CHAPTER 3. GENERAL TOPICS IN HYPOTHESIS TESTING AND POWER ANALYSIS WHEN THE POPULATION STANDARD DEVIATION IS KNOWN: THE CASE OF TWO GROUP MEANS

This chapter is a review of population-based hypothesis testing (the z-test) for the difference in means between two groups. We assume we know the population variance (σ^2) to keep things simple. In the subsequent chapters, we consider using samples (for which we have to estimate the population variance) to test differences between two groups, which impacts the shape of the distributions that influence power.

The difference in means as a normally distributed random variable when the population standard deviation is known

The beauty of inferential statistics is that the difference between two group means (as estimated from random samples) is also normally distributed when the population standard deviation is known. Suppose two groups, group 0 (e.g., a control group) and group 1 (e.g., a treatment group). The population difference between two groups is defined as $\Delta = \mu_1 - \mu_0$. Its standard error (SE_Δ), which is the standard deviation of the distribution of differences between groups from many samples, is a function of the shared population standard deviation (σ) from which the groups are drawn, the proportion of observations in each group (P_0 and $P_1, P_0 + P_1 = 1, 1 - P_0 = P_1$), and the total number of observations (N)

$$SE_\Delta = \frac{\sigma}{\sqrt{NP_1\left(1-P_1\right)}} \tag{3.1}$$

This expression, conceptually, means that the standard deviation of the sampling distribution is based on the standard deviation of the data divided by the square-root of a function involving the sample size. The difference in group means for some variable y from a random sample is distributed normally such that

$$\bar{y}_1 - \bar{y}_0 \sim N\left(\Delta, SE_\Delta\right) \tag{3.2}$$

Hypothesis testing with the difference between two group means when the population standard deviation is known

The task with hypothesis testing is to determine the likelihood of observing the data given assumptions about the population. Qualitatively, this is a process where we center the shape of sampling distribution of the statistic on an assumed *null* hypothesis, then find the probability that our data (or more extreme) came from such a sampling distribution. If the probability of observing the data on hand (or more extreme) is low, we then usually reject the null hypothesis as unlikely. The general logic is that either our data are wrong or the null hypothesis is wrong, and we use the process of quantifying the likelihood of our data (assuming the null hypothesis is true) to make our decision.

Hypothesis testing using Neyman & Pearson's method

Hypothesis testing, as presented by the statisticians Neyman and Pearson (1933), is about making a decision between two hypotheses. The null hypothesis is generally noted as H_0, and the alternative hypothesis is noted as H_a. For example, we may state that the null hypothesis is that the mean of group 0 and group 1 are the same

$$H_0 : \mu_1 = \mu_0,$$

which implies that the difference between the two group means, or Δ, is 0

$$H_0 : \mu_1 - \mu_0 = \Delta = 0.$$

When we are testing the difference in group means, this is typically the null hypothesis.

The alternative hypothesis (H_a), takes one of three forms when looking at the means of two groups. First, we can assert an alternative hypothesis that states that the mean of group 1 is greater than the mean of group 0

$$H_a : \mu_1 > \mu_0,$$

which implies that $\Delta = \mu_1 - \mu_0$ is a positive value, or

$$H_a : \mu_1 - \mu_0 > 0.$$

or

$$H_a : \Delta > 0.$$

Second, we can assert an alternative hypothesis that states that the mean of group 1 is less than the mean of group 0

$$H_a : \mu_1 < \mu_0,$$

which implies that $\Delta = \mu_1 - \mu_0$ is a negative value, or

$$H_a : \mu_1 - \mu_0 < 0.$$

or

$$H_a : \Delta < 0.$$

These first two alternative hypotheses are so-called one-tail hypotheses, the reason for which will be clear below. A final alternative hypothesis common to the examination of two group means is that the group means are just different, without a specification of direction

$$H_a : \mu_1 \neq \mu_0,$$

which implies that the difference $\Delta = \mu_1 - \mu_0$ is non-zero

$$H_a : \mu_1 - \mu_0 \neq 0.$$

or

$$H_a : \Delta \neq 0.$$

This is a so-called two-tail test, and why it is called a two-tailed test will be explained below.

The decision with hypothesis testing is this: should the null hypothesis (H_0) be rejected in favor of the alternative hypothesis (H_a)? A general way to think about rejecting the null hypothesis is to state that the observed data are unlikely to have occurred if the null were true, since the observed Δ is so different than the value stated in the null hypothesis relative to the sampling distribution.

One way to do this is to make the assumption that Δ is the value stated in the null hypothesis (typically, the null states that $\Delta = 0$), then draw a sampling distribution that is normally distributed with a mean of $\Delta = 0$ and a standard deviation of SE_Δ. Remember, SE_Δ is based on the known population standard deviation, the sample size, and the proportion of observations in each group (3.1). To put it another way, our sample size and known standard deviation allow us to draw a normal curve centered on the value of Δ assumed by the null hypothesis.

We then consider how likely it is to have observed our Δ compared to the null hypothesis. This approach, however, is not standardized and depends the ability to plot sampling distributions in the unit of the outcome being considered. While this is possible, it is difficult to do and does not lend itself well to general rules of thumb.

Another possibility is to perform the same procedure, but instead of using Δ and its standard error to draw a sampling distribution, we use Δ and its standard error to calculate a standard test statistic and then use the standard normal distribution to consider the likelihood of the data in light of the null hypothesis. The test statistic is implied in the z-scores used in (2.2), (2.3), and (2.4), and is

$$z = \frac{\Delta - \Delta_0}{SE_\Delta} \tag{3.3}$$

where Δ_0 is the difference assumed by the null hypothesis (which is typically 0). With this test in hand, we can return to the standard normal distribution in Figure 2.1 to set up our areas of rejection that relate to Type I error.

We now return to the types of errors that were introduced in Chapter 1. The first type of error is Type I, which is the chance of rejecting the null hypothesis in when in fact it is true. We note this error as α, and typically consider $\alpha = 0.05$ as an acceptable chance of Type I error.

We use the Type I error (α) rate with the sampling distribution under the null hypothesis to create regions of rejection based on the specification of the alternative hypothesis. These rejection regions are simply areas of the null hypothesis' sampling distribution that encompass the proportion of area under the distribution curve that represent an acceptable risk of error. The value of z where the rejection region begins is the critical value, and if the test statistic (3.3) exceeds this value, then the null hypothesis is rejected.

Figure 3.1 presents these critical regions for each of the three types of alternative hypotheses that set $\alpha = 0.05$. As you can see, the first two plots only shade one of the tails with 5 percent ($\alpha = 0.05$) of the area (which is why they are called "one-tailed" tests). The shading for the first plot, the "greater-than" alternative hypothesis that $\mu_1 > \mu_0$, starts at $z = 1.64$. This means that if (3.3) is greater than 1.64 that we reject the null hypothesis in favor of the alternative hypothesis. A similar logic is used for the alternative hypothesis used for in the second plot of Figure 3.1, that $\mu_1 < \mu_0$, except here the test statistic must be less than -1.64 to reject the null hypothesis.

Finally, if the alternative hypothesis is that the means are simply different, that $\mu_1 \neq \mu_0$, we split the 5 percent of the shaded area into both tails, which is why they are "two-tailed" tests. Since we now shade the tails with only 2.5 percent, the critical values are more extreme, $z = -1.96$ and $z = 1.96$. If the test statistic (3.3) exceeds 1.96 in either direction, we reject the null hypothesis. Of course, as we change α to another value, for example $\alpha =$

0.01, the percent of the sampling distribution that becomes the region of rejection changes, and thus the critical values change. Critical values for various levels of α and alternative hypotheses are found in almost every introduction to statistics text book.

What is a p-value?

Most statistical software does not use the Neyman and Pearson method of hypothesis testing directly. Instead, the software finds the two-tailed area of the sampling distribution that exceeds the magnitude of a test statistic, assuming that the null hypothesis is $\Delta_0 = 0$. For example, if the test statistic (3.3) were computed to be $z = 3$, the software uses the equivalent of one minus (2.4) to compute a probability that the data's test result (or a result that is more extreme) would occur assuming a null hypothesis of $\Delta = 0$:

$$p(z) = 1 - (\Phi(z) - \Phi(-z))$$

$$p(z) = 1 - (\Phi(3) - \Phi(-3))$$

$$p(z) = 0.003$$

Another way to express the two-tailed p-value is

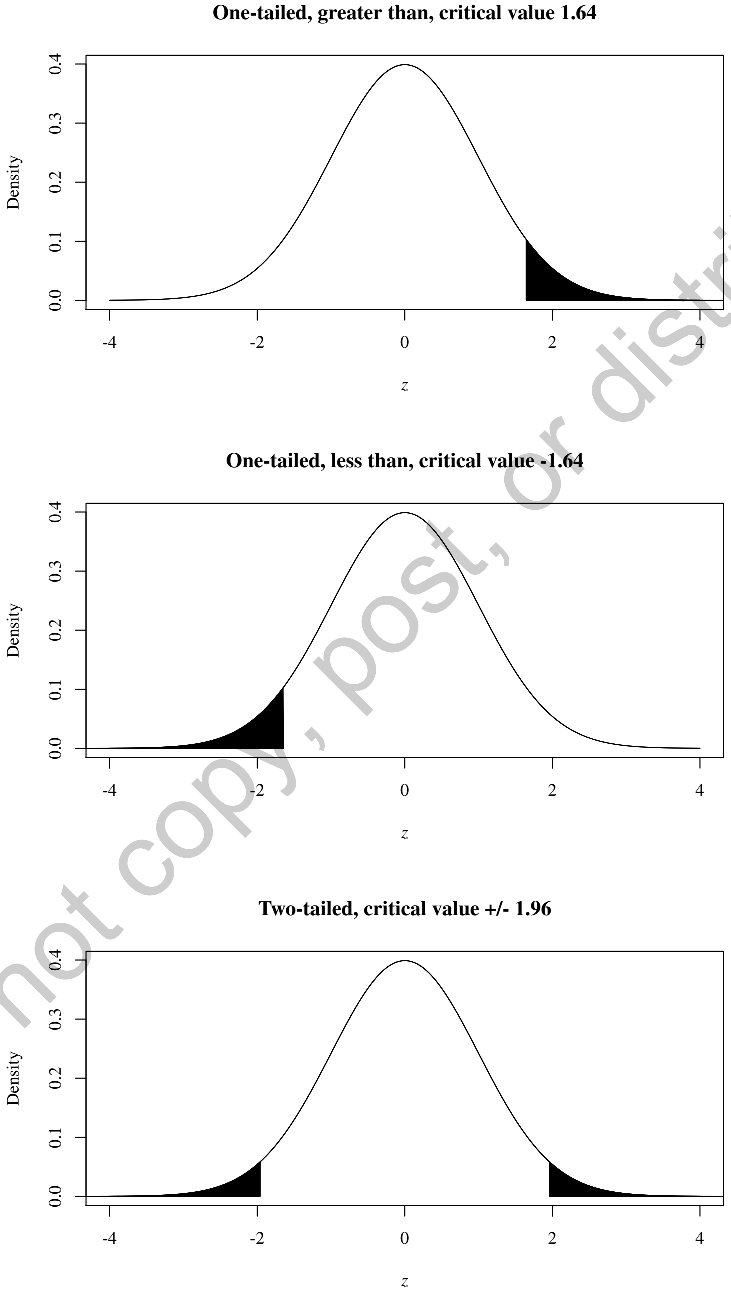
$$p(z) = 2 \times \left(1 - \Phi\left(\frac{|\Delta|}{SE_\Delta}\right)\right), \tag{3.4}$$

and for the one-tailed test it is,

$$p(z) = 1 - \Phi\left(\frac{|\Delta|}{SE_\Delta}\right) \tag{3.5}$$

The idea behind the "test" that uses the p-value is if $p(z)$ is less than α, then we consider the difference to be "significant."

It is important to realize that p-values are not the probability that the data per se occur if the null hypothesis were true. The data are a single point along a continuous distribution. Probabilities from a continuous distribution require a *range*, and the probably of a single point is 0. That is why we talk about the p-value as the chance the "the data's test result or a result that is more extreme" occurs if the null hypothesis were in fact true.



One-tailed, greater than, critical value 1.64

One-tailed, less than, critical value -1.64

Two-tailed, critical value +/- 1.96

Figure 3.1 Null z distribution with one- and two-tailed critical regions shaded ($\alpha = 0.05$)

What is useful for power analysis?

As we will see below, power analysis is essentially the manipulation of areas under curves given various ranges of values. Thus, p-values are less useful for power analyses since we never know the exact p-value a future data set will generate. Instead, we can pre-specify a level of α, which generates critical values, which allow us to find the probabilities of ranges along normal distributions (and other distributions when we have to estimate the population standard deviation, σ)

Power analysis for testing the difference between two group means when the population standard deviation is known

Now that we have established how to think about hypothesis testing using the normal distribution, we are ready to think about power analysis. Power analysis is tied to Type II error (β), or the chance that we accept the null hypothesis when in fact the alternative is true. One way to think about Type II error is to consider the chance that our test statistic falls before the critical value specified by α, analogous to the second simulation in Chapter 1 (see Figure 1.2). For example, if our alternative hypothesis states that $\Delta > 0$, and our one sided test sets $\alpha = 0.05$, our critical value of z is 1.64. Before collecting the data, Type II error asks what is the chance that our test statistic (3.3) winds up less than 1.64, even though the alternative hypothesis is true?

To answer this question, we suppose some expected test result, then draw a sampling distribution around that expected test result. Noting where the critical value falls, Type II error (β) for a one-tailed test is simply the area before the right critical value (if the alternative is positive) or after the left critical value (if the alternative is negative). Type II error (β) for a two-tailed test is the area between the critical values. The power of the test is then the area outside this region, or $1 - \beta$. Thus, power analysis is about imposing a sampling distribution that assumes the alternative hypothesis is true (the *alternative* or non-central distribution) on to the distribution that assumes the null hypothesis is true (the *null* or central distribution).

In order to find areas under the curve between points of the the alternative distribution, we need to know the mean and standard deviation of the alternative distribution. Using the standard normal distribution, we can use the same standard deviation as the null curve for the alternative curve (which is 1). However, while the mean of the null distribution is 0 (because the null assumes that $\Delta = 0$), we need to find an appropriate average for the alternative distribution. The mean for the alternative distribution is the expected test statistic, or what we expect (3.3) to be if the alternative were true. We generally note this parameter to be λ and call it the non-centrality parameter. Rearranging (3.3) slightly, using Δ and (3.1), gives a general expression for the non-centrality parameter for a standard normally distributed test statistic

$$\lambda = \frac{|\Delta|}{\sigma}\sqrt{NP_1(1-P_1)}. \tag{3.6}$$

Note that in this volume we will consider the difference between means in absolute values. This allows us to organize our analyses around only positive critical values.

Once we have a value for λ, we can take advantage of the fact that a variable Z, which is distributed standard normal, minus a non-centrality value λ is normally distributed (i.e., $Z - \lambda \sim N(0,1)$). The way we use this fact is to use the CDF of the standard normal distribution (Φ) to find the area before a critical value of $z_{1-\alpha}$ given a level of α to estimate the Type II error of a one-tailed test

$$\beta = \Phi\left(z_{1-\alpha} - \lambda\right), \tag{3.7}$$

or the area between critical values to estimate the Type II error for two-tailed tests

$$\beta = \Phi\left(z_{1-\alpha/2} - \lambda\right) - \Phi\left(z_{\alpha/2} - \lambda\right). \tag{3.8}$$

With Type II error (β) in hand, power is then simply $1 - \beta$.

Example

Suppose that a difference in means between two groups for some standardized test is expected to be $\Delta = 25$ from a sample of 100 observations where the population standard deviation is $\sigma = 75$. Further, suppose a balanced design where 50 observations are in each group, so $P_1 \frac{50}{100} = 0.5$. Using (3.6), we can calculate λ as

$$\lambda = \frac{25}{75}\sqrt{100 \times 0.5 \times (1-0.5)} = 1.67$$

If we were to conduct a one-tailed test with $\alpha = 0.01$, then the critical value of standard normal is 2.326, or that the $1 - \alpha = 1 - .01 = 0.99$ quantile value of the z distribution is $z_{1-\alpha} = z_{0.99} = 2.326$. We then find the Type II error by finding the area of the standard normal distribution before $z_{1-\alpha} - \lambda = 2.326 - 1.667 = 0.659$, or $\beta = \Phi(.659) = 0.745$. This means that power is about $1 - \beta = 1 - 0.745 = 0.255$. The one-tailed example is featured in Figure 3.2. As you can see, the gray shaded area is the $\beta \times 100 = 75$ percent of the alternative distribution, leaving the area beyond the right critical value (where the black shading starts) as the power of the test.

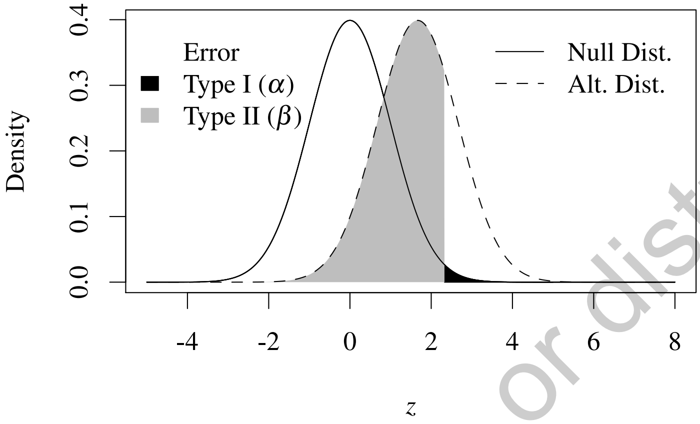


Figure 3.2 The standard normally distributed null distribution (solid line) with one-tail shaded corresponding to $\alpha = 0.01$ where the critical value $z_{\alpha/2} = 2.326$, and the alternative distribution (dashed line) with a non-centrality parameter $\lambda = 1.667$ and $\beta = 0.745$

If we were to use the same value of λ and conduct a two-tailed test with $\alpha = 0.01$, then the critical value of z is 2.576, or that the $1 - \alpha/2 = 1 - .01/2 = 0.995$ quantile value of the z distribution is $z_{1-\alpha/2} = z_{0.995} = 2.576$. We find the Type II error by finding the area of normal distribution between $z_{1-\alpha/2} - \lambda$ and $z_{\alpha/2} - \lambda$, which is $\beta = \Phi(2.576 - 1.667) - \Phi(-2.576 - 1.667) = 0.818$. This means that power is about $1 - \beta = 1 - 0.818 = 0.182$. The two-tailed example is featured in Figure 3.3. As you can see, the gray shaded area is the $\beta \times 100 = 82$ percent of the alternative distribution, leaving the area beyond the left and right critical values (where the black shading starts) as the power of the test.

Useful relationships

Given a standard normal distribution, we can use (3.7) and (3.8) to calculate useful quantities. For example, consider the Type II error for a one-tailed

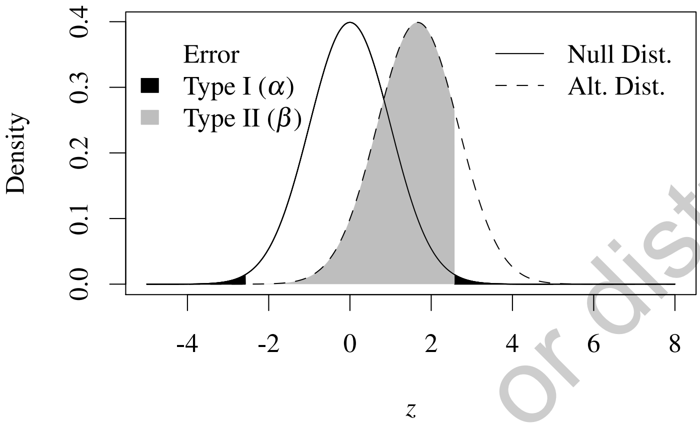


Figure 3.3 The standard normally distributed null distribution (solid line) with two-tails shaded corresponding to $\alpha = 0.01$ where the critical value $z_{\alpha/2} = \pm 2.576$, and the alternative distribution (dashed line) with a non-centrality parameter $\lambda = 1.667$ and $\beta = 0.818$

test.

$$\beta = \Phi\left(z_{1-\alpha} - \lambda\right).$$

Here, β is an area of the normal curve. Thus, we can reverse the Φ on both sides of the equation to reveal that

$$z_{\beta} = z_{1-\alpha} - \lambda,$$

which is reorganized as

$$\lambda = z_{1-\alpha} - z_{\beta}.$$

This means that in the context of the standard normal test, the expected test statistic is a function of two quantiles (at points $1-\alpha$ and β) of the standard normal distribution. The two-tailed test is a little more difficult, because we are adding two CDFs to calculate β. However, the second term, $\Phi\left(z_{\alpha/2} - \lambda\right)$ is usually very small (it is the area of the test distribution before the left critical value). In the example, the value was 0.00001. If we ignore that portion, we can use the same method for the one tailed test to find

$$\lambda \approx z_{1-\alpha/2} - z_{\beta}$$

for the two tailed test. Thus, we can generalize this relationship with the notation of the positive value of $z_{critical}$ based on whether the test is one or two tailed:

$$\lambda \approx z_{critical} - z_{\beta} \tag{3.9}$$

If $\alpha = 0.05$, the value of $z_{critical} = 1.96$ for two-tailed tests and $z_{critical} = 1.68$ for one-tail tests. For power of 0.8, $z_{\beta} = z_{0.2} = -0.84$, for power of 0.9, $z_{\beta} = z_{0.1} = -1.28$. For example, the expected test statistic (λ) for an α two-tailed test with power of 0.8 is approximately $1.96 - (-0.84) = 2.8$, and that number is often used in different types of power analyses as shown below.

Distribution quantiles

The values of $z_{critical}$ and z_{β} are quantiles of the standard normal distribution. A quantile is the inverse of a CDF. It is the value of z given a cumulative probability. Finding the quantiles of the normal distribution is possible with a computer program. For ease of using this text, I present important quantiles of the standard normal distribution on the last row of Table 4 in the Appendix.

The first set of columns in Appendix Table 4 are useful in finding quantiles of the normal distribution associated with values of β. For example, for a Type II error of 0.1, use $z_{\beta} = z_{0.1} = -1.282$. The columns to the right are useful for finding the positive critical values associated with the type of test. For example, if $\alpha = 0.05$ for a one-tail test, use $z_{1-\alpha} = z_{0.95} = 1.645$. For a two-tailed test, use $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Scale-free parameters

Up until this point, almost all of the components of our power analysis have been in the units of the outcome. The values of μ_1, μ_0, Δ, and σ have all been in the units of the variable in question, whether it be a standardized test, body mass index, or Galaga scores. The actual difference in values in the means between groups is itself difficult to anticipate, and in samples the variance is also difficult to anticipate. Power analysis becomes much easier if we deal with so-called “scale-free” parameters. Our first scale free parameter to be introduced in this volume is the effect size. Cohen’s d is an effect size that standardizes the difference between group means, Δ, by dividing it by the population standard deviation, σ (Cohen, 1988). This metric can be expressed with the lower-case greek letter δ, and is noted as

$$\delta = \frac{\mu_1 - \mu_0}{\sigma} = \frac{\Delta}{\sigma}, \tag{3.10}$$

which is the first term in (3.6).

The meaning of δ is relatively clear: it is the difference in means in standard deviation units. Thus, $\delta = 0.5$ indicates that one group is half a standard deviation greater, on average, than another group. The purpose of using δ instead of $\frac{\Delta}{\sigma}$ in planning studies is twofold. First, it allows researchers to think about the results of their future study in more general terms; what is the general difference between groups in standard units?

Second, and probably more important, is that effect sizes allow researchers to use prior research to manage expectations (see Chapter 8), even though they use different metrics. For example, suppose researchers are planning a study on how child savings accounts for college influence the financial literacy of parents. A previous study may use scale A, while the current researchers may plan to use scale B. If scale A ranges from 1 to 100, it is hardly possible to translate that difference into scale B that ranges from 0 to 7. However, if the previous research published the means and standard deviations for treatment and control groups, it is possible to determine an effect size (more on this later). With an effect size in hand, researchers using scale B will have some insight about the expected effect in their study.

Thus, we can express the non-centrality parameter using the effect size as follows:

$$\lambda = \underbrace{\delta}_{\text{Effect size}} \underbrace{\sqrt{NP_1 (1 - P_1)}}_{\text{Sample size}} \tag{3.11}$$

which allows for a power analysis that has no information about the scaled parameters in the future study.

Balance or unbalanced?

Another issue often encountered in planning studies is *balance*. A balanced study is one in which each group has the same number of cases, whereas unbalanced studies have differing numbers of cases in each group. Balance was once a crucial consideration in studies, especially before modern computing, because it made the calculations easier. Today, it is rarely the case that the final data researchers work with will actually be balanced; but it may be close.

However, in planning studies balance is an important consideration for several reasons. First, holding total sample size constant, a balanced sample has more power than an unbalanced sample. Second, as you will see, the non-centrality parameter and the calculations that employ it are easier to handle with a balanced sample. Third, sometimes balance may not be palatable: we may want to have a smaller control group to maximize the number of treatments applied to the study population.

Non-centrality parameter without balance

We have already seen the non-centrality parameter for the unbalanced case (3.11). In this case, group 1 has a sample size of n_1 and group 0 has a sample size of n_0, the total sample size is $n_1 + n_0 = N$, and the proportion of observations in group 1 is $\frac{n_1}{N} = P_1$. Thus, the calculations in the unbalanced case need information about the effect size, δ, the *total sample*, N, and the proportion of observations in one of the groups (I prefer group 1 P_1, but it makes no difference).

Non-centrality parameter with balance

The non-centrality parameter simplifies somewhat when $n_1 = n_0 = n$. In this case, $P_1 = (1 - P_1) = 0.5$ and the second term in the non-centrality parameter (3.11) is

$$\sqrt{N \times 0.5^2} = \sqrt{N \times 0.25} = \sqrt{\frac{N}{4}} = \sqrt{\frac{n}{2}}.$$

This makes the non-centrality parameter in balanced case

$$\lambda = \underbrace{\delta}_{\text{Effect size}} \underbrace{\sqrt{\frac{n}{2}}}_{\text{Sample size}} . \tag{3.12}$$

Thus, the calculations in the balanced case need only information about the effect size, δ and the *sample size in each group*, n.

Types of power analyses

Three types of questions are usually posed when planning the sample design of a study:

- **Finding power a priori:** What will be the power for a given α, effect size, and sample size?

- **Finding the necessary sample size:** What will the sample size need to be for a given level of power $(1-\beta)$, α, and effect size?
- **Finding the minimum detectable effect size:** What is the minimum detectable effect size for a given level of power $(1-\beta)$, α, and sample size?

Finding power a priori

Much of this volume so far as been about finding power given an effect size, sample size, and level of α. To briefly recap, once we have an expected effect size δ as defined in (3.10), the total sample size N, proportion of cases in one group P, and α, we can find the Type II error (β) for a one tail test using the CDF of the standard normal distribution:

$$\beta = \Phi(z_{1-\alpha} - \lambda) = \Phi\left(z_{1-\alpha} - \delta\sqrt{NP_1(1-P_1)}\right)$$

and a similar process for a two-tailed test

$$\beta = \Phi\left(z_{1-\alpha/2} - \lambda\right) - \Phi\left(z_{\alpha/2} - \lambda\right)$$
$$\beta = \Phi\left(z_{1-\alpha/2} - \delta\sqrt{NP_1(1-P_1)}\right) - \Phi\left(z_{\alpha/2} - \delta\sqrt{NP_1(1-P_1)}\right)$$

where z_a is the quantile of the standard normal distribution at point a.

When the data are expected to be balanced, these formulas simplify to the following for a one-tailed test

$$\beta = \Phi(z_{1-\alpha} - \lambda) = \Phi\left(z_{1-\alpha} - \delta\sqrt{\frac{n}{2}}\right)$$

and for a two-tailed test

$$\beta = \Phi\left(z_{1-\alpha/2} - \lambda\right) - \Phi\left(z_{\alpha/2} - \lambda\right)$$
$$\beta = \Phi\left(z_{1-\alpha/2} - \delta\sqrt{\frac{n}{2}}\right) - \Phi\left(z_{\alpha/2} - \delta\sqrt{\frac{n}{2}}\right)$$

These calculations require either a computer with the standard normal CDF function or a table of values. In any case, once the value of β is determined, the power of the test is $1-\beta$.

Finding the necessary sample size

To find the necessary sample size, we return to the useful relationship presented above, namely that $\lambda \approx z_{critical} - z_\beta$. With this relationship, and unpacking the non-centrality parameter λ, we can find the sample size necessary to detect a given effect size for a given level of α, which drives $z_{critical}$, and power, which drives z_β.

Unbalanced case

In the unbalanced case, the non-centrality parameter (3.11) for testing the difference between two groups with a known standard deviation is

$$\lambda = \delta\sqrt{NP_1\left(1-P_1\right)}$$

and so we can express all the ingredients for this calculation as

$$\delta\sqrt{NP_1\left(1-P_1\right)} \approx z_{critical} - z_{\beta}.$$

We can rearrange this to isolate N, which leaves us with the formula

$$N \approx \frac{\left(z_{critical} - z_{\beta}\right)^2}{\left(P_1\left(1-P_1\right)\right)\delta^2} \tag{3.13}$$

Looking at this expression, we can see that holding constant the power and significance level (the numerator), the sample size will increase as P_1 deviates from 50 percent, and will decrease as the effect size (δ) increases.

Example

For example, suppose we were going to perform a two-tailed test with $\alpha = 0.05$ thus $z_{critical} = z_{1-\alpha/2} = z_{0.975} = 1.96$, and wished to find the sample size for power of 0.8, so $\beta = 1 - 0.8 = 0.2$ thus $z_{\beta} = z_{0.2} = -0.842$, with an effect size of $\delta = 0.3$, where a quarter of the sample was in group 1 thus $P_1 = 0.25$. Putting these values into (3.13), we arrive at

$$N \approx \frac{\left(z_{critical} - z_{\beta}\right)^2}{\left(P_1\left(1-P_1\right)\right)\delta^2} \approx \frac{\left(1.96 - (-0.842)\right)^2}{\left(0.25\left(1 - .025\right)\right)0.3^2} \approx 465.26$$

or about 466 observations (we always round up).

Balanced case

If the observations are balanced, i.e. $n_1 = n_0 = n$ and $2 \times n = N$, the non-centrality parameter reduces to (3.12)

$$\lambda = \delta\sqrt{\frac{n}{2}} \approx \left(z_{critical} - z_{\beta}\right),$$

which can be arranged to find

$$n \approx \frac{2\left(z_{critical} - z_{\beta}\right)^2}{\delta^2} \tag{3.14}$$

Like the unbalanced case, we find again this expression indicating that the sample size will decrease as the effect size (δ) in the denominator increases.

Example

For example, suppose we were going to perform a two-tailed test with $\alpha = 0.05$ thus $z_{critical} = z_{1-\alpha/2} = z_{0.975} = 1.96$, and wished to find the sample size for power of 0.8, so $\beta = 1 - 0.8 = 0.2$ thus $z_\beta = z_{0.2} = -0.842$, with an effect size of $\delta = 0.3$. Putting these values into (3.14), we arrive at

$$n \approx \frac{2\left(z_{critical} - z_\beta\right)^2}{\delta^2} \approx \frac{2\left(1.96 - (-0.842)\right)^2}{0.3^2} \approx 174.47$$

or about 175 cases per group for a total of $2n = N = 350$. Notice that all the parameters in this example were the same as in the unbalanced case, and we arrived at a smaller value of N.

Finding the minimum detectable effect size

The minimum detectable effect size (Bloom, 1995) is a metric designed to summarized the sensitivity of a given sample. It is the smallest effect size (δ) that can be detected at a given level of power, assuming a sample size and level of α. We again return to the useful relationship presented above, namely that $\lambda \approx z_{critical} - z_\beta$. With this relationship, and unpacking the non-centrality parameter λ, we can find the effect size necessary to result in a significant test for a given sample size, with a given level of α, which drives $z_{critical}$, and power, which drives z_β.

Unbalanced case

Rearranging the non-centrality parameter (3.11) and $\left(z_{critical} - z_\beta\right)$ we can isolate the minimum detectable effect size

$$\delta_m \approx \frac{z_{critical} - z_\beta}{\sqrt{NP_1\left(1-P_1\right)}}. \tag{3.15}$$

This expression indicates that the detectable effect will decrease as the sample size (N) increases, but will increase as the sample becomes more unbalanced (i.e. P_1 moves away from 0.5).

Example

For example, suppose we were going to perform a two-tailed test with $\alpha = 0.05$ thus $z_{critical} = z_{1-\alpha/2} = z_{0.975} = 1.96$, and wished to find the sample size for power of 0.8, so $\beta = 1 - 0.8 = 0.2$ thus $z_\beta = z_{0.2} = -0.842$, with a sample size of $N = 500$, where a quarter of the sample was in group 1 thus $P_1 = 0.25$. Putting these values into (3.15), we arrive at

$$\delta_m \approx \frac{z_{critical} - z_\beta}{\sqrt{NP_1\left(1-P_1\right)}} \approx \frac{1.96 - (-0.842)}{\sqrt{500 \times 0.25\left(1 - 0.25\right)}} \approx 0.289.$$

This effect size is a little smaller than the previous example of 0.3 because we increased the sample from 466 to 500.

Balanced case

If the observations are balanced, i.e. $n_1 = n_0 = n$ and $2 \times n = N$, the noncentrality parameter can be rearranged to find

$$\delta_m \approx \left(z_{critical} - z_\beta\right)\sqrt{\frac{2}{n}} \tag{3.16}$$

This expression indicates that the minimum detectable effect will decrease as the sample size increases.

Example

For example, suppose we were going to perform a two-tailed test with $\alpha = 0.05$ thus $z_{critical} = 1.96$, and wished to find the sample size for power of 0.8, so $\beta = 1 - 0.8 = 0.2$ thus $z_\beta = 0.842$, with a sample size of $n = 175$ per group. Putting these values into (3.16), we arrive at

$$\delta_m \approx \left(z_{critical} - z_\beta\right)\sqrt{\frac{2}{n}} \approx (1.96 - (-0.842))\sqrt{\frac{2}{175}} \approx 0.300$$

or an effect size of 0.3, which is what associated with 175 cases per group above.

Power tables

The focus of this volume is on calculating results using formulas, which in turn will inform the use of software. Before software was widely available for power analysis, many researchers relied on books of tables such as Cohen's seminal volume (1988). These tables allowed the researcher to perform all three types of power analyses (with some effort). For example, examine Table 3.1, which is a reproduction of a similar table in Cohen's book (1988).

Power tables are typically organized around the three key pieces of information: sample size, effect size, and power. Separate tables are produced around the other assumptions, such as the number of tails and Type I error (α). With tables such as Table 3.1, users can move across rows that identify sample sizes, and columns, which identify effect sizes, to find the power in a given cell. For example, looking at table Table 3.1, we can see that for a sample size of 58 units in each group (last row) and an effect size of 0.4 (second to last column), the power of that study for a two-tailed test with $\alpha = 0.05$ is 0.57 (power tables typically do not print the decimal point).

	δ				
n	0.1	0.2	0.3	0.4	0.5
50	8	17	32	51	70
52	8	17	33	52	71
54	8	18	34	54	73
56	8	18	35	55	75
58	8	19	36	57	76

Table 3.1 Portion of Power Table 2.3.5 for two-tailed tests with $\alpha = 0.05$ from Cohen (1988, page 37). Small differences in power values are due to the approximation used in Cohen (1988) compared to the software used by the present author.

Summary

In this chapter we focused on how power analysis works from the point of view of understanding the sampling distributions under two scenarios: one where the null hypothesis the true (the central distribution) and the other where the alternative is true (the non-central or alternative distribution). If the population standard deviation is known, we can use the standard normal distribution and its quantiles to perform a power analysis. I also introduced the idea of the effect size and the need for scale-free parameters. In the next chapter, we move to the more common scenarios in which the population standard deviation is not known and must be estimated. This means that we must use the t-distribution (Student, 1908), which depends on degrees of freedom, which in turn depends on the sample size. Since sample size is such an important aspect of power, we will find that this complicates matters.