Chapter twenty three

Special Topics: Statistical Power

Assessing the risk of Type II error

How is Statistical Power Defined?

How Do Significance Criteria Influence Statistical Power?

How Does Effect Size Influence Statistical Power?

How Does Sample Size Influence Statistical Power?

Estimating statistical power

How Do We Define the Significance Criteria and Effect Size in a Statistical Power Analysis?

How Do We Determine the Sample Size Needed to Ensure a Statistically Powerful Study?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_23,

As we have seen in earlier chapters, criminal justice researchers place a premium on statistical inference and its use in making decisions about population parameters from sample statistics. In assessing statistical significance, the focus is the problem of Type I, or alpha (α), error: the risk of falsely rejecting the null hypothesis. Paying attention to the statistical significance of a finding should keep researchers honest, because it provides a systematic approach for deciding when the observed statistics are convincing enough for the researcher to state that they reflect broader processes or relationships in the general population from which the sample was drawn. If the threshold of statistical significance is not met, then the researcher cannot reject the null hypothesis and cannot conclude that a relationship exists.

Another type of error that most criminal justice researchers are aware of, but pay relatively little attention to, is Type II, or beta (β), error: the risk of falsely failing to reject the null hypothesis that we originally introduced in Chapter 6. A study that has a high risk of Type II error is likely to mistakenly conclude that treatments are not worthwhile or that a relationship does not exist when in fact it does. Understanding the risk of a Type II error is crucial to the development of a research design that will give the researcher a good chance of finding a treatment effect or a statistical relationship, if those effects and relationships exist in the population. This is fundamentally what we mean by statistical power—given the current design of a study, does it have the ability (i.e., the power) to detect statistically significant effects and relationships?

Although researchers in criminal justice have placed much more emphasis on the statistical significance than on the statistical power of a study, research in fields such as medicine and psychology routinely reports estimates of statistical power.[1] Federal funding agencies (e.g., National Institutes of Health) typically require research proposals to estimate how powerful the proposed research design will be. The purpose of this chapter is to present an introductory discussion of the key components in an assessment of the statistical power of a research design

[1] See, for example, S. E. Maxwell, K. Kelley, and J. R. Rausch "Sample Size Planning for Accuracy in Parameter Estimation," *Annual Review of Psychology* 59 (2008): 537-563.

and to explain why it is important for criminal justice researchers to have a basic understanding of the importance of statistical power in designing and evaluating criminal justice research.

Statistical Power

Statistical power measures the probability of rejecting the null hypothesis when it is false, but it cannot be measured directly. Rather, statistical power is calculated by subtracting the probability of a Type II error—the probability of falsely failing to reject the null hypothesis—from 1:

Power = 1 – Probability (Type II error) = 1 – β.

For many sample statistics, the Type II error can be estimated directly from the sampling distributions commonly assumed for most test statistics. In contrast to a traditional test of statistical significance, which identifies for the researcher the risk of stating that factors are related when they are not (i.e., the Type I error), statistical power measures how often one would fail to identify a relationship that in fact does exist in the population. For example, a study with a statistical power level of 0.90 has only a 10% probability of falsely failing to reject the null hypothesis. Alternatively, a study with a statistical power estimate of 0.40 has a 60% probability of falsely failing to reject the null hypothesis. Generally as the statistical power of a proposed study increases, the risk of making a Type II error decreases.

Figure 23.1 presents the relationship between Type I and Type II errors graphically. Suppose that we are interested in a difference in group means, say between a control and treatment group in a criminal justice experiment, and based on prior research and theory, we expect to find a positive difference in the outcome measure. We would test for a difference in the group means by using a one-tailed *t*-test. If we have 100 cases in each group, then the critical *t*-value is 1.653 for $\alpha = 0.05$. The distribution on the left side of Figure 23.1 indicated by the solid line represents the *t*-distribution—the sampling distribution—under the null hypothesis, with the significance level (α) shaded in the right tail of the distribution and denoted with α—the vertical line represents the critical *t*-value.

Figure 23.1 *Graphical Representation of Type I and Type II Errors in a Difference of Means Test (100 Cases Per Sample)*

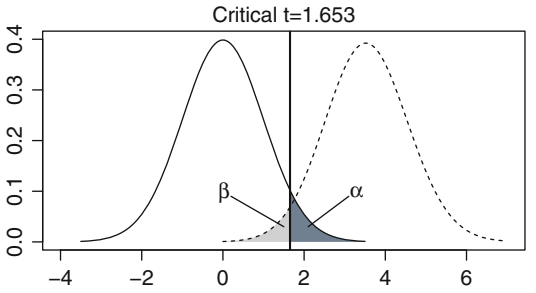
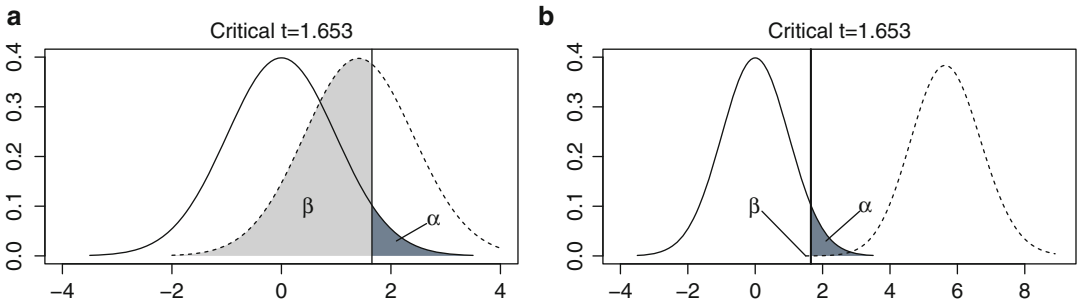


Figure 23.2 *Graphical Representation of Type I and Type II Errors in a Difference of Means Test—Changing the Difference in Mean Values. (a) Smaller Difference in Means—Fixed Sample Size. (b) Larger Difference in Means—Fixed Sample Size*



The distribution on the right side of Figure 23.1 and indicated by the dashed line represents the hypothesized sampling distribution based on prior research and theory and our expectations for the expected differences in the two group means. The hypothesized sampling distribution is also a *t*-distribution, but it is known as a non-central *t*-distribution—we illustrate below how this distribution is used to compute statistical power. The probability of making a Type II error (β) is denoted in the figure and is the cumulative probability in the distribution on the right up to the critical *t*-value (i.e., $t < 1.653$). The statistical power of this difference of means test is represented in the figure by the area under the dashed line that falls to the right of the critical value—the difference between 1 and β—and represents *t*-values of 1.653 and above, which would be interpreted as evidence in favor of rejecting the null hypothesis.

It is important to note that our estimate of β is fully dependent on our estimate of the expected magnitude of the difference between the two groups. Figure 23.2 illustrates the differences for two alternative effect sizes while assuming that the sample sizes remain fixed at 100 cases per group. For example, if we expect the difference of means to be smaller, we would shift the hypothesized sampling distribution (the dashed line) to the left, increasing our estimate of β (see Panel (a)). If we expect a larger difference, we would shift the hypothesized sampling distribution to the right, reducing the estimate of β (see Panel (b)).

If the statistical power of a research design is high and the null hypothesis is false for the population under study, then it is very likely that the researcher will reject the null hypothesis and conclude that there is a statistically significant finding. If the statistical power of a research design is low, it is unlikely to yield a statistically significant finding, even if the research hypothesis is in fact true. Studies with very low statistical power are sometimes described as being "designed for failure," because a study that is underpowered is unlikely to yield a statistically significant result, even when the outcomes observed are consistent with the research hypothesis.[2]

[2] For an extended discussion of this, see D. Weisburd "Design Sensitivity in Criminal Justice Experiments," *Crime and Justice* 17 (1991): 337-379.

Consider the implications for theory and practice in criminal justice of a study that has low statistical power. Suppose that a promising new program has been developed for dealing with spouse assault. If that program is evaluated with a study that has low statistical power, then the research team will likely fail to reject the null hypothesis based on the sample statistics, even if the program does indeed have the potential for affecting spouse assault. Although the research team is likely to say that the program does not have a statistically significant impact on spouse assault, this is not because the program is not an effective one, but because the research team designed the study in such a way that it was unlikely to be able to identify program success. Conceptually, this same problem occurs in the analysis of other types of data when trying to establish whether a relationship exists between two theoretically important variables. The relationship may exist in the population of interest, but a study with low statistical power will be unlikely to conclude that the relationship is statistically significant.

One might assume that researchers in criminal justice would work hard to develop statistically powerful studies, because such studies are more likely to support the research hypothesis proposed by the investigators. Unfortunately, statistical power is often ignored altogether by criminal justice researchers, which results in many criminal justice studies having a low level of statistical power.[3]

Setting the Level of Statistical Power

What is a desirable level of statistical power? There is no single correct answer to this question, since it depends on the relative importance of Type I and Type II errors for the researcher. That said, one of the more common suggestions in the statistical power literature has been that studies should attempt to achieve a power level of 0.80, meaning that the chances of a Type II error are $\beta = 0.20$. There are many ways in which this is an arbitrary threshold. At the same time, it implies a straightforward gauge for the relative importance of both types of error. If we use a conventional level of statistical significance ($\alpha = 0.05$) and statistical power (0.80, $\beta = 0.20$), it implies that the researcher is willing to accept a risk of making a Type II error that is four times greater than the risk of a Type I error:

$$\beta/\alpha = 0.20/0.05 = 4.0.$$

If the target level of statistical power is 0.90, then $\beta = 0.10$, and the ratio of probabilities decreases to $0.10/0.05 = 2.0$. What this means is that for a fixed level of statistical significance (α), increasing the level of statistical power reduces the chances of a Type II error (β) at the same time that the ratio of β/α moves closer to 1.0, where the chances of both types of error are viewed as equally important.

What happens if we reduce the desired level of statistical significance? For example, suppose we were particularly concerned about our chances of making a Type I error and reduced α from 0.05 to 0.01. For a statistical power level

---

[3] See S. E. Brown "Statistical Power and Criminal Justice Research," *Journal of Criminal Justice* 17 (1989): 115-122 and D. Weisburd "Design Sensitivity in Criminal Justice Experiments," *Crime and Justice* 17 (1991): 337-379.

of 0.80, this would imply that we are willing to accept a probability of making a Type II error that is 20 times greater than the probability of a Type I error. If we simultaneously increase the level of statistical power to 0.90 at the same time we reduce the significance level, the β/α ratio decreases to 10, but it still implies a much greater likelihood of a Type II error. If we wanted to keep the ratio of error probabilities at 4.0, we would need a study with a power level of 0.96 ($=1-4(\alpha)=1-0.04$). Intuitively, this makes good sense though: if we are going to make it more difficult to reject the null hypothesis by reducing α, we will simultaneously increase our chances of failing to reject a false null hypothesis unless we have a more powerful study.

Components of Statistical Power

The level of statistical power associated with any given test of a sample statistic is influenced by three key elements:

- Level of statistical significance, including directional tests when appropriate
- Sample size
- Effect size

The level of statistical significance and sample size are assumed to be within the control of the researcher, while the estimated effect size is not. The following discussion briefly describes the links between each element and the statistical power of any given test.

Statistical Significance and Statistical Power

The most straightforward way to increase the statistical power of a test is to change the significance level used. As we reduce the chances of making a Type I error by reducing the level of statistical significance from 0.10 to 0.05 to 0.01, it becomes increasingly difficult to reject the null hypothesis. Simultaneously, the power of the test is reduced. A significance level of 0.05 results in a more powerful test than a significance level of 0.01, because it is easier to reject the null hypothesis using the more lenient significance criteria. Conversely, a 0.10 level of significance would make it even easier to reject the null hypothesis.

As a simple illustration, Table 23.1 presents *z*-scores required to reject the null hypothesis for several levels of statistical significance using a two-tailed test. It would take a *z*-score greater than 1.645 or less than −1.645 to reject the null hypothesis with $p=0.10$, a *z*-score greater than 1.960 or less than −1.960 with $p=0.05$, and a *z*-score greater than 2.576 or less than −2.576 for $p=0.01$. Clearly,

Table 23.1 *z*-Scores Needed to Reject the Null Hypothesis in a Two-Tailed Test of Statistical Significance by Level of α

α	0.20	0.10	0.05	0.01	0.001
z-score	±1.282	±1.645	±1.960	2.576	3.291

Table 23.2 *z*-Scores Needed to Reject the Null Hypothesis in One-Tailed and Two-Tailed Tests of Statistical Significance

β	0.20	0.10	0.05	0.01	0.001
z-score (one-tail test)	–0.842 or 0.842	–1.282 or 1.282	–1.645 or 1.645	–2.326 or 2.326	–3.090 or 3.090
z-score (two-tail test)	±1.282	±1.645	±1.960	2.576	3.291

it is much easier to reject the null hypothesis with a 0.10 significance threshold than with a 0.01 significance threshold.

This method for increasing statistical power is direct, but it means that any benefit we gain in reducing the risk of a Type II error is offset by an increase in the risk of a Type I error. By setting a more lenient significance threshold, we do indeed gain a more statistically powerful research study. However, the level of statistical significance of our test also declines. Since a 0.05 significance level has become the convention in much of the research in criminology and criminal justice, it is important for authors to note why a more (or less) restrictive level of statistical significance is used.

Directional Hypotheses

A related method for increasing the statistical power of a study is to limit the direction of the research hypothesis to either a positive or a negative outcome, which implies the use of a one-tailed statistical test. A one-tailed test will provide greater statistical power than a two-tailed test for the same reason that a less stringent level of statistical significance provides more power than a more stringent one. By choosing a one-tailed test, the researcher reduces the absolute value of the test statistic needed to reject the null hypothesis by placing all of the probability of making a Type I error in a single tail of the distribution.

We can see this in practice again with the z-test. Table 23.2 lists the *z*-scores needed to reject the null hypothesis in one- and two-tailed tests for five different levels of statistical significance. (For the sake of simplicity, we assume in the one-tailed test that the outcome will be positive.) At each level, as in other statistical tests, the test statistic required to reject the null hypothesis is smaller in the case of a one-tailed test. For example, at $p = 0.05$, a *z*-score greater than or equal to 1.960 or less than or equal to −1.960 is needed to reject the null hypothesis in the two-tailed test. In the one-tailed test, the *z*-score needs only to be greater than or equal to 1.645. When we reduce the significance level to $p = 0.01$, a *z*-score greater than or equal to 2.576 or less than or equal to −2.576 is needed to reject the null hypothesis in the two-tailed test, but in the one-tailed test, the *z*-score needs only to be greater than or equal to 2.326.

Although the researcher can increase the statistical power of a study by using a directional, as opposed to a nondirectional, research hypothesis, there is a price for shifting the rejection region to one side of the sampling distribution. Once a one-directional test is defined, a finding in the direction opposite to that originally predicted cannot be recognized. To do otherwise would bring into question the integrity of the assumptions of the statistical test used in the analysis.

Sample Size and Statistical Power

The method used most often to change the level of statistical power in social science research is to vary the size of the sample. Similar to specifying the level of statistical significance, sample size can be controlled by the researcher. Modifying the size of the sample is typically a more attractive option for increasing statistical power than modifying the level of statistical significance, since the risk of a Type I error remains fixed—presumably at the conventional $p = 0.05$.

The relationship between statistical power and sample size is straightforward. All else being equal, larger samples provide more stable estimates of the population parameters than do smaller samples. Assuming that we are analyzing data from random samples of a population, the larger sample will have smaller standard errors of the coefficients than will the smaller sample. As the number of cases in a sample increases, the standard error of the sampling distribution (for any given statistical test) decreases. For example, we illustrated in Chapter 10 that the standard error for a single-sample *t*-test is computed as

$$\sigma_{se} = \frac{\sigma}{\sqrt{N-1}}.$$

As N gets larger, irrespective of the value of the standard deviation (σ) itself, the standard error of the estimate (σ_{se}) gets smaller. As the standard error of a test decreases, the likelihood of achieving statistical significance grows, because the test statistic for a test of statistical significance is calculated by taking the ratio of the difference between the observed statistic and the value proposed in the null hypothesis (typically 0) to the standard error of that difference. If the difference is held constant, then as the sample size increases, the standard error decreases, and a larger test statistic is computed, making it easier to reject the null hypothesis.

The effect of sample size on statistical power for a *t*-test of the difference of two independent sample means is illustrated in Table 23.3. The last column of Table 23.3 indicates the number of statistically significant outcomes expected in 100 two-sample *t*-tests in which a mean difference of two arrests between groups ($\sigma = 1$) is examined for four different scenarios (using a 5% significance threshold and a two-tailed test). In the first scenario, the sample size for each group is only 35 cases; in the second scenario, the sample size is 100; in the third, 200; and in the fourth, fully 1,000. Table 23.3 shows that the likelihood of rejecting the null hypothesis changes substantially with each increase in sample size, even though all

Table 23.3 Number of Statistically Significant Outcomes Expected in 100 Two-Sample *t*-Tests for Four Scenarios

SCENARIO	SAMPLE SIZE(PER GROUP)	μ1 – μ2	σ	EXPECTED SIGNIFICANT OUTCOMES
1	35	0.2	1	13
2	100	0.2	1	29
3	200	0.2	1	51
4	1,000	0.2	1	99

other characteristics are held constant across the four scenarios. Under the first scenario, we would expect only about 13 statistically significant outcomes in 100 tests. In the second scenario, 29 significant outcomes would be expected and in the third, 51. In the final scenario of samples of 1,000, nearly every test (99 out of 100) would be expected to lead to a significant result.

Sample size is often a primary concern in statistical power analysis because (1) it is directly related to statistical power, (2) it is a factor usually under the control of the researcher, and (3) it can be manipulated without altering the criteria for statistical significance of a study.

In most cases, researchers maximize the statistical power of a study by increasing sample size. The concern with sample size is also reflected in the number of publications focused on advising researchers in all behavioral and social science fields on how to determine the appropriate sample size for a proposed research study.[4]

Although sample size should be under the control of the researcher, it is important to be aware of the unanticipated consequences of simply increasing sample size may have on other factors that influence statistical power, particularly in evaluation research.[5] For example, suppose a researcher has developed a complex and intensive method for intervening with high-risk youth. The impact of the treatment is dependent on each subject receiving the "full dosage" of the treatment for a six-month period. If the researcher were to increase the sample size of this study, it might become more difficult to deliver the treatments in the way that was originally intended by the researcher. More generally, increasing the sample size of a study can decrease the integrity or the dosage of the interventions that are applied and result in the study showing no effect of the treatment. Increasing the size of a sample may also affect the variability of study estimates in other ways. For example, it may become more difficult to monitor implementation of treatments as a study grows. It is one thing to make sure that 100 subjects receive a certain intervention but quite another to ensure consistency of interventions across hundreds or thousands of subjects. Also, studies are likely to include more heterogeneous groups of subjects as sample size increases. For example, in a study of intensive probation, eligibility requirements were continually relaxed in order to meet project goals regarding the number of participants.[6] As noted earlier, as the heterogeneity of treatments or subjects in a study grows, it is likely that the standard deviations of the outcomes examined will also get larger. This, in turn, leads to a smaller effect size for the study and thus a lower level of statistical power.

---

[4] For a range of examples, see P. Dattalo *Determining Sample Size*, (New York: Oxford University Press, 2008); H. C. Kraemer and S. Thiemann *How Many Subjects: Statistical Power Analysis in Research*, (Newbury Park, CA: Sage, 1987); and, K. R. Murphy and B. Myors *Statistical Power Analysis*, 2 ed., (Mahwah, NJ: Lawrence Erlbaum, 2003).

[5] D. Weisburd "Design Sensitivity in Criminal Justice Experiments," *Crime and Justice* 17 (1991): 337-379.

[6] J. Petersilia "Randomized Experiments: Lessons from BJA's Intensive Supervision Project" *Evaluation Review* 13 (1989): 435-458.

Effect Size and Statistical Power

Effect size (ES) is a component of statistical power that is unrelated to the criteria for statistical significance used in a test. Effect size measures the difference between the actual parameters in the population and those hypothesized in the null hypothesis. In computing effect size, it is important to take into account both the raw differences between scores and the degree of variability found in the measures examined. Taking into account variability in effect size is a method of standardization that allows for the comparison of effects across studies that may have used different scales or slightly different types of measures. It has also allowed for the standardization of estimates of statistical power across a wide range of studies and types of analyses.

Generally, effect size is defined as

$$ES = \frac{Parameter - H_0}{\sigma}. \qquad \textbf{Equation 23.1}$$

The relationship between effect size and statistical power should be clear. When the standardized population parameters differ substantially from those proposed in the null hypothesis, the researcher should be more likely to observe a significant difference or effect in a particular sample. Effect size is dependent on two factors: (1) the difference between the actual parameter and the hypothesized parameter under the null hypothesis and (2) the variability (i.e., standard error) in the measure examined. Effect size will increase when the difference between the population parameter and the hypothesized parameter increases and the standard error is held constant or when the difference is held constant and the standard error is decreased, perhaps through the use of a larger sample of cases.[7]

A difference of means test for two independent samples provides a simple illustration for these relationships. In the difference of means test, effect size would be calculated by first subtracting the population difference as stated in the null hypothesis ($H_0\mu_1 - H_0\mu_2$) from the difference between the true means in the population ($\mu_1 - \mu_2$). When comparing these two populations, variability is defined as the pooled or the common standard deviation of the outcome measures in the two populations (σ). Consequently, ES would be computed as

$$ES = \frac{(\mu_1 - \mu_2) - (H_0\mu_1 - H_0\mu_2)}{\sigma}. \qquad \textbf{Equation 23.2}$$

---

[7] Effect size can also be calculated for observed differences in a study. This is a common approach in meta-analysis, where a large group of studies are summarized in a single analysis. For example, in calculating effect size for a randomized experiment with one treatment and one control group, the researcher would substitute the outcome scores for both groups in the numerator of the ES equation and the pooled standard deviation for the two outcome measures in the denominator. For a more detailed discussion of effect size and its use generally for comparing effects across different studies, see M. Lipsey and D. Wilson *Practical Meta-Analysis*, (Thousand Oaks, CA: Sage, 2001), and R. Rosenthal *Meta-Analytic Procedures for Social Research*, (Beverly Hills: Sage, 1984).

Table 23.4 Number of Statistically Significant Outcomes Expected in 100 Two-Sample *t*-Tests for Six Different Scenarios (100 Cases in Each Sample)

SCENARIO	μ1	μ2	σ	EXPECTED SIGNIFICANT OUTCOMES
(a) Means differ; standard deviations constant				
1	0.3	0.5	2	10
2	0.3	0.9	2	56
3	0.3	1.3	2	94
(b) Means constant; standard deviations differ				
4	0.3	0.5	0.5	80
5	0.3	0.5	1	29
6	0.3	0.5	2	10

Since the null hypothesis for a difference of means test is ordinarily that the two population means are equal (i.e., $H_0\mu_1 - H_0\mu_2 = 0$), we can simplify this formula and include only the difference between the actual population parameters:

$$ES = \frac{(\mu_1 - \mu_2)}{\sigma}. \qquad \textbf{Equation 23.3}$$

Thus, the ES for a difference of means test may be defined simply as the raw difference between the two population parameters, divided by their common standard deviation. To reiterate an earlier comment, when the difference between the population means is greater, the ES for the difference of means will be larger. Also, as the variability of the scores of the parameters grows, as represented by the standard deviation of the estimates, the ES will get smaller.

Table 23.4 presents a simple illustration of the relationship between effect size and statistical power in practice. The last column of Table 23.4 presents the number of statistically significant outcomes expected in 100 *t*-tests (using a 0.05 significance threshold and a nondirectional research hypothesis, resulting in a two-tail test), each with 100 cases per sample, and illustrated for six different scenarios. In the first three scenarios, the mean differences between the two populations are varied and the standard deviations for the populations are held constant. In the last three scenarios, the mean differences are held constant and the standard deviations differ.

As Table 23.4 shows, the largest number of statistically significant outcomes is expected in either the comparisons with the largest differences between mean scores or the comparisons with the smallest standard deviations. As the differences between the population means grow (scenarios 1, 2, and 3), so too does the likelihood of obtaining a statistically significant result. Conversely, as the population standard deviations of the comparisons get larger (scenarios 4, 5, and 6), the expected number of significant outcomes decreases.

As this exercise illustrates, there is a direct relationship between the two components of effect size and statistical power. Studies that examine populations in which there is a larger effect size will, all else being equal, have a higher level of statistical power. Importantly, the relationship between effect size and statistical power is unrelated to the significance criteria we use in a test. In this sense, effect

size allows for increasing the statistical power of a study (and thus reducing the risk of Type II error) while minimizing the risk of Type I error (through the establishment of rigorous levels of statistical significance).

Although effect size is often considered the most important component of statistical power, it is generally very difficult for the researcher to manipulate in a specific study.[8] Ordinarily, a study is initiated in order to determine the type and magnitude of a relationship that exists in a population. In many cases, the researcher has no influence at all over the raw differences or the variability of the scores on the measures examined. For example, a researcher who is interested in identifying whether male and female police officers have different attitudes toward corruption may have no idea prior to the execution of a study the nature of these attitudes or their variability. It is then not possible for the researcher to estimate the nature of the effect size prior to collecting and analyzing data—the effect size may be large or small, but it is not a factor that the researcher is able to influence.

In contrast, evaluation research—in which a study attempts to assess a specific program or intervention—the researcher may have the ability to influence the effect size of a study and thus minimize the risk of making a Type II error. There is growing recognition, for example, of the importance of ensuring the strength and integrity of criminal justice interventions.[9] Moreover, many criminal justice evaluations fail to show a statistically significant result simply because the interventions are too weak to have the desired impact or the outcomes are too variable to allow a statistically significant finding.[10]

Statistical power suggests that researchers should be concerned with the effect size of their evaluation studies if they want to develop a fair test of the research hypothesis. First, the interventions should be strong enough to lead to the expected differences in the populations under study. Of course, the larger the differences expected, the greater the statistical power of an investigation. Second, interventions should be administered in ways that maximize the homogeneity of outcomes. For example, interventions applied differently to each subject will likely increase the variability of outcomes and thus the standard deviation of those scores. Finally, researchers should recognize that the heterogeneity of the subjects studied (and thus the heterogeneity of the populations to which they infer) will often influence the statistical power of their tests. Different types of people are likely to respond in different ways to treatment or interventions. If they do respond differently, the variability of outcomes will be larger, and thus the likelihood of making a Type II error will increase.

As a caution, we note that a wide range of research in criminology and criminal justice has increasingly made use of archival data sets that result in researchers analyzing populations rather than samples. Examples of this would include studies that rely on archival data on all punishment decisions made in the US

---

[8] M. Lipsey *Design Sensitivity: Statistical Power for Experimental Research*, (Newbury Park, CA: Sage, 1990).

[9] J. Petersilia "Randomized Experiments: Lessons from BJA's Intensive Supervision Project" *Evaluation Review* 13 (1989): 435-458.

[10] D. Weisburd "Design Sensitivity in Criminal Justice Experiments," *Crime and Justice* 17 (1991): 337-379.

Federal District Courts or census data on all prisoners in a state on a specific date. A number of years ago, Maltz (1994) noted that the increased frequency with which populations are analyzed calls into question many of the assumptions about performing tests for statistical significance.[11] Put simply, the analysis of population data implies no need for statistical significance testing, since the researcher is not trying to generalize from a sample to a population. Clearly, issues of statistical power are not relevant when we analyze a population: setting a significance level makes little sense, the number of cases in the data set is as large as it possibly can be, and the effect size is simply what is observed (excluding measurement error).

Estimating Statistical Power and Sample Size for a Statistically Powerful Study

A number of texts have been written that provide detailed tables for defining the statistical power of a study.[12] All of these texts also provide a means for computing the size of the sample needed to achieve a given level of statistical power. In both cases—the estimation of statistical power or the estimation of necessary sample size—assumptions will need to be made about effect size and level of statistical significance desired. The following discussion provides a basic illustration for how to compute estimates of statistical power. (The computations reported in the following discussion have been performed with a variety of statistical software tools, several of which are freely available. More detail on several easily accessible resources to compute power estimates is provided in the computer problems section at the end of this chapter.)

The most common application of statistical power analysis in criminology and criminal justice research has been to compute the sample size needed to achieve a statistically powerful study (generally at or above 80%). As noted above, we need to be cautious about simply increasing the size of the sample, since a larger sample can affect other important features of statistical power. Thus, in using increased sample size to minimize Type II error, we must consider the potential consequences that larger samples might have on the nature of interventions or subjects studied, particularly in evaluation research. Nonetheless, sample size remains the tool most frequently used for adjusting the power of studies, because it can be manipulated by the researcher and does not require changes in the significance criteria of a test.

To define how many cases should be included in a study, we must conduct power analyses before the study is begun, generally referred to as prospective or

[11] M. D. Maltz "Deviating from the Mean: The Declining Significance of Significance," *Journal of Research in Crime and Delinquency* 31 (1994): 434-463.

[12] Among some of the more widely used examples are J. Cohen *Statistical Power Analysis for the Behavioral Sciences*, 2 ed., (Hillsdale, NJ: Lawrence Erlbaum, 1988); H. C. Kraemer and S. Thiemann *How Many Subjects: Statistical Power Analysis in Research*, (Newbury Park, CA: Sage, 1987); M. Lipsey *Design Sensitivity: Statistical Power for Experimental Research*, (Newbury Park, CA: Sage, 1990); and, K. R. Murphy and B. Myors *Statistical Power Analysis*, 2 ed., (Mahwah, NJ: Lawrence Erlbaum, 2003).

a priori power analysis, and where our attention has been focused thus far in this chapter. Some authors have advocated the use of power analysis to evaluate whether studies already conducted have acceptable levels of statistical power, based on the sample statistics, referred to as retrospective or **post hoc** power analysis. Although there is much agreement about the utility of prospective power analysis, there is little consensus about the appropriateness of retrospective power analysis.[13] The widespread use of secondary data sources in the study of crime and criminal justice further complicates the interpretation of results from a statistical power analysis. Since it is not possible for researchers to augment the original study's sample, results from a power analysis will still be informative in the sense that the results will indicate to the researchers using these data sources what the archived data set can and cannot tell them about the statistical relationships they may be most interested in.

To define the sample size needed for a powerful study, we must first clearly define each of the components of statistical power other than sample size. These include:

1. The statistical test
2. The significance level
3. The research hypothesis (whether directional or nondirectional)
4. The effect size

The first three of these elements should be familiar, since they are based on common assumptions made in developing any statistical test. The statistical test is chosen based on the type of measurement and the extent to which the study can meet certain assumptions. For example, if we want to compare three sample means, we will likely use analysis of variance as our test. If we are comparing means from two samples, we will likely use a two-sample *t*-test. If we are interested in the unique effects of a number of independent variables on a single interval-level dependent variable, we will likely use OLS regression and rely on *t*-tests for the individual coefficients and *F*-tests for either the full regression model or a subset of variables from the full model.

To calculate statistical power, we must also define the significance level of a test and its research hypothesis. By convention, we generally use a 0.05 significance threshold, and thus we are likely to compute statistical power estimates based on this criterion. The research hypothesis defines whether a test is directional or nondirectional. When the statistical test allows for it, we will typically choose a nondirectional test to take into account the different types of outcomes that can be found in a study.[14] If we were evaluating an existing study, we would use the decisions as stated by the authors in assessing that study's level of statistical power.

---

[13] For an example, see the exchange between J. P. Hayes and R. J. Steidl "Statistical Power Analysis and Amphibian Population Trends," *Conservation Biology* 11 (1997): 273-275 and L. Thomas "Retrospective Power Analysis," *Conservation Biology* 11 (1997): 276-280.

[14] J. Cohen *Statistical Power Analysis for the Behavioral Sciences*, 2 ed., (Hillsdale, NJ: Lawrence Erlbaum, 1988).

The fourth element, defining effect size, is perhaps the most difficult component. If we are trying to estimate the magnitude of a relationship in the population that has not been well examined in the past, how can we estimate the effect size in the population? It may be useful to reframe this criterion. The purpose of a power analysis is to see whether our study is likely to detect an effect of a certain size. Usually, we define that effect in terms of what is a meaningful outcome in a study. A power analysis, then, tells us whether our study is designed in a way that is likely to detect that outcome (i.e., reject the null hypothesis on the basis of our sample statistics). This is one of the reasons why statistical power is sometimes defined as design sensitivity.[15] It assesses whether our study is designed with enough sensitivity to be likely to reject the null hypothesis if an effect of a certain size exists in the population under study.

The task of defining effect size has been made easier by identifying broad categories of effect size that can be compared across studies. Cohen's suggestions have been the most widely adopted by other researchers and simply refer to classifying effect sizes as small, medium, and large.[16] The numeric value associated with an effect size classified as small, medium, or large is contingent on the specific statistical test being considered. For example, if our focus is on a difference of means test for two independent samples, the standardized effect size estimate is known as d (explained below) and is considered to be a small effect if it is 0.2, a medium effect if it is 0.5, and a large effect if it is 0.8. In contrast, if we are considering the statistical power of an OLS regression model, the standardized effect size estimate is known as f^2 and is considered to be a small effect if it is 0.02, a medium effect if it is 0.15, and a large effect if it is 0.35. Other authors have followed suit and attempted to define similar types of standardized effects for more complex statistical models not addressed in Cohen's work or this book.

The following illustration turns to a discussion of the computation of statistical power for several common situations in criminology and criminal justice research: difference of means test, ANOVA, correlation, and OLS regression—all of which have been the focus of previous chapters.

The computation of statistical power estimates requires the comparison of a sampling distribution under the null hypothesis with a sampling distribution under the alternative or the research hypothesis (see again Figure 23.1, above). The sampling distribution under the research hypothesis is referred to as a non-central distribution. Recall from our discussion of Figure 23.1 that the sampling distribution under the null hypothesis is the t-distribution, while the sampling distribution under the research hypothesis is the non-central t-distribution.

The non-central sampling distribution is computed based on a "non-centrality" parameter, which in all cases is a function of the standardized effect for the statistical test under consideration. For each of the statistical tests discussed below, we describe both the standardized effect and the non-centrality parameter and explain

---

[15] M. Lipsey *Design Sensitivity: Statistical Power for Experimental Research*, (Newbury Park, CA: Sage, 1990).

[16] J. Cohen *Statistical Power Analysis for the Behavioral Sciences*, 2 ed., (Hillsdale, NJ: Lawrence Erlbaum, 1988).

how to use these values to estimate the statistical power of a sample as well as the size of sample needed to meet a target level of statistical power.

Difference of Means Test

Throughout this chapter, we have pointed to the difference of means test as an example for many of the points we wanted to make about statistical power. More directly, the standardized effect size d is

$$d = \frac{\mu_1 - \mu_2}{\sigma},$$

which is identical to the equation noted earlier for computing a standardized difference of means for two independent samples. Recall that σ represents the pooled, or common, standard deviation for the difference of means.

The non-centrality parameter δ for the t-distribution is

$$\delta = d\sqrt{\frac{N}{4}}, \qquad \textbf{Equation 23.4}$$

where $N = n_1 + n_2$ when there are equal numbers of cases in each group (i.e., $n_1 = n_2$). For the situation where $n_1 \neq n_2$, the non-centrality parameter δ is

$$\delta = d\sqrt{\frac{N_H}{2}}, \text{ where } N_H = \frac{2n_1 n_2}{n_1 + n_2}. \qquad \textbf{Equation 23.5}$$

To illustrate the computation of a statistical power estimate, suppose that we want to assess the effectiveness of a treatment program for drug offenders. Our design calls for random assignment of 100 cases to each group. We expect the program to be effective at reducing recidivism in the treatment group and so can assume a one-tailed t-test with a significance level of 5%. What is the statistical power of our design for detecting standardized effects at the small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) levels?

For all three scenarios, the critical t-value will be 1.653, based on a one-tailed test with a significance level of 0.05 and $df = N - 2 = 198$. For a small effect, the non-centrality parameter δ is 1.414 ($=0.2*\sqrt{(200/4)}$). This provides us with an estimate for risk of making a Type II error of $\beta = 0.593$, suggesting that we have a probability of 59.3% of making a Type II error and fail to reject the null hypothesis when it is false.[17] The corresponding estimate of statistical power is $1 - 0.593 = 0.407$. Substantively, this result suggests that if we have only 100 cases in each group, our probability of rejecting the null hypothesis when it is false is only about 40.7%. In regard to a medium effect size, $\delta = 3.536$, $\beta = 0.030$, and power = 0.970. For a large effect size, $\delta = 5.657$, $\beta < 0.0001$, and power > 0.9999.

[17] It is not possible to include copies of non-central t-distribution tables in the same way that we have for the Student's t-distribution in Appendix 4. We will illustrate in the Computer Exercises at the end of this chapter how to work with the non-centrality parameter to obtain estimates of β from various statistical packages.

Putting these results together indicates that our design with 100 cases assigned to each group provides a high level of statistical power for detecting medium effects and larger but an inadequate level of power for detecting small effects.

Alternatively, we may be interested in determining the sample size needed to provide us with a statistical power estimate of 80% for each of the three effect sizes: small, medium, and large. In the case of a small effect, we find that we need a total sample of 620 cases—310 in each group—to assure us that we will be able to reject the null hypothesis when it is false about 80% of the time. To achieve a power estimate of 80% for a medium effect, we only need 102 cases (51 in each group). For a large effect, the sample size drops to 40 (20 in each group).

ANOVA

For a simple ANOVA, where we are looking only at fixed effects and assume equal sample sizes across groups, the standardized effect size *f* is defined as

$$f = \frac{\sigma_m}{\sigma}, \qquad \textbf{Equation 23.6}$$

where $\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(m_i - m)^2}{k}}$, k is the number of groups, m is the grand mean, and m_i represents each of the group means with $n_1 = n_2 = \ldots = n_k$.

The non-centrality parameter λ for the *F*-distribution is

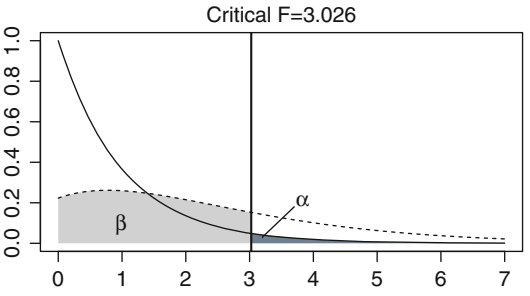
$$\lambda = f^2 N, \qquad \textbf{Equation 23.7}$$

where f^2 refers to the square of the standardized effect size (f) and N refers to the total sample size.

As an illustration of the calculation of statistical power estimates for a fixed-effects ANOVA model, assume that we have three groups, each with 100 cases participating in an experiment aimed at reducing recidivism among violent offenders: a control group and two different kinds of treatment groups. Assume that the significance level has been set at 5%. What is the level of statistical power of our design for detecting standardized effects at the small ($f = 0.1$), medium ($f = 0.25$), and large ($f = 0.4$) levels?

For each of the three scenarios, the critical value of the *F*-statistic is 3.026 ($df_1 = 2$, $df_2 = 297$). For a small effect, the non-centrality parameter λ is 3 ($= 0.12 * 300$). This provides us with an estimate for risk of making a Type II error of $\beta = 0.681$, suggesting that we have a probability of 68.1% of making a Type II error and fail to reject the null hypothesis when it is false. The corresponding estimate of statistical power is $1 - 0.681 = 0.319$, meaning that we have only a 31.9% chance of rejecting the null hypothesis when it is false. This result is presented graphically in Figure 23.3. Similar to the layout in Figure 23.1, the vertical line indicates the critical F of 3.026, the solid line the *F*-distribution, and the dashed line the non-central *F*-distribution. Below the two curves, represented by two different shades of grey, alpha is indicated by the darker shading in the right tail of

Figure 23.3 *Graphical Representation for Power Analysis in a One-Way ANOVA*



the *F*-distribution beyond the critical value, and beta is represented by the lighter shaded area to the left of the critical value and under the non-central *F*-distribution.

For the medium and large effect size analyses, the *F*-distribution remains the same, but the non-central *F*-distribution is shifted further to the right. For the medium effect size, $\lambda = 18.75$, $\beta = 0.022$, and power = 0.978. The large effect size has $\lambda = 48$, $\beta < 0.0001$, and power > 0.9999. Similar to the previous analysis comparing the means for only two groups, our research design with 100 cases assigned to each of the three groups provides a high level of statistical power for detecting medium and large effects but an inadequate level of power for detecting small effects.

If our concern is focused on the size of the sample needed for a power level of 80% for each of the three effect sizes—small, medium, and large—then we would again proceed in the same way as in the two-sample *t*-test. To have an 80% chance of detecting a small effect ($f = 0.10$), we would need a sample of 969 cases (323 in each group). For the medium effect, we would need only 159 cases (53 in each group) and for the large effect, only 66 cases (22 in each group).

Correlation

To test the statistical power of a correlation coefficient, we can use either the correlation coefficient (r) or the Fisher r-to-Z transformation of the correlation coefficient (r_Z) as the standardized effect size. Although the estimates of statistical power will not be identical, they will tend to be very close, typically differing only at the second or the third decimal.

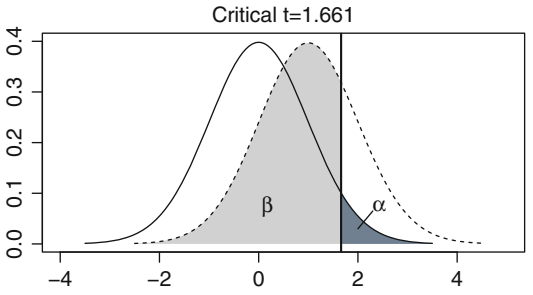
The non-centrality parameter δ for the correlation coefficient is

$$\delta = \sqrt{\frac{r^2}{1-r^2} \times N}, \qquad \textbf{Equation 23.8}$$

where r is either the sample correlation coefficient (r) or the Fisher transformed (r_Z) and N is the sample size.

We can again illustrate the calculation of statistical power for correlations by assuming that we have 100 observations that would allow us to compute a correlation between two variables. For example, suppose we interview a random sample of police officers and are interested in the correlation between the number of

Figure 23.4 *Graphical Representation for Power Analysis of a Correlation*



years on the police force and a scale that measured hostility toward judges. We might expect that more years on the police force will have a positive correlation with hostility toward judges, implying that we can conduct a one-tailed *t*-test of statistical significance. As with the preceding examples, assume that the level of statistical significance is 5%. What is the level of statistical power of our design for detecting standardized effects at the small ($r = 0.1$), medium ($r = 0.3$), and large ($r = 0.5$) levels?

The critical *t*-value for all three scenarios is 1.661, based on $df = N - 2 = 98$. For a small effect size ($r = 0.1$), the non-centrality parameter is $\delta = 1.005$. This provides us with an estimate for risk of making a Type II error of $\beta = 0.741$, suggesting that we have a probability of 74.1% of making a Type II error and would fail to reject the null hypothesis when it is false. The corresponding estimate of statistical power is 0.259, indicating that we would only reject the null hypothesis when it was false about 26% of the time. Figure 23.4 presents these results graphically. The statistical power analysis of the medium effect indicates that $\delta = 3.145$, $\beta = 0.070$, and power = 0.930. The large effect shows an even greater level of statistical power, where $\delta = 5.774$, $\beta < 0.0001$, and power > 0.9999.

The sample size required to detect each of the three effect sizes—small, medium, and large—with a statistical power of 80% again requires the use of the *t*-distribution. To achieve a power level of 80% for a small effect ($r = 0.1$), a sample of 614 cases would be needed. For the medium effect ($r = 0.3$), the required number of cases drops to 64, while for the large effect ($r = 0.5$), only 21 cases are required to have an 80% chance of rejecting the null hypothesis when it is false.

Least-Squares Regression

The statistical power analysis of least-squares regression can take two different, but related, forms. One question asks about the ability to detect whether a regression model—a single dependent variable and two or more independent variables—has a statistically significant effect on the dependent variable. This means that the null hypothesis is focused on whether the regression model in its entirety has an effect on the dependent variable. A second question asks about the ability to detect the effect of a single variable or a subset of variables added to a regression model. This addresses the more common substantive question in much of the published

research: Once the other relevant independent and control variables have been taken into account statistically, does variable *X* add anything to the overall model?

Whether we are analyzing the full model or a subset of the full model, the standardized effect size (denoted as f^2) is based on either the R^2 for the full model or the partial R^2 for the subset of variables we are interested in analyzing. Specifically,

$$R^2 = \frac{f^2}{1+f^2}. \qquad \textbf{Equation 23.9}$$

As indicated above, Cohen's recommendations for small, medium, and large standardized effect sizes in a regression model are 0.02, 0.15, and 0.35, respectively. To provide some context to these values, an f^2 value of 0.02 corresponds to an R^2 of 0.02, while $f^2 = 0.15$ implies that $R^2 = 0.13$, and $f^2 = 0.35$ implies that $R^2 = 0.26$. Statistical power analysis for least-squares regression uses the *F*-distribution.

As noted in the discussion of statistical power analysis for ANOVA models, the non-centrality parameter λ for the *F*-distribution is

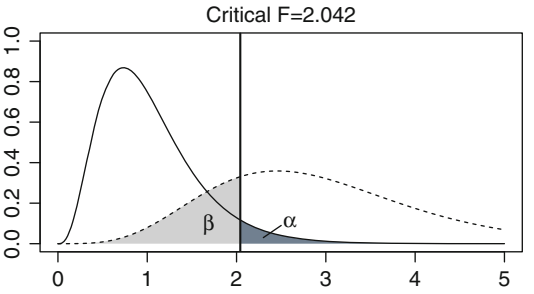
$$\lambda = f^2 N.$$

To assess the statistical power for the full regression model consider the following simple example. Suppose that we are interested in the effects of various case and defendant characteristics on the amount of bail required by a court. Typical analyses of bail decisions would consider some of the following characteristics (as well as others not listed): (1) severity of the prior record, (2) severity of the current offense, (3) number of counts of the current offense, (4) type of attorney, (5) whether the defendant was under criminal justice supervision at the time of the current offense, (6) age of the defendant, (7) race of the defendant, and (8) gender of the defendant. This provides us with a regression model with eight independent variables.

As a point of illustration, we may want to estimate the statistical power of the regression model assuming that we have a sample of only 100 cases and have set a significance level of 5%, giving us a crucial *F*-value of 2.024. For the small effect size ($f^2 = 0.02$), we have non-centrality parameter $\lambda = 2.0$ (=0.02*100). We then find $\beta = 0.876$, meaning that with only 100 cases, we have a probability of making a Type II error of just under 88%. Alternatively, the estimate of statistical power is 0.124, meaning that we have a probability of only 12.4% of rejecting the null hypothesis when it is false. The results for the medium effect ($f^2 = 0.15$) appear in Figure 23.5, which is based on $\lambda = 15.0$, $\beta = 0.242$, and power = 0.758. This is still an inadequate level of power but is much closer to the target of 80%. For the large effect ($f^2 = 0.35$), $\lambda = 35.0$, $\beta = 0.007$, and power = 0.993, which is well beyond the desired level of 80%.

For a regression model with eight independent variables, what sample size is required to achieve a statistical power level of 80% for detecting effects at the small ($f^2 = 0.02$), medium ($f^2 = 0.15$), and large ($f^2 = 0.35$) levels? For the small effect, we would require a sample of 759 cases to achieve a power level of 80%. For the medium and large effects, we would require samples of 109 and 52 cases, respectively.

Figure 23.5 *Graphical Representation for Power Analysis of a Regression Model (with Eight Independent Variables)*



The number of cases required to detect a statistically significant effect at either the medium or the large effect level may strike many readers as small. It is important to keep in mind that we have only been assessing the full model—the number of cases required for detecting individual effects will tend to be different than the number of cases required for detecting whether the full model is significant.

The assessment of statistical power for a single independent variable or a small subset of independent variables proceeds in much the same way as the analysis for the full model. The key difference is in the degrees of freedom required for the *F*-distribution. In the case of a single independent variable, the numerator $df = 1$, while the denominator *df* remains the same as in the full model. For a subset of independent variables, the numerator *df* = the number of variables in the subset (the denominator *df* remains the same).

If we return to the bail example above, the analysis of statistical power for any one of the independent variables will be identical. We continue to keep the sample size at 100 cases, the level of statistical significance at 5%, and the definition of small, medium, and large effects the same as before. For the small effect ($f^2 = 0.02$), $\lambda = 2.0$, $\beta = 0.712$, and power = 0.288, meaning that we would only be able to reject the null hypothesis of no relationship between the independent and dependent variables about 28.8% of the time. For the medium effect ($f^2 = 0.15$), $\lambda = 15.0$, $\beta = 0.031$, and power = 0.969, while for the large effect ($f^2 = 0.35$), $\lambda = 35.0$, $\beta < 0.0001$, and power > 0.9999.

Similarly, we may be interested in assessing the statistical power of a subset of variables. For example, in the bail example, the subset of demographic characteristics (age, race, and gender) may be important in testing some aspect of a theory predicting differential treatment of defendants within the courts. We find a similar pattern to the results. For the small effect ($f^2 = 0.02$), $\lambda = 2.0$, $\beta = 0.814$, and power = 0.186, again indicating a low level of statistical power for detecting a statistically significant relationship between demographic characteristics and bail amount. For the medium effect ($f^2 = 0.15$), $\lambda = 15.0$, $\beta = 0.095$, and power = 0.905, while for the large effect ($f^2 = 0.35$), $\lambda = 35.0$, $\beta = 0.001$, and power = 0.999.

Sample size calculations work in the same way as for the full model. If we hope to achieve a power level of 80%, what size sample is necessary to detect

small, medium, and large effects for either single variables or subsets of variables? Continuing the bail example, we assume that there are eight independent variables. For the single variable, the number of cases required to detect a small effect with a probability of 80% is 395. A medium effect requires only 55 cases, while a large effect requires only 26 cases. It is worth noting that sample size calculations for single variable effects are not affected by the number of variables included in the full regression model.

In practice, many of the individual effects that researchers are trying to assess in their multivariate models will tend toward the small effect size. For example, much survey research aimed at trying to explain attitudes toward a particular topic will often incorporate 10–20 independent variables and have a full model R^2 typically between 0.15 and 0.20. This implies that many of the effects of individual variables will tend to be quite small in magnitude. In order for an analysis to detect a statistically significant relationship, a much large sample becomes necessary.

Summing Up: Avoiding Studies Designed for Failure

The statistical power of a test can be compared to the sensitivity of a radiation meter. A very sensitive meter will be able to identify even the smallest deposits of radioactivity. A meter that is not very sensitive will often miss such small deposits, although it likely will detect very large radiation signals from areas rich in radioactivity. Similarly, a statistically sensitive study will be able to identify even small effects. This is usually because the researcher has increased the sample size of the study to make it more statistically powerful. Conversely, a study that has little sensitivity is unlikely to yield a statistically significant result even when relatively large differences or program impacts are observed. Such studies may be seen as "designed for failure," not because of inadequacies in the theories or the programs evaluated, but because the investigator failed to consider statistical power at the outset of the study.

You might question why we would even bother to define the size of the sample needed for statistically powerful studies. Why not just collect 1,000 or more cases in every study and be almost assured of a statistically powerful result? The simple answer is that although you should try to sample as many cases as you can in a study, there are generally constraints in developing samples. These constraints may be monetary, related to time, or associated with access to subjects. It is often important to know the minimum number of cases needed to achieve a certain threshold of statistical power so that you can try, within the constraints of the research setting, to reach an adequate level of statistical power in your study. It is also important to be able to assess whether studies that you read or evaluate were designed in such a way that they are reasonable tests of the hypotheses presented. If such studies are strongly underpowered, then you should have much less confidence in findings that do not support the research hypothesis.

Chapter Summary

A statistically powerful test is one for which there is a low risk of making a Type II error. Statistical power can be defined as 1 minus the probability of falsely accepting the null hypothesis. A test with a statistical power of 0.90 is one for which there is only a 10% probability of making a Type II error. If the power of a test is 0.10, the probability of a Type II error is 90%. A minimum statistical power level of at least 0.50 is recommended. However, it is generally accepted that in better studies, the level of statistical power will be at least 0.80. A study with a low level of statistical power can be described as "designed for failure," as it is unlikely to produce a statistically significant result even if the expected effect exists in the population under study.

There are several ways in which statistical power can be maximized. First, we may raise the significance threshold. Doing so, however, also increases the risk of a Type I error. Second, we may limit the direction of the research hypothesis and conduct a one-tailed test. Doing so, though, will necessarily ignore outcomes in the opposite direction. Third, we may try to maximize the effect size. The greater the differences betwen the populations and the smaller the variability of those differences, the larger the population effect size will be. Effect size, however, is usually beyond the control of the researcher. Fourth, we may increase the sample size. A larger sample produces a smaller standard error for the sampling distribution and a larger test statistic. The larger the sample, all else being equal, the greater the chance of rejecting the null hypothesis.

Sample size is generally the most useful tool for maximizing statistical power. A power analysis before a study is begun will define the number of cases needed to identify a particular size effect—small, medium, or large. A power analysis of an existing study will help to identify whether it was well designed to assess the questions that were examined. To identify a small effect size, the overall sample must be very large. For a large effect size, a much smaller sample will suffice.

Key Terms

design sensitivity The statistical power of a research study. In a sensitive study design, statistical power will be maximized, and the statistical test employed will be more capable of identifying an effect.

effect size (ES) A standardized measure derived by taking the effect size (e.g., the difference between two populations), measured in the raw units of the outcome measure examined, and dividing it by the pooled or common standard deviation of the outcome measure.

statistical power One minus the probability of a Type II error. The greater the statistical power of a test, the less chance there is that a researcher will mistakenly fail to reject the null hypothesis.

Symbols and Formulas

ES: Effect Size
D: Standardized effect size
F: Standardized effect size in ANOVA and OLS regression
N: Sample size
n_i: Number of cases in group *i*
δ: Non-centrality parameter for the *t*-distribution
λ: Non-centrality parameter for the *F*-distribution (used for ANOVA and OLS regression power estimates)

To calculate effect size:

$$ES = \frac{Parameter - H_0}{\sigma}.$$

To calculate the effect size for a difference of means test:

$$ES = \frac{(\mu_1 - \mu_2) - (H_0\mu_1 - H_0\mu_2)}{\sigma}, \text{ which simplifies to } ES = \frac{(\mu_1 - \mu_2)}{\sigma}.$$

To calculate the non-centrality parameter for the *t*-distribution for difference of means:

When groups have same number of cases:

$$\delta = d\sqrt{\frac{N}{4}}.$$

When groups have different number of cases:

$$\delta = d\sqrt{\frac{N_H}{2}}, \text{ where } N_H = \frac{2n_1n_2}{n_1 + n_2}$$

To calculate the standardized effect size for an ANOVA:

$f = \frac{\sigma_m}{\sigma}$, where $\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(m_i - m)^2}{k}}$, k is the number of groups, m is the grand mean, and m_i represents each of the group means with $n_1 = n_2 = \cdots = n_k$.

To calculate the non-centrality parameter λ for the *F*-distribution:

$$\lambda = f^2N.$$

To calculate the non-centrality parameter for the *t*-distribution for correlation coefficient:

$$\delta = \sqrt{\frac{r^2}{1 - r^2} \times N}.$$

To calculate R^2 for OLS regression using the standardized effect size f as defined above for a subset of independent variables (one or more) in OLS regression:

$$R^2 = f^2\left(1+f^2\right).$$

Computer Exercises

In contrast to many of the other computer exercises in this text, the computation of statistical power estimates is not easily performed in any of the large stand-alone statistical packages. There are a variety of software packages available for computing statistical power as well as a number of websites that host power calculators for a wide range of statistical tests. All of the analyses presented in this chapter were performed with G*Power (version 3.1.7). G*Power 3 is freely available to download from the Institut fur Experimentelle Psychologie at Universitat Dusseldorf (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). G*Power 3 is a specialized package devoted to statistical power estimation and offers a wide range of tests beyond those discussed here. G*Power 3 also features the simple creation of powerful graphs that will plot power estimates across a range of sample sizes, effect sizes, and statistical significance levels. The figures presented in this chapter are similar to what are produced with G*Power 3. Faul et al. provide a useful overview of the capabilities of G*Power 3[18].

Power and Precision v. 2.0 is a commercially available software package designed to compute power estimates for a wide range of statistical models in a user-friendly environment.[19] As a commercial software package, its range of capabilities is significantly greater than G*Power 3. A particularly useful feature is that all of the output—text and graphs—can be easily exported to other programs.

In the case that one simply wants to compute a small number of power estimates without bothering to learn a new software package, a reasonably comprehensive list of Web-based power calculators can be found at http://statpages.org/#Power. The list of websites hosting power calculators is categorized by the type of statistical test that the user is searching for—one-sample *t*-test, two-sample *t*-test, correlation, regression, and so on.

On a technical note, it is worth highlighting that there will be slight differences across statistical software packages and power calculators in the estimated sample sizes needed to achieve a given level of statistical power. The primary reason for this appears to be focused on rounding the estimated sample size to an integer, since we cannot sample a fraction of a case in any research study. Some packages round up so that the estimated statistical power as always at least as great as the target entered into the computation. Other packages and calculators will round to the closest integer (regardless of whether it is larger or smaller), so the overall estimate of statistical power may be slightly less than the initial target.

---

[18] F. Faul, E. Erdfelder, A. Land and A. Buchner "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences," *Behavior Research Methods* 39 (2007): 175-191.

[19] M. Borenstein, H. Rothstein, and J. Cohen *Power and Precision*, (Englewood, NJ: Biostat, Inc., 2001).

Stata

Two-Sample Difference of Means Test

In Stata, one- and two-sample difference of means tests are performed with the **sampsi** command:

sampsi Mean1 Mean2 , **sd1(#) sd2(#) n1(#) n2(#) power(#) onesided**

where Mean1 and Mean2 refer to the expected population means for the two samples being compared, **sd1(#)** and **sd2(#)** refer to the expected standard deviations for each sample (values inserted in the parentheses), **n1(#)** and **n2(#)** refer to the expected number of cases (values inserted in the parentheses) in each sample, **power(#)** is a designated level of power for sample size estimation (the default is a power level of 0.9), and **onesided** indicates that a one-tail test should be used (a two-tail test is the default). In the situation where we are trying to estimate power and assume constant standard deviations and sample sizes across the two samples, this can be simplified to

sampsi Mean1 Mean2, **sd(#) n(#)**

Upon entering the command, the output will list all of the assumptions (alpha level, etc.) and then compute the power. For example, the first line of Table 23.3 indicates that the expected number of significant results (out of 100 tests) is 13, meaning that the estimate of power for a situation involving two samples of 35 cases each, a difference of population means of 0.2, a common standard deviation of 1.0 is 0.13 (with rounding). Using the Stata **sampsi** command, we would enter the following command:

sampsi 0 0.2 , **sd(1) n(35)**

The use of 0 and 0.2 for the two sample means is a convenient way to represent the difference. It would make no difference what two numbers we inserted here so long as the difference was 0.2 (try it).

Stata produces the following output:

```
. sampsi 0 .2, sd(1) n(35)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
            m1 =        0
            m2 =       .2
           sd1 =        1
           sd2 =        1
sample size n1 =       35
            n2 =       35
         n2/n1 =     1.00

Estimated power:

         power =   0.1332
```

The power estimate of 0.1332 would lead us to expect 13 significant results in 100 tests (=0.1332*100).

The remainder of Table 23.3 can be reproduced simply by changing the expected sample size in the command above from 35 to 100, 200, and 1,000, respectively.

If our interest is in estimating the sample size required to achieve a given level of statistical power, we would alter the **sampsi** command by omitting the sample size values (**n(#)**). For example, in our more detailed example for the difference of means power analysis, we assumed a small standardized effect (i.e., $d = 0.2$) and a one-tail test. To compute the sample size needed to detect this effect, we would enter the following command:

sampsi 0 .2 , **sd(1) power**(.8) **onesided**

The result is 310 cases per group, just as we reported above. To reproduce the sample size estimates for the medium and large effects, simply increase the value of the second mean from 0.2 to 0.5 and 0.8.

ANOVA

Unfortunately, there is no built-in command in Stata to compute power in a simple one-way ANOVA. Although there are several user-written commands that can be installed and used, it is often difficult to obtain the kind of information we may be most interested in gleaning from a power analysis. Based on existing procedures in Stata, as well as other statistical packages, we have written **anova_pwr**, which is a modest Stata procedure that you can use to replicate the results in this chapter and perform other simple one-way ANOVA estimates

of power and sample size. To install this command on your copy of Stata, type the following command:

net install anova_pwr,

from(http://myfiles.neu.edu/c.britt/stata/ado/power)

The basic components to the **anova_pwr** command are the following:

anova_pwr, ngp(#) f(#) min(#) max(#)

where **ngp(#)** represents the number of groups to be compared, **f (#)** represents the standardized effect size (the default is $f = 0.1$), **min(#)** is the minimum number of cases per group (the default is 10), **max(#)** is the maximum number of cases per group (the default is 10), and **byvalue(#)** represents a way to control how much output is generated by the command (default is every fifth line of output).

For our example above, we computed the power of a one-way ANOVA design with three groups (k) and 100 cases in each group for three different effect sizes f (0.1, 0.25, and 0.4).

The **anova_pwr** command to compute the power estimate for the small effect (i.e., f = 0.10) is

anova_pwr, ngp(3) f(.1) min(100) max(100)

Since we specified the minimum and maximum group sizes to be the same (100 cases per group, for a total sample size of 300), there will only be one line of output:

```
. anova_pwr, ngp(3) f(.1) min(100) max(100)

Number of Groups =  3  Effect Size (f) =  .1 alpha =  .05
```

nobs	ncp	Errordf	beta	power
100	3	297	.681357	.3186429

If you rerun this command, but change the value for f to reflect the medium and strong effect sizes, the power estimates reported above will also be reproduced.

It is not possible to estimate directly the sample size required for a designated level of statistical power in an ANOVA using Stata. The **anova_pwr** command can be used to represent a range of group sizes through the use of the **min(#)** and **max(#)** options that will estimate the power associated with a given effect size. If our goal is to achieve a power of 0.80, then we might start by estimating the power for a wide range of sample sizes and then narrowing down the range on a second run. For example, if we are interested in determining the sample size required to detect a small effect ($f = 0.1$) in a study with three groups, we might enter the following command using 100 cases as the minimum, since we already know that it has a power level below 0.8:

anova_pwr, ngp(3) f(.1) min(100) max(500) byvalue(10)

The output from this run is

```
. anova_pwr, ngp(3) f(.1) min(100) max(500) byvalue(10)

Number of Groups =  3  Effect Size (f) =  .1 alpha =  .05
```

nobs	ncp	Errordf	beta	power
100	3	297	.681357	.3186429
110	3.3	327	.6527492	.3472508
120	3.6	357	.6244401	.3755599
130	3.9	387	.5965289	.4034711
140	4.2	417	.569103	.430897
150	4.5	447	.5422376	.4577624
160	4.8	477	.5159978	.4840023
170	5.1	507	.4904379	.5095621
180	5.4	537	.4656031	.5343969
190	5.7	567	.44153	.55847
200	6	597	.4182473	.5817527
210	6.3	627	.395776	.604224
220	6.6	657	.374131	.625869
230	6.9	687	.353321	.646679
240	7.2	717	.3333495	.6666505
250	7.5	747	.314215	.685785
260	7.8	777	.295912	.704088
270	8.1	807	.2784314	.7215686
280	8.4	837	.2617608	.7382392
290	8.7	867	.2458848	.7541152
300	9	897	.2307859	.7692141
310	9.3	927	.2164447	.7835553
320	9.6	957	.2028401	.7971599
330	9.9	987	.1899495	.8100505
340	10.2	1017	.1777494	.8222506
350	10.5	1047	.1662157	.8337843
360	10.8	1077	.1553234	.8446766
370	11.1	1107	.1450476	.8549524
380	11.4	1137	.135363	.864637
390	11.7	1167	.1262442	.8737558
400	12	1197	.1176662	.8823338
410	12.3	1227	.1096042	.8903958
420	12.6	1257	.1020336	.8979664
430	12.9	1287	.0949306	.9050694
440	13.2	1317	.0882718	.9117282
450	13.5	1347	.0820342	.9179658
460	13.8	1377	.0761958	.9238042
470	14.1	1407	.0707351	.9292649
480	14.4	1437	.0656314	.9343686
490	14.7	1467	.0608648	.9391352
500	15	1497	.056416	.943584

As we move through the values in the output, we see that a power level of 0.8 falls between a group size of 320 (power = 0.797) and 330 (power = 0.810). If we rerun the command and limit the range to 320–330 cases per group, we find the following:

```
. anova_pwr, ngp(3) f(.1) byvalue(1) min(320) max(330)

Number of Groups =  3  Effect Size (f) =  .1 alpha =  .05
```

nobs	ncp	Errordf	beta	power
320	9.6	957	.2028401	.7971599
321	9.63	960	.2015193	.7984807
322	9.66	963	.2002056	.7997944
323	9.69	966	.198899	.801101
324	9.72	969	.1975994	.8024006
325	9.75	972	.1963069	.8036931
326	9.78	975	.1950215	.8049785
327	9.81	978	.1937431	.8062569
328	9.84	981	.1924716	.8075284
329	9.87	984	.1912071	.8087929
330	9.9	987	.1899495	.8100505

Consistent with the results reported above, a sample of 323 cases per group would be required to achieve a minimum power of 0.80. The table of results also illustrates how sample size estimates may vary across programs to compute statistical power. A sample of 322 cases per group has a power that is 0.0002 from 0.80, which some programs would round to 0.80. At the same time, it is technically below a value of 0.80 and a sample of 323 cases per group crosses that threshold.

Correlation

There is one user-written procedure that we are aware of for computing power estimates of correlation coefficients in Stata. The command is **sampsi_rho**, which bases power calculations on converting the correlation coefficient with the Fisher *z* formula and then using the normal distribution (instead of a *t*-distribution for the untransformed correlation coefficient). This command can be installed with the following command:

ssc install sampsi_rho

The basic structure of the **sampsi_rho** command is

sampsi_rho , null(#) alt(#) n(#) power(#) solve() alpha(#) onesided

where **null(#)** specifies the value of the correlation for the null hypothesis (default is 0), **alt(#)** specifies the alternative hypothesis value of the correlation (default is 0.5), **n(#)** specifies the sample size (default is 100), **power(#)** indicates the desired level of power (default is 0.9), **solve()** notes whether to solve for sample size (n, which is the default) or power, **alpha(#)** specifies the alpha level if different than 0.05 (the default), and onesided indicates that a one-tail test is to be performed (a two-tailed test is the default).

To replicate the values above in our analysis of power estimates for correlation coefficients for a sample size of 100, we would enter the following command in Stata to estimate the power to detect a weak correlation:

sampsi_rho, solve(power) n(100) alt(0.1) onesided

The estimated power is 0.260, very nearly the same as the estimate produced using the correlation coefficient and the *t*-distribution. If you were interested in reproducing the power estimates for the medium and strong effects, you would just need to change the value of **alt(#)** to **alt(0.3)** and **alt(0.5)**, respectively.

In a similar way, we can estimate the sample size needed to achieve a designated level of statistical power for a hypothesized effect size by making just a few changes to the **sampsi_rho** command. For example, if we wanted to estimate the sample size needed to detect a medium correlation (i.e., $\mathbf{r} = 0.3$) with a power level of 0.80, we would omit the sample size and solve options but insert **power(0.8)** and enter the following command:

sampsi_rho, alt(0.3) power(0.8) onesided

We find that the estimated sample size is 67.53. Since we cannot find a fraction of a case, we would typically round up to 68 in this case. The rationale, as we noted above, in rounding up is to ensure that a power level of no less than our target (e.g., 0.80) is achieved. Note, too, that the sample size estimated here (68) is slightly larger than that estimated above in the text (64). The difference is entirely due to the use of the Fisher-transformed value of the correlation and use of the normal distribution and is to be expected.

OLS Regression

Similar to computing power with ANOVA in Stata, it is necessary to rely on the user-written command **powerreg**. The installation of this command is similar:

net install powerreg, from(http://www.ats.ucla.edu/stat/stata/ado/analysis)

The basic structure to the powerreg command is

powerreg, r2f(value) r2r(value) nvar(#) ntest(#) alpha(value)

where **r2f(value)** is the hypothesized value of R^2 expected, **r2r(value)** is the R^2 for the reduced (null) model, **nvar(#)** refers to the total number of independent variables included in the regression model, and **ntest(#)** refers to the number of independent variables being tested. Alpha is assumed to be 0.05, and **nvar** and **ntest** are both set at a default of 1.

To reproduce the results reported above for power in OLS regression for a weak effect (i.e., $R^2 = 0.02$), we would use the command

powerreg, r2f(.02) r2r(0) n(100) nvar(8) ntest(8)

Note that the value for **r2r** is entered as 0—this is the expected value of R^2 without any of the independent variables included in the analysis. The power estimate reported by Stata is 0.1253, very close to the result of 0.124 reported above. Results for the moderate ($R^2 = 0.13$) and strong ($R^2 = 0.26$) effect sizes are obtained by simply altering the value of **r2f** in the **powerreg** command. Note that the power estimates reported by Stata vary slightly from those reported above using G*Power.

To compute the sample size needed to achieve a designated level of statistical power, we would omit the **n(#)** option but insert an option for **power(#)**:

powerreg, r2f(.02) r2r(0) power(0.8) nvar(8) ntest(8)

We find that the estimated sample size needed to detect a weak effect ($R^2 = 0.02$) is 740, which is different from the value of 759 reported above. This is due to the calculation of the standardized effect (f^2) and rounding error for the weak effect size[20]—the values for the medium and large effect sizes are nearly identical and differ by only 1 case.[21]

Problems

1. Compute the estimates of statistical power for each of the four scenarios in Exercise 21.1. Which scenario has the highest level of statistical power? Explain why.
2. Compute the estimates of statistical power for each of the four scenarios in Exercise 21.3. Which scenario has the highest level of statistical power? Explain why.
3. Compute the post hoc estimates of statistical power for the comparisons in Exercise 21.6. Was this study designed to have a high level of statistical power to identify small and medium effects? Explain why.
4. Compute the estimates of statistical power for each of the following one-way ANOVA studies. (For all scenarios, assume that the researcher is trying to detect a small effect.)
   - Scenario 1: Three groups with 75 cases per group.
   - Scenario 2: Four groups with 60 cases per group.
   - Scenario 3: Five groups with 55 cases per group.

   Which scenario would have the highest level of statistical power? Explain why.

---

[20] If the value of **r2r(#)** in the command is changed to **r2r(0.0196)**, the resulting estimate for required sample size is 760—a value that differs by 1 from the 759 cases reported above.

[21] The reason for this difference is that the **powerreg** command computes sample size estimates that have a corresponding power level of just under 0.80, while G*Power estimates a sample size that ensures that the power level is at least 0.80 and so the estimates reported above are greater by 1.

5. In attempting to design a correlation study looking at academic performance and delinquency, a researcher expects a small-to-moderate correlation among a population of adolescents he or she will sample from.

   a. If he or she computes estimates of statistical power assuming a two-tail test, what size sample would he or she need to detect a small correlation? Medium correlation?

   b. Do you think he or she could justify a one-tail test of the correlation? If a one-tail test was used, how does the estimated sample size change for both the small and medium correlations?

6. A research team is preparing to launch a statewide survey to gauge public sentiment about the incarceration of juvenile offenders, focusing primarily on support for more lenient punishments. Consistent with much public opinion research, expectations are that a combination of ten independent variables is likely to explain about 15% of the variation in views about juvenile punishment.

   a. What size sample would the researchers need to have to achieve a power of 0.80? 0.90?

   b. Of particular interest to the researchers is the effect of three different measures of experience with the justice system, but their expectation is that the overall effect of these three measures will be small. What size sample would the researchers need to achieve a power of 0.80? 0.90?

   c. What size sample should the researchers try to obtain? Explain why.