

Chapter 8

Power Analysis of Means

Introduction

This chapter introduces power analysis and sample size calculation for tests that compare means. In many situations, the results for each treatment group are summarized as means. There are many issues that must be considered when designing experiments for comparing means. For example, are the means independent or correlated? Which test statistic to use? Will a parametric or nonparametric test be used? Are the data normally distributed? Are there more than two treatment groups? The answers to these and other questions result in a large number of situations.

Specifying the Means

Assume that μ_1 is the mean of an experimental group and μ_2 is the mean of a control (standard or reference) group. Suppose δ represents their difference. That is, $\delta = \mu_1 - \mu_2$. In most hypothesis tests, the null hypothesis (H_0) is $\delta = 0$ and the alternative hypothesis (H_1) is $\delta \neq 0$. Since H_0 assumes that $\delta = 0$, all that is really needed to compute the power is the value of δ under H_1 . So, when the input screen asks for μ_1 and μ_2 , these values should be interpreted as follows. The value of μ_1 is actually the value of both μ_1 and μ_2 under H_0 . Under H_1 , the values of μ_1 and μ_2 provide the value of δ at which the power is calculated.

The above discussion is summarized in the following chart:

Input Parameter	Under H_0	Under H_1
Mean1	μ_1, μ_2	μ_1
Mean2	ignored	μ_2

Also, it is important to understand what we mean by ‘under H_1 ’ in the above discussion. H_1 defines a range of values for δ at which the power can be computed. To compute the power, the specific values of δ must be determined. Thus, there is not a single power value. Instead, there are an infinite number of power values possible, depending on the value of δ .

Selecting an appropriate value for μ_1 must be done very carefully. We recommend the following approach. Select a value of μ_1 that represents the change from μ_2 that you want the experiment to detect. When you calculate a sample size, it is interpreted as the sample size necessary to detect a difference of at least δ when the significance level is α and the power is $1 - \beta$.

Power Analysis of Means

It is important to realize that δ is not the value you anticipate obtaining from the experiment. Instead, it is that value of δ at which you want to compute the power. This is a very important distinction! This is why, when investigating the power after an experiment is run, we recommend that you do not simply plug in the values of μ_1 and μ_2 from that experiment. Rather, you use values that represent the size of the difference that you want to detect.

Specifying the Standard Deviation

Usually, statistical hypotheses about the means make no direct statement about the standard deviation. However, the standard deviation is a parameter in the normal distribution, so its value must be specified. For this reason, it is called a *nuisance parameter*.

Even though it is not of primary interest, an estimate of the standard deviation is necessary to perform a power analysis. Finding such an estimate is difficult not only because it is required before the data are available, but also because the physical interpretation of the standard deviation is vague. How do you estimate a quantity without data and without a clear understanding of what it is? This section will try to help.

Understanding the Standard Deviation

The standard deviation has two general interpretations. First, it is similar to the average absolute difference between each observation and the mean. Second, it is the average absolute difference between every pair of observations.

The standard deviation of a population of values is calculated using the formula

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}}$$

where N is the number of items in the population, X is the variable being measured, and μ_X is the mean of X . This formula indicates that the standard deviation is the square root of an average of the squared differences between each value and the mean. The differences are squared to remove the sign so that negative values will not cancel out positive values. After summing up these squared differences and dividing by N , the square root is taken to put the result back in the original scale. Bottom line—the standard deviation can be thought of as the average absolute difference between the data values and their mean.

Estimating the Standard Deviation

Our task is to find a rough estimate of the standard deviation to use in a power analysis. Several possible methods could be used. These include using the results of a previous study or a pilot study, using the range, using the coefficient of variation, etc. **PASS** includes a Standard Deviation Estimator procedure that will help you obtain a standard deviation estimate based on these methods. It is loaded from the Tools menu. Remember that as the standard deviation increases, the power decreases. Hence, an increase in the standard deviation will cause an increase in the sample size. To be conservative in sample size calculation, you should use a large value for the standard deviation.

Simulations

Most of the formulas used in **PASS** were derived by analytic methods. That is, based on a series of assumptions, a formula for the power and sample size is derived mathematically. This formula is then programmed and made available in **PASS**. Unfortunately, the formula is only as realistic as the assumptions upon which it is based. If the assumptions are inaccurate in a certain situation, the power calculations may also be inaccurate. An alternative to using analytic methods is to use *simulation* (or *Monte Carlo*) techniques. Because of the speed of modern computers, simulations can now be run in minutes that would have taken days or weeks only a few years ago.

In power analysis, *simulation* refers to the process of generating several thousand random samples that follow a particular distribution, calculating the test statistic from each sample, and tabulating the distribution of these test statistics so that the significance level and power of the procedure may be estimated.

The steps to a simulation study are

1. Specify how the study is carried out. This includes specifying the randomization procedure, the test statistic that is used, and the significance level that will be used.
2. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's significance level.
3. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The significance level is the proportion of simulated samples in step 2 that lead to rejection. The power is the proportion of simulated samples in step 3 that lead to rejection.

How Large Should the Simulation Be?

As the number of simulations is increased, the precision and running time of the simulation will be increased also. This section provides a method for estimating of the number simulations needed to achieve a given precision.

Each simulation iteration (or simulation) generates a binary outcome: either the null hypothesis is rejected or not. Thus, the significance level and power estimates each follow the binomial distribution. This knowledge makes it a simple matter to compute confidence intervals for the significance level and power values.

The following table gives one-half the width of a 95% confidence interval for the power when the estimated value is either 0.50 or 0.95.

Simulation Size M	Half-Width when Power = 0.50	Half-Width when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.014 when the true power is 0.95. Also, as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved. Since most sample-size studies require an accuracy of within one or two percentage points, simulation sizes from 2000 to 10000 should be ample.

You are Running Two Simulations

It is important to realize that when you run a simulation in **PASS**, you are actually running two separate simulations: one to estimate the significance level and the other to estimate the power. The significance-level simulation is defined by the input parameters labeled “|H0”. The power simulation is defined by the input parameters labeled “|H1”. Even though you have complete flexibility as to what distributions you use in each of these simulations, it usually makes sense to use the same distributions for both simulations—only changing the values of the means.

Unequal Standard Deviations

One of the subtle problems that can make the results of a simulation study misleading is to specify unequal standard deviations unknowingly when you are investigating another feature, such as the amount of skewness. It is well known that if the standard deviations differ (a situation called heteroskedasticity), the accuracy of the significance level and power is doubtful. When investigating the power of the t or F tests in non-normal situations, care must be taken to insure that the standard deviations of the groups remain about the same. Otherwise, the effects of skewness and heteroskedasticity cannot be separated.

Finding the Hypothesized Means

It is important to set the mean difference of each simulation carefully. In the case of analytic formulas, the mean difference is specified easily and directly. Usually, the mean difference is set to zero under the null hypothesis and to a non-zero value under the alternative hypothesis. You must make certain that you follow this pattern when setting up a simulation.

For most distributions, the means are set explicitly (the exception is the multinomial distribution, where this is impossible). Hence, for both the null and alternative simulations, it is relatively simple to calculate the mean difference. You must do this! We will now present two examples showing how this is done.

For the first example, consider the case of a simulation being run to compare two independent group means using the two-sample t-test. Suppose the **PASS** setup is as follows. Note that N(40 2) stands for a normal distribution with a mean of 40 and a standard deviation of 2.

<u>Distribution</u>	<u>PASS Input</u>	<u>Mean Value of Simulated Data</u>
Group 1 Distribution H0	N(40 2)	40.0
Group 2 Distribution H0	N(40 2)	40.0
Group 1 Distribution H1	N(42 2)	42.0
Group 2 Distribution H1	N(40 2)	40.0

The mean difference under H0 is $40 - 40 = 0$, which is as it should be. The mean difference under H1 is $42 - 40 = 2$. Hence, the power is being estimated for a mean difference of 2.

Next we will consider a more complicated example. Suppose the **PASS** setup is as follows. Note that N(40 2)[70];K(0)[30] specifies a mixture distribution made up of 70% from a normal distribution with a mean of 40 and a standard deviation of 2 and 30% from a constant distribution with a value of 30.

<u>Distribution</u>	<u>PASS Input</u>	<u>Mean Value of Simulated Data</u>
Group 1 Distribution H0	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$
Group 2 Distribution H0	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$
Group 1 Distribution H1	N(42 2) [70];K(0)[30]	$42(0.7) + 30(0.3) = 38.4$
Group 2 Distribution H1	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$

Power Analysis of Means

The mean difference under H_0 is $37.0 - 37.0 = 0$, which is as it should be for the null hypothesis. The mean difference under H_1 is $38.4 - 37.0 = 1.4$. Hence, the power is being estimated by simulation for a mean difference of 1.4.

You must always be aware of what the mean differences are under both the null and alternative hypotheses.

Adjusting the Significance Level

When faced with the task of designing an experiment that will have a specific significance level for a situation that does not meet the usual assumptions, there are several possibilities.

1. A statistician could be hired to find an appropriate testing procedure.
2. A nonparametric test could be run that (hopefully) corrects for the implausible assumptions.
3. The regular parametric test could be run, relying on the test's 'robustness' to correct for the implausible assumptions.
4. A simulation study could be conducted to determine an appropriate adjustment to the significance level so that the actual significance level is at the required value.

We will now present an example of how to do the simulation adjustment alluded to in item 4, above.

The two-sample t-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracies when the data contain outliers. A mixture of two normal distributions will be used to generate data with outliers. The mixture will draw 95% of the data from a normal distribution with a mean of 0 and a standard deviation of 1. The other 5% of the data will come from a normal distribution with a mean of zero and a standard deviation of 10. A simulation study using 10,000 iterations and a sample size of 100 per group produced the following results when the nominal significance level was set to 0.05.

Nominal Alpha	Actual Alpha	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Power
0.050	0.045	0.041	0.049	0.816
0.055	0.051	0.047	0.055	0.843
0.060	0.057	0.053	0.062	0.835

The actual alpha level of the t-test is 0.045, which is below the target value of 0.05. When the nominal alpha level is increased to 0.055, the actual alpha is 0.051—close to the desired level of 0.05. Hence, an adjustment could be applied as follows. Analyze the data with the two-sample t-test even though they contain outliers. However, instead of using an alpha of 0.05, use an alpha of 0.055. When this is done, the simulation shows that the actual alpha will be at the desired 0.05 level.

There is one limitation to this method: the resulting test procedure is not necessarily efficient. That is, it may be possible to derive a testing procedure that is more efficient (requires a smaller sample size to achieve the same power). For example, in this example, a test based on the trimmed mean may be more efficient in the presence of outliers. However, if you do not have the time or ability to derive an alternative test, this adjustment allows you to obtain reasonable testing procedure that achieves a desired significance level and power.