principal of Raven Analytics. Send comments to Rahul@RavenAnalytics.com

For the latest version of this article, please go to www.RavenAnalytics.com/articles.php