6

Statistical Power and Sample Size

We began the previous chapter by citing statistics from the What Works Clearinghouse (WWC) about the enormous number of completed empirical evaluations of educational interventions that were unable to support causal inference. For example, we noted that among 301 evaluations of the effectiveness of interventions in elementary mathematics, 97% of the studies reviewed could not support a causal conclusion. The most common reason was that the authors of the studies were unable to defend the assumption that participants who had been assigned to the treatment and control conditions were *equal in expectation* before the intervention began.

However, even in studies that meet this condition—for example, because the investigator has assigned members of the analytic sample randomly to treatment and control groups—the effort can be stymied by a sample of inadequate size. If you conduct otherwise well-designed experimental research in a too-small sample of participants, you may estimate a positive impact for your intervention, but be unable to reject the null hypothesis that its effect is zero, in the population. For example, the 3% of studies of elementary-mathematics interventions that met the WWC standards for supporting causal inferences included one evaluation of the causal impact of a curriculum entitled Progress in Mathematics 2006.[1] Had the sample size of this study been larger and all else remained the same, the modest positive results of the evaluation would have been statistically significant.

[1] http://ies.ed.gov/ncee/wwc/reports/elementary_math/promath_06/, accessed May 29, 2009.

Thus, early in the process of planning research, it makes good sense to decide how many participants you need to include in your sample in order to have a decent chance of detecting any effect that may indeed be present in the population. To make this sample size decision sensibly, you need to conduct what is known as a *statistical power analysis* as part of your research planning process. In this chapter, we explain how to do this. As you will see, an important guiding principle is that you can always manipulate the important facets of your research design, such as sample size, to create a stronger empirical "magnifying glass" for your work. With a more powerful magnifying glass, you can always see finer detail.

We devote this chapter and the next to explaining how to conduct statistical power analyses because we believe that many social-science investigators have been unaware of the true requirements for sample size in effective research design. As a result, much empirical research in education and the social sciences in the past has been underpowered. In this chapter, we describe the link between statistical power and sample size, and establish basic guidelines for figuring out the values that both should take on in high-quality research. We begin by defining the concept of *statistical power,* connecting it to the process of statistical inference with which you are already familiar. Then, we describe the link between power and sample size, and between power and other critical features of the research design. We do this all in the context of the "gold standard" research design for causal research—an experiment in which participants have been randomized individually to either a treatment or a control condition. Then, in the following chapter, we extend our presentation to include the more complex case in which groups of individuals—such as classrooms or schools—are sampled and randomly assigned to experimental conditions.

Statistical Power

Reviewing the Process of Statistical Inference

In introducing the concept of statistical power, we rely again on the example of the New York Scholarship Program (NYSP), which we introduced in Chapter 4. As we described earlier, the NYSP is an example of a two-group experiment in which individual participants were randomly assigned to either a treatment or a control group. Members of the experimental group received a private-school tuition voucher and members of the control groups did not. To facilitate our explanation of the critical statistical concepts in this chapter, we begin by narrowing our focus and addressing the implicit NYSP research question using the simplest appropriate analytic technique available to the empirical researcher. This is a two-group *t*-test of the null hypothesis that there is no difference, in the population, between the average academic achievement of African-American children in the experimental (voucher) and control (no voucher) conditions.

To simplify our explanation of the new statistical concepts in this chapter, we base our presentation on the application of a *one-sided t*-test. This means that—in our introduction of the concept of statistical power—we test the null hypothesis that the average academic achievement of treated children is equal to the average achievement of untreated children versus an *alternative* hypothesis that their achievement is *greater* than that of control children, in the population. This is a strictly *pedagogic* decision on our part and was made to simplify our technical presentation. It contrasts with our earlier *substantive* decision to rely on a two-sided *t*-test in our detailed presentation of the actual analyses and findings from the NYSP project in Chapter 4. There, we assumed that, if the null hypothesis were rejected, the average achievement of children in the population who were offered vouchers could be either greater than, or less than, the average achievement of children not offered vouchers. Generally, in conducting research, a one-sided test should only be used in circumstances in which you can defend a strong prior belief that, if the treatment did have an effect on the outcome of interest, you would know with certainty what the direction of the difference in outcomes would be. This is rarely true in practice, and we do not believe it would be true in the case of empirical analyses of the NYSP data. On the other hand, as we discuss in Chapter 8, an example in which we believe a one-sided test would be appropriate concerns the impact of college scholarship aid on the decisions of high-school seniors to enroll in college. Since scholarship aid reduces the cost of college enrollment, it seems compelling to assume that, if scholarships did have an impact on the percentage of high-school seniors who enrolled in college, that effect would indeed be positive.

Fortunately, whether you choose a directional or a nondirectional alternative for your hypothesis testing, the technical concepts and connections that we introduce in this chapter—and, in particular, the concept of statistical power itself—remain unchanged. Later in the chapter, we describe how critical features of the research design, the measurement of the variables, and the choice of a particular data-analytic approach affect the statistical power in any particular experiment. At that point, we reconsider the decision to adopt a directional versus a nondirectional alternative hypothesis and comment on how it impacts the magnitude of the statistical power.

First, it is useful to recall the steps in the process of statistical inference that we made use of in the top panel of Table 4.1. There, to test the null hypothesis that students who were offered a NYSP voucher had academic achievement three years later that was no different from students who lost out in the voucher lottery, we first adopted a suitable α-level (of 0.05) to fix the Type I error of our test at 5%. Second, we computed the value of an observed *t*-statistic, obtaining a value of 2.911, using the following formula:

$$t_{observed} = \frac{\left(\overline{POST\_ACH}_V - \overline{POST\_ACH}_{NV}\right)}{\sqrt{s^2\left(\frac{1}{n_V} + \frac{1}{n_{NV}}\right)}} \qquad (6.1)$$

where subscripts *V* and *NV* are intended to distinguish the voucher and no-voucher groups, and s^2 and *n* refer to the pooled variance of post-test academic achievement and the number of African-American children in the respective groups. Third, based on our adopted α-level, we determined a critical value of the *t*-statistic under the null hypothesis at the appropriate degrees of freedom (here, 519).[2] This critical value, in the case of a one-sided test favoring the experimental voucher group, is 1.648. Fourth, because the magnitude of the observed *t*-statistic (2.911) exceeded the critical value (1.648), we rejected the null hypothesis that African-American children with, and without, vouchers performed identically in academic achievement, on average, in the population. Hence, we concluded—because our research design was a randomized experiment—that voucher receipt caused the observed difference of about 5 points in academic achievement between members of the treatment and control groups.[3]

[2] There were a total of 521 children in the sample.

[3] You can also proceed by referring to the *p*-value associated with the statistic of interest. This estimates the probability that you could have obtained your single empirically obtained estimate of the parameter of interest, or something more extreme than it, by an accident of sampling from a population in which the value of the parameter was 0—that is, from a population in which the null hypothesis was true. In the *t*-test conducted here, for instance, the *p*-value was 0.004 (Table 4.1, upper panel), meaning it was unlikely that we could have obtained our single empirically obtained average treatment/control difference of 4.899 and its companion *t*-statistic of 2.911, or something larger, by an accident of sampling from a "null" population. So, we conclude that, in the reality of the actual experiment, we were probably not sampling from a null population, but from an alternative population in which there was indeed a relationship between academic achievement and voucher receipt.

Notice how the construction of the observed *t*-statistic, which we defined in Equation 6.1, is conceptually appealing. Its numerator is equal to the sample mean difference in academic achievement between the voucher and no-voucher groups. Its denominator is simply the standard error of the difference in means between the groups—that is, the standard error of the quantity that sits in the numerator.[4] So, the observed *t*-statistic is just the sample mean difference between the voucher and no-voucher groups expressed in appropriate standard error units. More compellingly, provided that the original achievement scores are normally distributed, theoretical work in statistics shows that all such statistics formed in this way have *t*-distributions. So, we were able to use our existing knowledge of the *t*-distribution to determine a critical value for comparison with the observed test statistic, in order to carry through on the test.

As you know, through a process of sampling from the underlying population, the observed *t*-statistic in which we are interested—that is, the "2.911" obtained in our NYSP analyses—derives its value implicitly from an underlying and critically important parameter representing the average difference in academic achievement between African-American children with, and without, vouchers in the population. We write this important population difference in means as $(\mu_V - \mu_{NV})$, where subscripts V and NV refer to the experimental "voucher" and control "no-voucher" groups, respectively and, in what follows, for convenience, we refer to it as $\Delta\mu$. If the mean difference in the population $\Delta\mu$ were large, the corresponding difference in mean academic achievement that we would obtain by drawing samples from that population—and the corresponding value of the accompanying observed *t*-statistic—would also tend to be large, except that the idiosyncrasies of random sampling might occasionally toss up some radically different value than we had anticipated. Conversely, if the important population mean difference $\Delta\mu$ were actually equal to zero in the population, then any corresponding sample mean difference observed in the sample—and, consequently, the value of the corresponding observed *t*-statistic—would tend to be close to zero, except for the idiosyncrasies of sampling.

The complete formal logic of hypothesis testing is actually a little more complex than intimated up to this point, and it is from this added complexity that the notion of statistical power derives. When we conduct a hypothesis test, we actually contrast what we have learned from the empirical data with what we might anticipate under a *pair* of hypothetical settings. The first of these settings we have commented upon earlier. It is

[4] Under the assumption of homoscedasticity for the population residual variance.

described by the null hypothesis H_0, and under it we imagine there exists a hypothetical "null" population in which the important population mean-difference parameter $\Delta\mu$ is actually equal to zero (i.e., we stipulate that H_0: $\Delta\mu = 0$). The second, and equally important, setting is provided by an alternative hypothesis H_A, in which we establish a second hypothetical population where the value of the important population mean-difference parameter is *not 0*, but equal to some non-0 value of magnitude δ (i.e., we will stipulate that $\Delta\mu = \delta$, under the alternative hypothesis H_A). In quantitative research, we are usually interested in rejecting the null hypothesis in favor of the alternative, and then interpreting δ substantively.

Classical hypothesis testing simply contrasts the vicissitudes of the empirical setting, as encapsulated in the single empirically obtained value of the $t_{observed}$ statistic, with the set of values that the test statistic could potentially take on if we were to sample repeatedly and independently from populations in which these null and alternative hypotheses are true, respectively. Of course, we expect that values of $t_{observed}$ obtained in repeated resamplings would be scattered randomly and naturally by the idiosyncrasies of sampling. But, we anticipate that they would be scattered around the value *zero* if we were sampling from the null population, and around some non-zero value that depends on δ if we were sampling from the alternative population.[5] Then, if we found that the actual value of $t_{observed}$ obtained in our actual experiment was close to zero, and fell within a range of values that we might naturally anticipate in the "idiosyncratic scattering from the null" case, we would prefer the "*It came from H_0*" explanation and consequently accept that $\Delta\mu = 0$. If our single empirically obtained value of $t_{observed}$ was large, on the other hand, and looked more like a value that we might have gotten in an "idiosyncratic scattering from the alternative" case, then we will prefer the "*It came from H_A*" explanation and accept that $\Delta\mu = \delta$. Picking a sensible α-level for our test is how we choose between these two potential explanations.

We summarize these aspects of the hypothesis-testing process in Figure 6.1. In the top panel, under the symmetric hill-shaped "envelope," we represent the distribution of the values that a *t*-statistic could potentially take on, in random resampling from a "null" population in which

[5] Unfortunately for the pedagogy of our example, the "some non-0 value" to which we refer in this sentence is not δ itself, but a linear function of it. This is because, under the alternative hypothesis, the observed *t*-statistic has a non-central *t*-distribution whose population mean is equal to δ *multiplied by a constant whose value is* $\sqrt{\frac{\upsilon}{2}}\left(\frac{\Gamma((\upsilon-1)/2}{\Gamma((\upsilon)/2}\right)$, where υ represents the degrees of freedom of the distribution and $\Gamma(\ )$ is the gamma function.

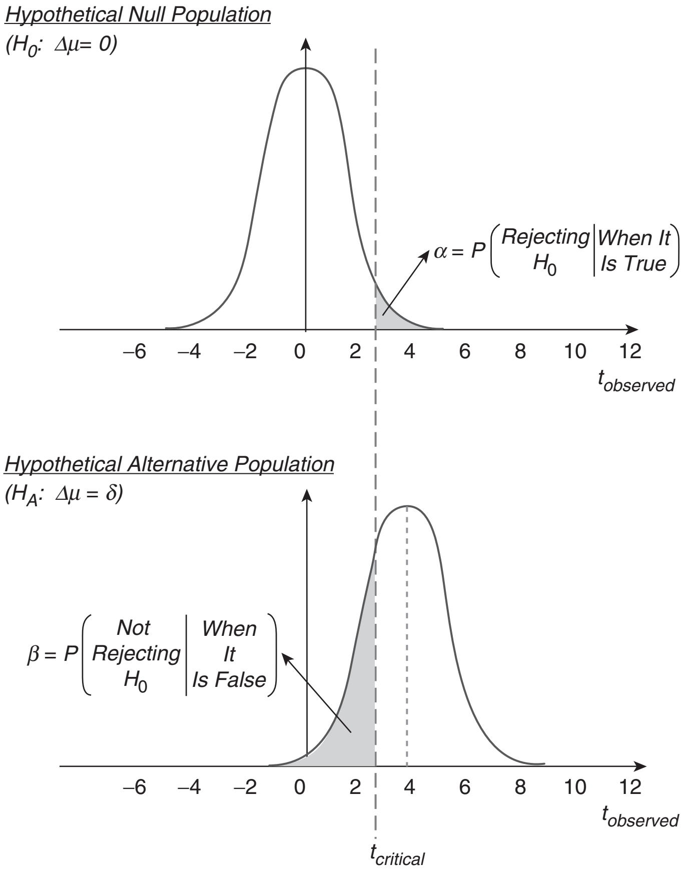

Figure 6.1 Distributions of the observed *t*-statistic ($t_{observed}$) under competing null (H_0) and alternative (H_A) hypotheses, showing the Type I error (α), Type II error (β), and placement of the critical value of the *t*-statistic ($t_{critical}$), for a one-tailed test of population outcome mean differences between a treatment and a control group.

the population mean-difference parameter $\Delta\mu$ was equal to zero. Although *zero* may not be the magnitude that we ultimately hope population outcome mean difference $\Delta\mu$ will have in our actual experiment (in fact, we usually hope that it is not zero), this idea of sampling repeatedly from a "null population" provides us with a useful baseline for subsequent comparison. Conceptually, the curve in the top panel represents something akin to a histogram of all the idiosyncratic values that $t_{observed}$ could possibly take on if we were to resample an infinite number of times from a population in which the population outcome mean difference between treatment and control conditions, $\Delta\mu$, was zero. As in any histogram, the horizontal axis represents the possible values that $t_{observed}$ could attain—actually, these values range from $-\infty$ to $+\infty$. The vertical axis represents the "frequency" with which each value has occurred during the resampling process. However, we are dealing with infinite resampling and a statistic that can take on values ranging continuously between plus and minus infinity. Consequently, we have drawn the exhibit as the envelope of a *probability density function* (or pdf) in which the histogram has been rescaled so that the total area under the envelope is equal to 1. Areas beneath the envelope represent the *probabilities* with which particular *ranges* of values of $t_{observed}$ would occur in infinite resampling from a null population. For instance, the probability that $t_{observed}$ will take on *any* value at all is obviously 1, a value equal to the total area beneath the pdf.[6] Similarly, because the pdf is symmetric and centered on zero, there is a probability of exactly one half—a 50% chance—that a value of $t_{observed}$ sampled at random from the null population will be larger than zero, or smaller than zero.[7]

In the bottom panel in Figure 6.1, we display the situation that would occur under the competing alternative hypothesis, H_A: $\Delta\mu = \delta$. The graphic is essentially identical to that displayed under H_0, but we have shifted the pdf of $t_{observed}$ to the right by an amount that depends on δ—the value we would anticipate for the population outcome mean difference between treatment and control groups if H_A were true.[8] Again, the displaced pdf represents the distribution of all the possible values of $t_{observed}$ that could be obtained if samples were drawn repeatedly and randomly from the alternative population.

To complete our test, we rely on a decision rule that derives from our decision to set the Type I error of our test at 5%. From this decision, we can derive a *critical value* against which to compare the value of the observed test statistic. We do this by determining the value that $t_{observed}$ would have to take on in order to split the null distribution of $t_{observed}$ in the top panel of Figure 6.1 vertically into two parts, with 5% percent of the area beneath its envelope falling to the right of the split and 95% falling to the left.[9] In the figure, we indicate the place at which this split occurs

[6] The area beneath the *t*-distribution is finite, and equal to 1, because its tails asymptote to zero.

[7] Not all distributions of test statistics are symmetric and zero at the center. However, the logic of our argument does not depend for its veracity on the particular shape of the pdf we have chosen to display. All that is required is that the pdf of the test statistic, under H_0, be known. Consequently, our argument applies equally well to cases in which distributions are asymmetric (as with the F and $\chi 2$ distributions).

[8] Again, under the alternative hypothesis, the pdf of the observed *t*-statistic is not centered on the value of δ itself, but on a value proportional to it. See footnote 5.

[9] Recall that this is a one-sided test.

by drawing a dashed vertical line. The place at which the vertical dashed line intersects the horizontal axis provides the required critical value of the test statistic $t_{critical}$ that we will use in our hypothesis test. Our decision is then straightforward. If $t_{observed}$ is greater than $t_{critical}$, then we conclude that it is probably too extreme to have come legitimately from the null distribution. Consequently, we reject H_0 in favor of H_A, and conclude that parameter $\Delta\mu$ is equal to δ, not zero, in the population from which we have sampled. On the other hand, if $t_{observed}$ is less than $t_{critical}$, we conclude that our single empirical value of $t_{observed}$ was probably sampled from a null population Consequently, we would not reject H_0 in favor of H_A. In other words, by choosing a particular α-level (5%, say) to fix the level of the Type I error, and combining this with our theoretical knowledge of the shape of the pdf of the *t*-statistic under the null hypothesis, we can carry out the desired test. It is the choice of the Type I error that provides us with the criterion that we need to make the testing decision.

Now focus on the lower second panel in Figure 6.1, which is aligned beneath the first. As we have noted, this lower panel illustrates the "alternative" side of the hypothesis testing situation. In it, we display the pdf of all possible values that an observed *t*-statistic could take on in repeated resampling from a population in which the alternative hypothesis was true, and parameter $\Delta\mu$ had a non-zero value of δ . Of course, because of sampling variation, it is entirely possible that, in some proportion of resamplings, $t_{observed}$ will take on very small values, perhaps even values less than $t_{critical}$—values that we typically associate with sampling from a null population—even though the alternative hypothesis is actually true. If this were to happen in practice, and we were to base our decision on an artificially small empirically obtained value, we would declare the null hypothesis true. In this case, we would have committed another kind of mistake—called a *Type II error*. Now, we would end up falsely accepting the null hypothesis even though the alternative was, in fact, true. The probability that $t_{observed}$ may be idiosyncratically less than $t_{critical}$, even when the alternative hypothesis is true, is represented by the shaded area under the "alternative" probability density function to the left of $t_{critical}$. Just as symbol α is used to represent the magnitude of Type I error, β is the symbol used to represent the probability of a Type II error.

Finally, notice the horizontal separation of the centers of the pdfs, under the competing null and alternative hypotheses, H_0 and H_A, in Figure 6.1. This separation reflects the difference in the potential values of $\Delta\mu$, under the alternative ($\Delta\mu$=δ) and null hypotheses ($\Delta\mu$=0).[10]

[10] Again, the horizontal distance between the centers of the H_0 and H_A pdfs is not equal to δ, but is proportional to it. See footnote 5.

Methodologists refer to the difference between the values of $\Delta\mu$ under H_0 and H_A—that is, δ or a sensible rescaling of it—as the *effect size*. If you conduct a statistical test and reject H_0 in favor of H_A, you can conclude that the important population outcome mean-difference parameter has magnitude δ, rather than zero. In other words, you will be ready to declare that you have detected an effect of the treatment. In analyses for our NYSP experiment, for instance, after rejecting H_0 in favor of H_A, we conclude that $\Delta\mu$ is certainly not zero, and we estimate its value under the alternative hypothesis—that is, δ—by the sample mean difference in the outcome between members of the treatment and control groups.

Under this definition, we could regard the effect size of the voucher treatment as simply equaling our best estimate of δ, and it would be measured in the same units as the outcome—student achievement, in the NYSP experiment. Of course, this scaling is arbitrary, because it is determined by the metric in which the outcome was measured. Two investigators could then end up with different values for the effect size if they chose to measure the same outcome on the same children using one achievement test rather than another. So, for greater uniformity and generality, effect size is usually redefined so that it can be communicated in standard deviation units. Thus, for each different test and test statistic, the mathematical features of the rescaling differ, but the consequences are the same. Once the rescaling is complete, investigators can refer to the effects of their experiments using statements like "a difference of a half standard deviation," "a quarter standard deviation difference," and so on. These kinds of statements can be understood by their colleagues and by remote audiences, regardless of the specific metric of the outcome measurement itself.

Based on these ideas, to facilitate communication, researchers have tended to adopt the set of loose standards that Jacob Cohen (1988) proposed for describing the magnitudes of effect sizes. Cohen proposed that in comparing an average difference in outcome between members of a treatment and a control group, we should regard a difference of eight-tenths (0.8) of a standard deviation a "large" effect, a difference of one-half (0.5) of a standard deviation a "moderate" effect, and two-tenths (0.2) of a standard deviation a "small" effect size.[11] For instance, in the case of the NYSP evaluation, recall that the difference in academic achievement

[11] Effect size can also be defined in terms of the *correlation* between outcome and predictor. In the NYSP evaluation, an effect size defined in this way would be the sample correlation between the academic achievement outcome and the dichotomous *VOUCHER* predictor, for the sample of African-American children. This correlation has a value of 0.127. When effect sizes are defined as correlations, a coefficient of magnitude 0.10 is regarded as a "small" effect size, 0.25 as a "medium" effect size, and 0.37 as a "large" effect size (Cohen, 1988, Table 2.2.1, p. 22).

between African-American children in the voucher and no-voucher conditions at the end of third grade was 4.899 points (Table 4.1, top panel). The standard deviation of academic achievement for these children was 19.209, and so we would say that effect size in the NYSP evaluation—which is then about a quarter of a standard deviation—was "small."[12] In our experiences, effect sizes of even the most successful interventions in education and the social sciences tend to be "small," when calibrated in Cohen's metric.

Defining Statistical Power

When conducting any hypothesis test, you have only two decisions to make. You can either reject H_0 because your obtained value of $t_{observed}$ is larger than the value of $t_{critical}$, or you can fail to reject it because $t_{observed}$ is smaller than $t_{critical}$. However, whichever of these two decisions you make, you can either be correct or you could have made a mistake. So, there is a "two-by-two" alignment of the testing decision with its consequences that leads to four possible decision scenarios. To two of these, by virtue of our definitions of Type I and Type II error, we can attach probabilities of occurrence. We summarize these four possible decision scenarios, and their associated decision probabilities, in the simplified graphical cross-tabulation in Figure 6.2.

In the figure, we have redisplayed the critical features of the H_0 and H_A pdfs that we displayed in Figure 6.1, along with the probabilities associated with their splitting vertically into parts by the placement of $t_{critical}$ (again represented by the vertical dashed line). The first row of the graphical cross-tabulation summarizes the distribution of $t_{observed}$ when H_0 is true; the second row summarizes its distribution when H_A is true. We comment briefly on each decision scenario below, beginning in the first row.

When H_0 *Is True and* $\Delta\mu$ *Is Equal to Zero (First Row)*

- *Right-hand cell.* Even though the null hypothesis is actually true in this row and there are no differences between the treatment and control group outcome means, in the population, you may find that your single empirically obtained value of $t_{observed}$ is idiosyncratically larger than the value of $t_{critical}$ simply by virtue of an accident of sampling. Then, you will reject H_0 by mistake and declare that

[12] Some argue that effect size is best scaled in terms of the standard deviation of the outcome for participants in the control condition only. In the NYSP evaluation, this would have led to an effect size of (4.899/17.172), or 0.285.

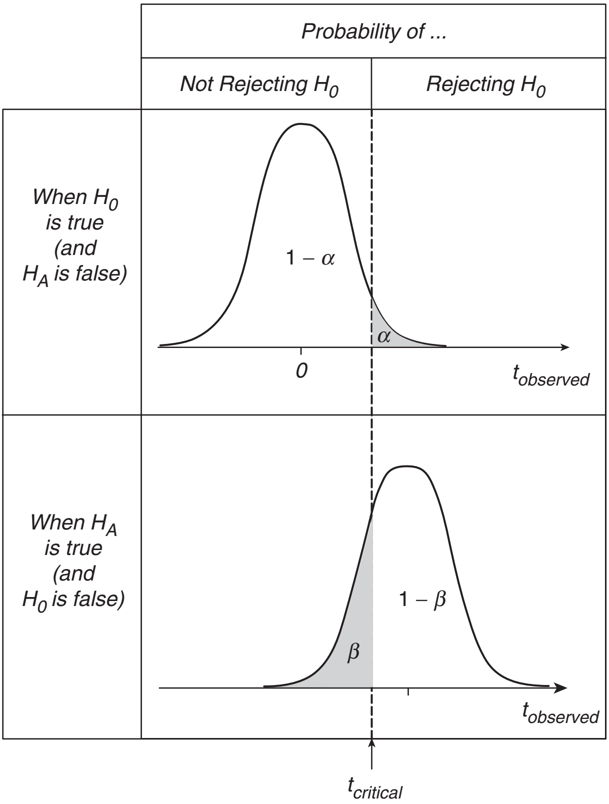

Figure 6.2 Four-way decision scenario, summarizing the probabilities of not rejecting H_0 (1st column) or rejecting H_0 (2nd column) when it is either True (1st row) or False (2nd row), showing the Type I error (α), Type II error (β), and placement of the critical value of the *t*-statistic ($t_{critical}$), for a one-tailed test of population outcome mean differences between a treatment and a control group.

$\Delta\mu$ is equal to δ incorrectly. In this case, you have made a Type I error, because you have falsely rejected your null hypothesis when it was correct. Such a decision scenario would occur if your actual experiment was one of those unfortunate occurrences in which a sample drawn from a truly null population generated a large value of $t_{observed}$ by an idiosyncratic accident of random sampling. However, because under this scenario H_0 is actually correct, the probability that you will make such a decision is equal to the area under H_0's pdf to the right of $t_{critical}$, which is of course the level of Type I error α that you yourself have picked in advance of the test. Thus, you have direct control over the Type I error probability, and you have an incentive to limit it by choosing a suitably small α-level, such as 0.05, for your test.

- *Left-hand cell.* On the other hand, you may find that your single empirically obtained value of $t_{observed}$ is appropriately smaller than your tabled value of $t_{critical}$, you will correctly fail to reject $H_{0,}$ and you will be right when you declare that $\Delta\mu$ is equal to zero. In this decision scenario, you have drawn a well-behaved small value of $t_{observed}$ from the null distribution, and it is appropriately less than $t_{critical}$. The probability that this decision scenario will occur is simply the area under H_0's pdf to the left of $t_{critical}$, or the *complement* of your self-selected Type I error, and is therefore equal to $(1 - \alpha)$.

When H$_A$ *Is True and* $\Delta\mu$ *Is Equal to* δ *(Second Row)*

- *Left-hand cell.* In this scenario, even though the alternative hypothesis is true and there are indeed differences between the treatment and control group outcome means, in the population, you may find that your single empirically obtained value of $t_{observed}$ is idiosyncratically smaller than the value of $t_{critical}$, again by an accident of sampling, and you will fail to reject H_0 even though it is false. Thus, you would incorrectly declare that $\Delta\mu$ is equal to zero. This would occur if your experiment was one of those occasions when a random sample from the alternative population happens to toss up an idiosyncratically small value of $t_{observed}$. Consequently, although H_A is actually true, your idiosyncratically small obtained value of $t_{observed}$ leads you to conclude that the sample was drawn from the null population. You have now made a Type II error. The probability that this decision scenario will occur is given by the area under H_A's pdf to the left of the value of $t_{critical}$. It is called the Type II error of the decision-making process, and we represent it by the symbol β. It is again a probability, just like α.
- *Right-hand cell.* Finally, you may find that your single empirically obtained value of $t_{observed}$ is appropriately larger than the tabled value of $t_{critical}$, and you will correctly reject H_0. In this scenario, your alternative hypothesis is true and you will be right when you declare that $\Delta\mu$ is equal to δ. The probability that this decision scenario will occur is equal to the area under H_A's pdf to the right of $t_{critical}$—it is the complement of Type II error, or $(1 - \beta)$.

This two-way cross-tabulation of the decision scenarios illustrates that the magnitudes of the several decision probabilities are interrelated.

To appreciate this fully, recall that, once the pdf of the test statistic has been specified under H_0, the value of $t_{critical}$ depends only on your selection of the α-level. So, if you were willing to entertain a larger Type I error, perhaps as high as 0.10, then your corresponding value of $t_{critical}$ would shrink, so that 10% of the area beneath H_0's pdf can now become entrapped to its right. With your new willingness to entertain this larger Type I error, you would find it easier to reject H_0 because the single empirically obtained value of your observed test statistic would be more likely to exceed the now smaller value of $t_{critical}$. This means that, if you can tolerate increased Type I error, you can more easily reject H_0 and more easily claim detection of a non-zero effect in the population. Of course, in enhancing your chances of claiming such a non-zero effect, you have increased the probability of Type I error—that is, you are now more likely to reject H_0 even when it is true! At the same time, shifting $t_{critical}$ to a smaller value has implicitly moved the vertical splitting of H_A's pdf to the left in Figure 6.2, and thereby reduced the value of the Type II error β. So, you are now more likely to accept H_A when it is true. This intimate—and inverse—connection between the magnitudes of the Type I and II errors is a central fact of statistical life. As you decide to make one type of error *less* likely, you force the other one to become *more* likely, and vice versa. So, you can correctly regard hypothesis testing as a trade-off between the probabilities of two competing types of error.

More importantly, the decision probability featured in the right-hand cell of the lower second row in Figure 6.2, which is of magnitude $(1 - \beta)$, is a central and important commodity in our empirical work. It is the *probability of rejecting H_0 when it is false*. Or, alternatively, it is the probability of accepting the alternative hypothesis when it is true. This is a highly preferred end result for most research—the rejection of the null hypothesis in favor of the alternative, when the alternative is true. For example, in designing the NYSP experiment, investigators were hoping to reject the null hypothesis of no causal connection between voucher receipt and student achievement in favor of an alternative hypothesis that stipulated voucher receipt had a causal effect on student achievement. This important quantity is defined as the *statistical power* of the study and, as you can see from Figure 6.2, it is simply the complement of the Type II error. This means that, knowing the pdfs of our test statistics—such as the *t*-statistic—under the null and alternative hypotheses, and being willing to set the Type I error level to some sensible value, means that we can actually estimate a value for the statistical power. This can be very useful both during the design of the research and also after the research has been completed. We follow up on these ideas in the section that follows.

Factors Affecting Statistical Power

Given this explanation, statistical power can be estimated prospectively for any research design, provided that you are willing to stipulate four things. First, you must be willing to anticipate the effect size that you hope to detect (e.g., Do you expect to be detecting a small, medium, or large effect?). Second, you must pick the type of statistical analysis you will eventually conduct (e.g., Will you use a *t*-test of differences in means, or more sophisticated methods of data analysis?). Third, you must pick an α-level for your future statistical inference (Are you happy with the 0.05 level?). Fourth, you must decide on the number of participants you want to include in your sample (Can you afford to recruit 200, 300, 400, or more participants?). The reason that these four decisions determine the statistical power of your prospective analysis is as follows. By choosing the method of statistical analysis, you identify the statistic that will be used to test your hypotheses. Knowing the test statistic and the prospective sample size determines the shape of the test statistic's pdf under H_0. Choice of the effect size then determines the pdf of the test statistic under H_A (typically, by displacing the pdf to the right).[13] Finally, overlaying the α-level on the test statistic's pdf under H_0 then fixes the critical value of the test statistic, which consequently determines the statistical power. We call this a *statistical power analysis*.

Often of greater interest, if you are willing to anticipate the effect size, specify a type of analysis, pick an α-level, and decide on the statistical power you want, you can figure out the sample size that will permit you to reach your analytic objectives. The actual computations underlying such statistical power analyses are complex, and they make use of theoretical knowledge of the mathematical shapes of the pdfs of the different test statistics under the null and alternative hypotheses, and of integral calculus. Consequently, we do not describe them here. But they are available for reference in standard statistical texts, and are most easily carried out by dedicated computer software, much of which is now available free on the Internet.[14] Instead, our purpose here is to give you a ballpark sense of the kinds of sample sizes that are needed for successful experimental research design in education and the social sciences, and the levels of

[13] Depending on the type of analysis, the test statistic's pdf under H_A may also have a different shape from its pdf under H_0.

[14] All the power analyses in this chapter were conducted using the G*Power freeware, v2.0, *GPOWER: A-Priori, Post-Hoc and Compromise Power Analyses for MS-DOS*, Dept. of Psychology, Bonn University, Germany, http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/.

statistical power that they typically provide. In addition, we hope to guide you toward the kinds of design decisions that will enable you to achieve your objectives as an investigator of cause and effect.

To provide you with some intuition about the sizes of sample that are required typically in a successful two-group experiment, we now present estimates of statistical power for an experiment in which we assign a sample of participants randomly and individually to either a treatment or a control condition, so that groups of equal size are formed. We again assume that a one-sided *t*-test will eventually be used to test a null hypothesis of no group differences in the outcome mean, in the population. For this empirical set-up, in Figure 6.3, we plot the obtained values of statistical power (vertical axis) at different values of the total sample size (the total number of participants in the treatment and control groups combined, on the horizontal axis). We do this for both small effect sizes ($ES = 0.2$, lower pair of curves) and medium effect sizes ($ES = 0.5$, upper pair of curves), at α-levels of 0.05 (solid lines) and 0.10 (dashed lines), respectively. We have not provided plots for the large effect size ($ES = 0.8$) condition because such effect sizes occur rarely in experimental research in education and the social sciences. You can replicate these plots by downloading standard statistical power analysis software from the Internet and inserting these values we have provided for effect size and Type I error (see footnote 14).

Inspecting the figure, you can discern three important relationships between statistical power and the other quantities involved. First, notice that statistical power is always greater when you adopt a more liberal α-level in your statistical testing. In Figure 6.3, at any pairing of effect and sample size, power is always greater when the α-level is 0.10 rather than 0.05. For instance, if you want to detect a small effect size ($ES = 0.2$) with a total sample size of 300, then choosing an α-level of 0.10 rather than 0.05 increases your statistical power from approximately 0.53 to 0.67, an improvement of more than 25%. Our earlier description of the nature of statistical power provides an explanation for why this occurs. Returning to Figure 6.2 and focusing on the first row, you will see that it is the choice of α-level that splits the area beneath the test statistic's pdf under H_0 and determines the test statistic's critical value $t_{critical}$. So, if you deliberately increase the value of α, from 0.05 to 0.10. say, the value of $t_{critical}$ must "shift to the left," so that a larger area (10%) can be entrapped under the H_0 pdf to its right. But, if $t_{critical}$ is shifted, any areas entrapped beneath the alternative probability density function in the second row of the figure must be affected. Specifically, the area to the left of $t_{critical}$ under the H_A pdf will be reduced, decreasing the value of the Type II error β, and increasing its complement, the statistical power.

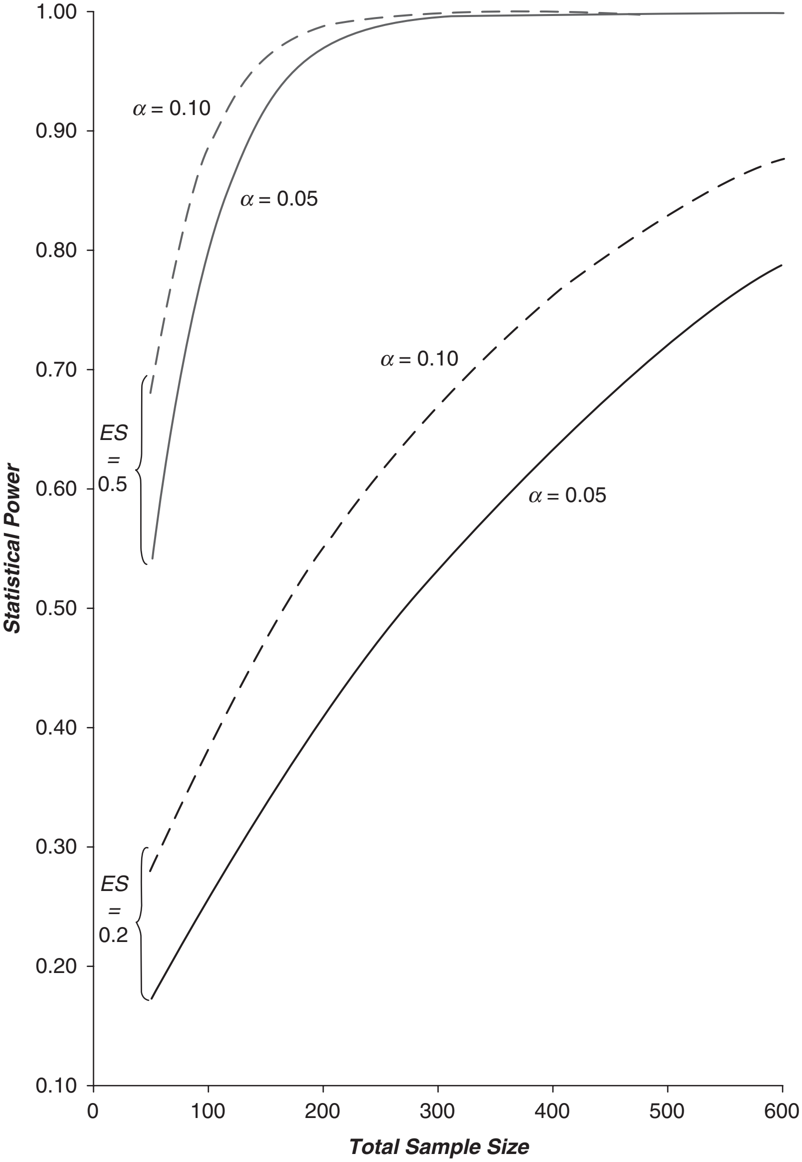

Figure 6.3 Statistical power as a function of total sample size, effect size (0.2 versus 0.5), and α-level (0.05 versus 0.10), for a one-tailed test of population outcome mean differences between a treatment and a control group.

The second important relationship evident in our figure is that, all else remaining equal, you will always have more power to detect a larger effect. In Figure 6.3, with a total sample size of 100 participants randomized to treatment conditions, say, and an α-level of 0.05, you have a power of just over 0.25 to detect a small effect ($ES = 0.2$) and a power of almost 0.80 to detect a medium effect ($ES = 0.5$). Again, the reason for this link between effect size and power can be deduced from our decision-scenario description in Figure 6.2. As we have noted already, the effect size determines the horizontal separation of the test statistic's pdfs under H_0 and H_A. So, if a larger effect size is accommodated, the H_0 and H_A pdfs must be more widely separated along the horizontal axis. But, in the first row of the figure, the center of H_0's pdf is fixed at zero (because it represents the "null" condition). So, as effect size is increased, the pdf of the test statistic under H_A shifts to the right, in the second row of the figure, sliding past the location of $t_{critical}$, the placement of which has been fixed by the earlier choice of α-level under the H_0 pdf. Consequently, the area beneath the alternative distribution to the right of $t_{critical}$ must rise, and statistical power is again increased.

Third, and most important, statistical power is always greater when the total number of participants included in the experiment is larger, all else being equal. This is quite a dramatic effect, as evidenced by the slopes of the power/sample size relationships in Figure 6.3. Notice, for instance, in research to detect a medium effect size ($ES = 0.5$) at an α-level of 0.05, statistical power can be increased from about 0.55 to more than 0.80 by increasing the total sample size from 50 to 100 participants! Although it is more difficult to understand, the reason for this dependency can again be deduced from Figure 6.2. As sample size increases, the pdf associated with any test statistic always becomes slimmer and taller because its values enjoy greater precision—and less scatter on repeated sampling—at larger sample sizes. However, the location of the center of the distribution remains unchanged.[15] So, as the H_0 and H_A pdfs in Figure 6.2 slim down and become more pointy, there are two important consequences, one for each featured pdf. First, in the H_0 pdf in the first row of the figure, the location of $t_{critical}$ must move to the left—that is, the critical value must get smaller—in order to accommodate the fixed choice of α-level adopted for the test. (Recall that choice of α-level splits the pdf under the null distribution vertically, so that an area equal to the Type I error must fall to the right of $t_{critical}$. In a rapidly narrowing distribution, this can only continue

[15] You can check out this claim using one of the simulations of the distribution of the sample mean as a function of sample size available on the Internet.

to occur if $t_{critical}$ shifts to the left, thereby becoming smaller.) Second, in the second row of the figure, the corresponding narrowing of the H_A pdf causes the area beneath it to be reapportioned on either side of the now fixed value of $t_{critical}$, with less of the area falling to the left and more falling to the right. The shift of the value of $t_{critical}$ to the left in the H_0 pdf in the first row and the reapportioning of the area beneath the H_A pdf in the second row both lead to a reduction in the area beneath the H_A distribution to the left of $t_{critical}$. Consequently, Type II error β is reduced and statistical power $(1 - \beta)$ is increased. We suggest that you download software for computing statistical power from the Internet, and try out some of these computations for yourself, based on experiments that you think you may want to conduct in your own area of substantive interest.

One of our own great concerns as social scientists and methodologists has always been that many empirical investigators do not have a realistic vision of the actual sample sizes that are required to conduct powerful and effective research. It is common that researchers underestimate the numbers of participants required for empirical success. For instance, if you were designing research to estimate the impact of private-school tuition voucher receipt on academic achievement, and you suspected that the effect size you might detect was small (as in the NYSP experiment), you could set your α-level to the "usual" 0.05 level of statistical significance and strive for moderate power in the region of 0.80. From the plot in Figure 6.3, with these values set, you can see that you require a total sample size of about *620 participants*—distributed randomly into equally sized treatment and control groups—to have a reasonable hope of successfully detecting an effect size of 0.20. If you are unhappy with the idea that there is a 20% chance that you would declare the null hypothesis to be true when in fact this is not the case (a Type II error of 0.20), you might want to shoot for a power of 0.90. Then, you would need a total of 860 participants in your sample. This makes even the NYSP experiment a little underpowered for investigating the causal impact of tuition vouchers on African-American children, as there are only 521 of these children in the sample.

If you need more power for your experiment, or if the predicted effect size is smaller than 0.20, or if you want to detect the same treatment effect in multiple subgroups (e.g., among different race/ethnicities), then your sample must be considerably larger than these targets. Do not underestimate the sample size that you will require for your research. In underpowered research, you will never know whether you have failed to reject the null hypothesis because it is true, or because you simply did not have sufficient power to confirm the alternative. This problem plagued many of the studies of elementary-school mathematics interventions that the WWC reviewed.

The Strengths and Limitations of Parametric Tests

Keep in mind that the magnitude of the statistical power available in a particular investigation also depends on the type of statistical technique selected for data analysis. In our earlier examples, to underpin our technical presentation and form a basis of our "ball-park" estimates of power and sample size, we have focused on the simplest possible kind of statistical analysis that you can conduct in data drawn from a two-group experiment—the two-group *t*-test. In focusing on the use of this simple technique, we intended to provide a "baseline" set of recommendations about sample size and statistical power in research design.

However, many other statistical techniques are available for analyzing data, even for analyzing data from a simple two-group experiment, and some of them are more powerful than others. As a guiding principle, statistical techniques are more powerful when they incorporate more information into the analysis. Other than simply collecting data on more participants, there are two straightforward ways to achieve this—you can either make stronger assumptions about the data and the statistical model upon which the analysis is based, or you can add covariates to the analysis. Generally, analytic techniques that make more stringent assumptions are more powerful than those with weaker assumptions. The reason is that the assumptions themselves constitute a kind of information that is incorporated in the analysis. For example, among techniques for comparing the average outcomes of a treatment and a control group, the *t*-test is intrinsically more powerful than the nonparametric Wilcoxon rank test. In fact, as a general principle, parametric statistical tests are always more powerful than the corresponding nonparametric tests. This is because the *t*-test, and other traditional parametric tests like those that automatically accompany ordinary least-squares (OLS) regression analysis and the analysis of variance, make stronger assumptions about the distribution of the outcome in the analyses. The *t*-test, for instance, assumes that participants' values of the outcome are independently and normally distributed with homoscedastic variance in the treatment and control groups.[16] These stringent assumptions provide additional information that contributes greatly to the power of the analysis. Of course, you don't get anything for nothing. In choosing to use a test like the *t*-test over the Wilcoxon rank test, you are relying heavily on the validity of these additional parametric assumptions. This means that the added assumptions must be valid in order for the results of your analysis to be correct. If the assumptions are

[16] Some versions of the *t*-test relax the population homoscedasticity assumption.

violated, then your answer may be wrong no matter how powerful the technique!

The Benefits of Covariates

A second direct way to bolster the statistical power of your analysis is to add covariates to your statistical models. Techniques like multiple-regression analysis, for instance, are more powerful than simpler techniques like a *t*-test of differences in means, for this reason. As we described in Chapter 4, a research question about the equality of average academic achievement between a voucher and a no-voucher group can be addressed in data either by a *t*-test of differences in sample means or by regressing the achievement outcome on a dichotomous "question" predictor that distinguishes participants' membership in the treatment or control group. If no covariates were included in the regression model, both approaches would provide identical answers and have identical power.

However, the regression analysis approach lets you include judiciously selected additional variables—exogenous measures of the children's demographic background, home life, and prior achievement—as covariates or control predictors to the analysis, without increasing the sample size. Providing the new covariates are well behaved—that is, reliably measured, linearly related to the outcome, uncorrelated with the "treatment" predictor,[17] and independent of the existing residuals in the model (exogenous)—their inclusion will tend to increase the proportion of the outcome variation that is predicted when the model is fitted (i.e., increase the value of the R^2 statistic) and thereby reduce residual variance. A reduction in residual variance necessarily implies a shrinking of the standard errors associated with the estimation of regression parameters, and an (inversely proportional) increase in the magnitude of *t*-statistics associated with the predictors. A larger *t*-statistic means that you are more likely to reject the null hypothesis and therefore your analysis has greater power at the same sample size. This is evident in the third panel of Table 4.1, where the standard error associated with the *VOUCHER* predictor has declined from 1.683 to 1.269 on inclusion of student pre-test scores as a covariate, and the *t*-statistic associated with the impact of the voucher treatment increased correspondingly from 2.911 to 3.23. In general, the impact of covariates on power can be dramatic. For instance, Light, Singer, and Willett (1990) comment that, if you include in your regression analyses

[17] If treatment status is assigned randomly by the investigator, then the treatment predictor will necessarily be uncorrelated with *all other* exogenous covariates.

a set of covariates that predict about half the variation in the outcome jointly, then you can maintain the same statistical power for your analyses at half the sample size.

The message is clear. There is always an analytic advantage to preferring a more complex statistical analysis over a less complex one because it provides you with an opportunity to increase precision by including covariates. Greater precision brings increased statistical power, and the ability to detect a smaller effect at the same sample size. However, significant knowledge is needed to use complex statistical analyses appropriately. In doing so, you are relying more heavily on the hypothesized structure of the statistical model. You have to ensure that additional assumptions are met. You have to do a good job, analytically speaking, with the new terms in the model. You need to worry about whether the new covariates meet the underlying requirements of the analysis in terms of the quality of their measurement, the functional form of their relationship with the outcome, whether they interact with other predictors in the model, and whether they are truly independent of the existing residuals, as required. Clearly, everything has its price! However, if it is a price that you can pay, the rewards are great.

The Reliability of the Outcome Measure Matters

An additional factor to consider when figuring out how large a sample you will need for your research is the *reliability* of your outcome measure. To this point, we have assumed that the measurement of the outcome variable has been perfectly reliable. Of course, this is rarely the case in practice. All measures of observed quantities suffer from some level of unreliability as a result of the presence of random measurement error. Standardized measures of student achievement, such as those administered in the NYSP experiment, may have reliabilities above 0.90. Measures of many other constructs, particularly those with less precise definitions, or those that seek to document participants' self-reported beliefs and opinions, may have reliabilities that fall as low as 0.60.

Although psychometricians define the reliability parameter formally as a ratio of the population variances of the true and observed scores (Koretz, 2008), you can regard measurement error as being the random "noise" that obscures the true "signal" in an outcome variable. Measures that are less reliable obscure the true signal to a greater extent and therefore make it more difficult to detect treatment effects. This means that one simple approach for assessing the impact of outcome unreliability on statistical power computations is to view it from the context of effect size. Ultimately, we are conducting research so that we can detect the presence of true effects, and so we must account for measurement unreliability in our designation of observed effect size for the purposes of statistical power computation. In other words, because measurement fallibility undermines our ability to detect effects, we must plan our research in anticipation of even smaller effects than we would hope to detect in a world of perfect measurement.

Specifically, if you want to detect a true effect of a particular size, then you must design your research to seek an observed effect size that is equal to the anticipated true effect size, *multiplied by the square root of the reliability of the outcome variable*. The newly attenuated effect size thus obtained can then be incorporated into your power computations in the usual way. To give you some sense of the magnitude of the correction, imagine that you set your α-level at the 0.05 level of statistical significance and are planning to design a two-group randomized experiment that will have a statistical power of 0.80 to detect a small effect ($ES = 0.2$). We noted earlier that you should anticipate requiring a total sample size of 620 participants. If your outcome reliability were less than perfect, but at the level of most published achievement tests—around 0.95, say—then you would need to conduct power analyses in anticipating the detection of a new effect size of 0.195—that is, 0.2 multiplied by the square root of 0.95. To compensate for this small decline in effective effect size, total sample size would have to increase by 32 participants to 652. However, if your outcome reliability fell as low as 0.85, then your sample size would need to rise by 112 participants to 732. Notice that, because we take the square root of the outcome reliability (an estimate that always falls between 0 and 1) before conducting the new power analysis, the impact of measurement reliability—in its typical ranges (0.85 to 0.95)—is mitigated and the impact on sample size is of the order of a few percent. Outcome reliability would have to fall to 0.16, for instance, before measurement unreliability would force you to reclassify a "medium" effect as "small."

Although the impacts on power and sample size are not enormous when the reliability of measurement is reasonably high, it is worth paying attention to the potential impact of measurement reliability on your power analyses. Specifically, we suggest that you incorporate two steps in your research planning in order to deal with reliability of measurement. First, you should always make sure—by pre-research piloting, detailed item analysis, and prior editing and refinement of your instruments—that you administer measures of the highest reliability possible for the construct, audience, and context in your research. Second, you should always anticipate the presence of measurement error in your assessment of effect size and conduct your power analyses at that smaller effect size. Fortunately, with any decently constructed and reasonably reliable measure, this will probably mean that you will only have to increase your anticipated total sample size by a few percent.

The Choice Between One-Tailed and Two-Tailed Tests

Finally, we return to the question of whether it makes sense to adopt a one-tailed (directional) or a two-tailed (nondirectional) test when conducting data analyses. Earlier, in our replication of the original analyses of the NYSP data in Chapter 4, we made use of a two-tailed test. The reason was that we wanted to retain an open mind and proceed as though the jury were still out on the effectiveness of educational vouchers. If we were ultimately to reject a null hypothesis of no group difference in outcome between those randomly assigned vouchers and those not, we did not want to prejudge whether any detected effect favored the voucher recipients or control-group members.

In contrast, when we reviewed the concept of hypothesis testing and introduced the notion of statistical power in this chapter, we made use of a one-tailed test. We did this to make our pedagogic explanations of Type I and Type II error simpler. In particular, this decision allowed us to focus only on the single *upper* tail of the pdf of the test statistic, under H_0, and the area trapped beneath it, in Figures 6.1 and 6.2. Now that these concepts have been established, it makes sense to consider the consequences for statistical power analysis of the choice between a non-directional (two-tailed) and a directional (one-tailed) test. The answer is straightforward.

When you adopt a one-tailed test, essentially you place your entire reservoir of Type I error—typically, 5%—into the area trapped beneath the upper tail of the pdf of the test statistic under H_0 and the critical value. This is what we are illustrating in the first row of Figure 6.2. By adopting an α-level of 5%, say, and insisting on a one-tailed test, we fix the critical value of the *t*-statistic at the place already displayed in the figure.

If we were to now change our minds and opt for a two-tailed test, we would need to adopt a new critical value for the *t*-statistic, and this would affect both our Type II error and statistical power. For instance, under the non-directional testing option, we would need to accept that Type I error could potentially occur at either end of the pdf of the test statistic under H_0. We could falsely reject the null hypothesis because the value of $t_{observed}$ was driven to be either too large or too small as a result of the idiosyncrasies of sampling. Either way, we would reject H_0 incorrectly, and commit a Type I error. As a result, we need to split our adopted Type I error level—usually, 5%—into two halves, each of 2.5%. We would then choose a new critical value of the *t*-statistic, so that 2.5% of the area beneath the pdf of the test statistic (under H_0) was entrapped to the right of its positive value at the upper end and 2.5% of the area was entrapped to the left of its negative value at the lower end.[18] As a consequence, the magnitude of the new $t_{critical}$ must be larger than that currently displayed.

In going from the existing critical value of the *t*-statistic obtained under the one-tailed test of our initial explanation to the new larger critical value, we have effectively moved the vertical dashed reference line in Figure 6.2—the line that also splits the pdf of the *t*-statistic under H_A, in the second row of the figure—to the right. Thus, the Type II error (β)—represented by the area entrapped beneath the pdf of the test statistic (under H_A) to the left of the dashed vertical line—will have increased. Concurrently, the statistical power—the complement of that area, to the right of the vertical dashed line—must be reduced. Thus, switching from a one-tailed to a two-tailed test implicitly reduces the power of a statistical test.

We conclude by reminding you then that, in most research, two-tailed tests are the order of the day, even though they are implicitly less powerful than one-tailed tests. Only when you can mount a compelling defense of the argument that a particular policy or intervention can have only a directed impact (positive or negative) on the outcomes of interest, in the population, is the use of one-tailed tests justified.

What to Read Next

If you want to learn more about statistical power, we suggest that you consult the classic text by Jacob Cohen entitled *Statistical Power Analysis for the Behavioral Sciences* (1988, 2nd edition).

[18] Implicitly, in the two-tailed case, because the pdf of the *t*-statistic (under H_0) is symmetric, $t_{critical}$ will take on two values of the same magnitude—one positive and the other negative—which are equally spaced on either side of the center of the pdf. During the subsequent test, if the value of the observed *t*-statistic is positive, it will be compared to the upper positive value of $t_{critical}$; if it is negative, it will be compared to the lower negative value.