CHAPTER 16

Statistical Power

CHESTER L. BRITT AND DAVID WEISBURD

INTRODUCTION

Criminal justice researchers have placed a premium on statistical inference and its use in making decisions about population parameters from sample statistics. In assessing statistical significance, the focus is on 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. 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 what we fundamentally mean by statistical power – given the current design of a study, does it have the ability (i.e., power) to detect statistically significant effects and relationships?

Although researchers in criminal justice have placed much more emphasis on statistical significance than on the statistical power of a study, research in fields, such as medicine and psychology routinely report estimates of statistical power (see, e.g., Maxwell et al. 2008). Funding agencies (e.g., National Institutes of Health) are increasingly likely to require research proposals to estimate how powerful the proposed research design will be. The purpose of this chapter is to present the key components in an assessment of statistical power, so that criminal justice researchers will have a basic understanding of how they can estimate the statistical power of a research design or to estimate the size of sample necessary to achieve a given level of statistical power. Toward that end, our discussion is organized as follows. The next two sections present the basic and conceptual background on statistical power and the three key components to statistical power, respectively. We then focuses on the computation of statistical power estimates, as well as estimates of sample size, for some of the most common

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_16,

types of statistical tests researchers will confront.[1] Finally, we conclude by noting some of the more recent developments and future directions in estimating statistical power in criminology and criminal justice.

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:

$$\text{Power} = 1 - \text{Probability(Type II error)} = 1 - \beta.$$

For many sample statistics, the Type II error can be estimated directly from the sampling distributions, commonly assumed for each statistic. In contrast to a traditional test of statistical significance, which identifies the risk of stating that factors are related when they are not (i.e., the Type I error), for the researcher 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 10% probability of falsely failing to reject the null hypothesis. Alternatively, a study with a statistical power estimate of 0.40 has 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 16.1 presents the relationship between Type I and Type II errors graphically. Suppose that we are interested in a difference in groups 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 group means 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 Fig. 16.1 represents the t-distribution – the sampling distribution – under the null hypothesis, with the significance level (α) noted in the right tail of the distribution – the vertical line represents the critical

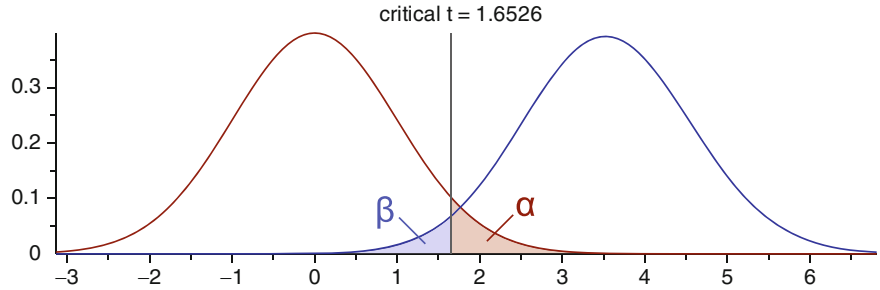

FIGURE 16.1. Graphical representation of Type I and Type II errors in a difference of means test (100 cases per sample).

[1] Since the actual computation of statistical power varies with the sample statistic being tested, there is voluminous literature on how to compute statistical power for a wide range of statistical models and our discussion is necessarily limited.

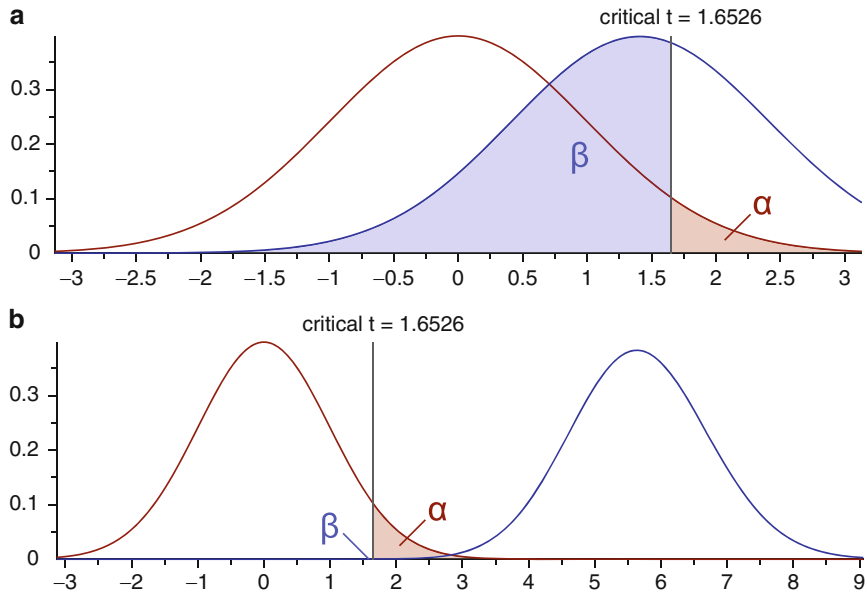

FIGURE 16.2. Graphical representation of Type I and Type II errors in a difference of means test – changing the difference in mean values. (**a**) smaller difference in means – fixed sample size. (**b**) larger difference in means – fixed sample size.

value. The distribution on the right side of Fig. 16.1 represents the hypothesized sampling distribution based on prior research and theory and our expectations for differences in the two group means. The probability of making a Type II error (β) is the cumulative probability in the figure on the right, which represents the chances of finding a difference in the two group means that is less than the critical value. The power of the difference of means test is represented in the figure by the area under the curve on the right that falls to the right of the critical value – the difference between 1 and β.

It is important to note that our estimate of β is fully dependent on our estimate of the magnitude of the difference between the two groups. Figure 16.2 illustrates the differences for alternative effect sizes, 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 to the left, increasing our estimate of β (see Fig. 16.2a). If we expect a large difference, we would shift the hypothesized sampling distribution to the right, reducing the estimate of β (see Fig. 16.2b).

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 (Weisburd 1991).

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 offender reentry to the community following incarceration. 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 improving the outcomes of offenders after release from prison. Although the research team is likely to say that the program does not have a statistically significant impact on offender reentry, 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. Although the relationship may exist in the population of interest, 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, resulting in many criminal justice studies having a low level of statistical power (Brown 1989; Weisburd 1991).

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 the chances of a Type II error are $\beta = 0.20$. In many ways, 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 $0.10/0.05 = 2.0$. Consequently, 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 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 want to maintain the ratio of error probabilities at 4.0, we 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 our 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 within the control of the researcher, while the estimated effect size is not. The following discussion briefly highlights 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 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 16.1 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, 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 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

TABLE 16.1. z-Scores needed to reject the null hypothesis in a two-tailed test of statistical significance

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

TABLE 16.2. z-Scores needed to reject the null hypothesis in one-tailed and two-tailed tests of statistical significance

α	0.20	0.10	0.05	0.01	0.001
z-score (1-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 (2-tail test)	±1.282	±1.645	±1.960	2.576	3.291

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 16.2 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 need only 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 need only 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 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, it is well known that the standard error (se) for a single-sample t-test is computed as:

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

TABLE 16.3. Number of statistically significant outcomes expected in 100 two-sample *t*-tests for four scenarios

Scenario	Sample size (per group)	$\mu_1 - \mu_2$	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

As N gets larger, irrespective of the value of the standard deviation (*sd*) itself, the standard error of the estimate 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 16.3. The last column of Table 16.3 indicates the number of statistically significant outcomes expected in 100 two-sample t-tests in which there is 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 16.3 shows that the likelihood of rejecting the null hypothesis changes substantially with each increase in sample size, even though all 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 on how to determine the appropriate sample size for a proposed research study (see, e.g., Dattalo 2008; Kraemer and Thiemann 1987; Murphy and Myors 2003).

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 (Weisburd 1991). 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 6-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 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 (Petersilia 1989). 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.

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 (ES) is defined as:

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

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.[2]

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

[2] 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 Lipsey and Wilson (2001) and Rosenthal (1984).

populations, variability is defined as the pooled or common standard deviation of the outcome measures in the two populations (σ). Consequently, ES would be computed as:

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

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:

$$\text{ES} = \frac{(\mu_1 - \mu_2)}{\sigma}.$$

Thus, 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 16.4 presents a simple illustration of the relationship between effect size and statistical power in practice. The last column of Table 16.4 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, under 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 16.4 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

TABLE 16.4. Number of statistically significant outcomes expected in 100 two-sample *t*-tests for six different scenarios (100 cases in each sample)

Scenario	μ_1	μ_2	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

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 (Lipsey 1990). 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 (Petersilia 1989). 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 (Weisburd 1991).

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 grow.

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 (Cohen 1988; Kraemer and Thiemann 1987; Lipsey 1990; Murphy and Myors 2003). 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 discussion 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 to interested researchers. A later section of the chapter highlights several easily accessible resources that that will perform the power estimates reported in 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 a priori power analysis, and where our attention is focused 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 (see, e.g., Hayes and Steidl, 1997; Thomas 1997). 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 dataset 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 the 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 (Cohen 1988). 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.

The fourth element, defining effect size, is more difficult. 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 the statistical power is sometimes defined as the design sensitivity (Lipsey 1990). 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 the effect size has been made easier by identifying broad categories of effect size that can be compared across studies. Cohen's (1988) suggestions have been the most widely adopted by other researchers and simply refer to classifying effect sizes as small, medium, and large. The specific value of 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 (1988) work. For example, Raudenbush and Liu (2000) develop criteria for defining small, medium, and large standardized effects in hierarchical linear models.

The following discussion turns to the computation of statistical power for several common situations in criminology and criminal justice research: difference of means test, ANOVA, correlation, and OLS regression. Of course, there are a variety of other statistical models used in criminological research. In some cases, there are well-developed means for determining statistical power. In other applications, there are no clear guidelines for estimating the statistical power, and it is up to the researcher to conduct simulation studies to estimate the level of statistical power for a given model and sample. For example, structural equation models are notoriously complex, and with slight changes to the model being estimated, the statistical power estimates can be wildly different and necessitate the use of simulations to estimate power (see, e.g., Muthén and Muthén 2002).

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 research hypothesis (see again Fig. 16.1, above). The sampling distribution under the research hypothesis is referred to as a noncentral distribution. For example, in Fig. 16.1, the sampling distribution under the null hypothesis is the t-distribution, while the sampling distribution under the research hypothesis is the noncentral t-distribution.

The noncentral sampling distribution is computed based on a "noncentrality" 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 noncentrality parameter and explain 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 noncentrality parameter δ for the t-distribution is

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

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 noncentrality parameter δ is

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

To illustrate the computation of a statistical power estimate, suppose that we want to assess the effectiveness of a treatment program for property 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 ($d = 0.2$) at the medium ($d = 0.5$), and at the 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 noncentrality parameter δ is 1.414 $\left(= 0.2 \times \sqrt{(200/4)}\right)$. 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. 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. 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

$$f = \frac{\sigma_m}{\sigma},$$

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 noncentrality parameter λ for the F-distribution is

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

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 ($f = 0.1$), at the medium ($f = 0.25$), and at the 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 noncentrality parameter λ is 3 ($=0.1^2 \times 300$). 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. 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 situation with only two groups, our research design with 100 cases assigned to each of 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 third decimal.

The noncentrality parameter δ for the correlation coefficient is:

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

where r is either the sample correlation coefficient (r) or the 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 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 ($r = 0.1$), at the medium ($r = 0.3$), and at the 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 noncentrality 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. 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 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,

$$R^2 = f^2(1 + f^2).$$

Cohen's (1988) recommendations for small, medium, and large standardized effect sizes 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 noncentrality 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 and control 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%. For the small effect size ($f^2 = 0.02$), we have noncentrality parameter $\lambda = 2.0$ ($=0.02\times 100$). 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. For the medium effect ($f^2 = 0.15$), $\lambda = 15.0$, $\beta = 0.242$, and power $= 0.758$, which is still an inadequate level of power, although it 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 and control variables, what sample size is required to achieve a statistical power level of 80% for detecting effects at the small ($f^2 = 0.02$), at the medium ($f^2 = 0.15$), and at the 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. 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 and/or control 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 and control 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 to 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 and control 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 incorporate 10 to 20 independent and control variables and have a full model R^2 typically between 0.15 and 0.20. This implies that many of the individual-variable effects will tend to be quite small in magnitude and in order for an analysis to detect a statistically significant relationship, a large sample becomes necessary.

Statistical Software

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 two different software packages: (1) G*Power (version 3.0.10) and (2) the `pwr` library available for the *R* system. 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. All of the figures presented in this chapter were produced with G*Power 3. Faul et al. (2007) provide a useful overview of the capabilities of G*Power 3.

The *R* system is a comprehensive mathematical and statistical package that is modeled on S-Plus (R Development Core Team 2009). *R* can be freely downloaded from the Comprehensive *R* Archive Network website (http://cran.r-project.org/). The pwr library (Champely 2007) is a user-contributed set of routines that computes the basic estimates of statistical power discussed in this chapter and is based on Cohen's 1988) work. The *R* system also includes a powerful set of graphics routines and libraries. Although not as user-friendly for computing complicated graphs of power estimates as G*Power 3, every graph produced in this chapter can be reproduced in the *R* system.

Power and Precision v. 2.0 (Borenstein et al. 2001) is a commercially available software package designed to compute power estimates for a wide range of statistical models in a user-friendly environment. Its range of capabilities is 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 addition to the software options noted above, more widely available all-purpose statistical packages – SAS, SPSS, and Stata – also include routines for estimating statistical power. For readers interested in pursuing the estimation of statistical power with one of these three packages, consult the appropriate software websites.

For the reader who 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 web sites 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.

Finally, it is worth noting 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.

SUMMARY AND CONCLUSION

The focus of this chapter has been the presentation of the key components and the rationale underlying the computation and the estimation of statistical power and sample size. Due to the wide range of statistical models used in the study of crime and criminal justice, our discussion necessarily touched on a limited number of more common situations. That said, the key components of statistical power – level of statistical significance, size of sample, and effect size – apply to estimating statistical power across all types of linear models, regardless of the specific form that a model may take. The manner in which statistical power is computed may vary widely by the type of model, but the components of statistical power remain the same.

Beyond the issues discussed in this chapter, there are four other important issues that we wish to highlight. First, in recent years, there has been a significant growth in the computation of statistical power for a much wider range of statistical models. For example, multilevel models that involve the analysis of clustered data raise a number of issues related to statistical

power: How many groups? How many observations within each group? What are the estimates of statistical power, given different configurations for the number of groups (at different levels of aggregation) and the number of individual observations? Raudenbusch and Liu's (2000) simulation study focusing on multilevel models is a nice example for how power estimates vary across a range of small, medium, or large effects that they tried to make comparable to Cohen's (1988) recommendations.

In a similar way, the complexity of structural equation models makes the computation of statistical power estimates that much more challenging. Muthén and Muthén (2002) illustrate the application of simulation studies to structural equation models as a means for estimating statistical power. In general, simulation studies will likely grow in their use, as researchers confront increasingly complex multivariate models and are either unsure about the sampling distributions of the statistics being tested or are attempting to measure the sampling distribution. Using a similar approach, Kleiber and Zeileis (2008) provide an example for how to compute the power of the autocorrelation coefficient in an autoregression model.

Second, much of the written work on tests of statistical power notes that we make important assumptions about the sampling distributions for the statistics we estimate. Since the assumptions for many of the statistical models used by researchers in criminology and criminal justice are often violated, it is not clear that the assumed sampling distributions remain appropriate. Bootstrapping techniques offer a way around assuming a specific sampling distribution for a test statistic by allowing the researcher to generate an empirical sampling distribution through resampling the current sample data and performing the same statistical analysis (see, for example, Efron and Tibshirani (1993) and Dattalo (2008)). The empirical sampling distribution can then be used to provide a better sense of the statistical power of a study, since it will be based on the data, methods, and models used by the researcher.

Third, Maltz (1994) noted several years ago that the expanded use of archival datasets in criminology and criminal justice resulted in many researchers analyzing populations, rather than samples. The increased frequency with which populations are analyzed calls into question many of the assumptions about performing tests for statistical significance. 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 dataset is as large as it possibly can be, and the effect size is simply what is observed (excluding measurement error).

Finally, we think it is important for all researchers to be sensitive to issues of statistical power. Although much of the research published in criminology and criminal justice journals relies on secondary data sources, we think that researchers analyzing these datasets will still benefit from thinking through issues of statistical power prior to setting off on a complicated analysis. Clearly, researchers cannot alter the size of the sample in a secondary dataset. At the same time, the use of the techniques discussed in this chapter and elsewhere should provide researchers with a better sense of the potential findings from any given dataset, which in the long run should improve the quality of the research enterprise.

REFERENCES

Borenstein M, Rothstein H, Cohen J (2001) Power and precision. Biostat, Inc., Englewood, NJ
Brown SE (1989) Statistical power and criminal justice research. J Crim Justice 17:115–122
Champely S (2007) pwr: Basic functions for power analysis. Rpackage version 1.1.

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, NJ
Dattalo P (2008) Determining sample size. Oxford University Press, New York
Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
Faul F, Erdfelder E, Land AG, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191
Hayes JP, Steidl RJ (1997) Statistical power analysis and amphibian population trends. Conserv Biol 11:273–275
Kleiber C, Zeileis A (2008) Applied econometrics in R. Springer, New York
Kraemer HC, Thiemann S (1987) How many subjects: statistical power analysis in research. Sage, Newbury Park, CA
Lipsey MW (1990) Design sensitivity: statistical power for experimental research. Sage, Newbury Park, CA
Lipsey M, Wilson D (2001) Practical meta-analysis. Sage, Thousand Oaks, CA
Maltz MD (1994) Deviating from the mean: the declining significance of significance. J Res Crime Delinq 31: 434–463
Maxwell SE, Kelley K, Rausch JR (2008) Sample size planning for accuracy in parameter estimation. Annu Rev Psychol 59:537–563
Murphy KR, Myors B (2003) Statistical power analysis, 2nd edn. Lawrnce Erlbaum, Mahwah, NJ
Muthén LK, Muthén BO (2002) How to use a monte carlo study to decide on sample size and determine power. Struct Equ Model 9:599–620
Petersilia J (1989) Randomized experiments: lessons from BJA's intensive supervision project. Eval Rev 13:435–458
Raudenbusch S, Liu X (2000) Statistical power and optimal design for multisite randomized trials. Psychol Methods 5:199–213
R Core Development Team (2009) R: a language and environment for statistical computing. http://www.R-project.org
Rosenthal R (1984) Meta-analytic procedures for social research. Sage, Beverly Hills
Thomas L (1997) Retrospective power analysis. Conserv Biol 11:276–280
Weisburd D (1991) Design sensitivity in criminal justice experiments. Crim Justice 17:337–379