Steidl, R.J. and L. Thomas. 2001. Power analysis and experimental design. Pp 14-36 In Scheiner, S.M. and J. Gurevitch (eds) Design and Analysis of Ecological Experiments. 2nd Edition. Oxford University Press, New York.

2

Power Analysis and Experimental Design

ROBERT J. STEIDL

LEN THOMAS

2.1 Introduction

Ecologists conduct research to gain information about ecological patterns and processes (chapter 1). An underlying, fundamental goal for all research, therefore, is to generate the maximum amount of information from a given input of effort, which can be measured as time, money, and other similarly limited resources. Consequently, after we develop a clear set of questions, objectives, or hypotheses, the most critical aspect of ecological research is design.

Designing a new study involves making a series of interrelated choices, each of which will influence the amount of information gained and, ultimately, the likelihood that study objectives will be met. For a manipulative experiment, choices must be made about the number of treatments to apply and the way in which treatments are assigned to experimental units. Similarly, for an observational study, samples must be selected in some way from the larger population of interest. In both types of research, a critical decision is the number of replicates (experimental units receiving the same treatment) or samples to choose. When considering these issues, it is helpful to have a tool to compare different potential designs. Statistical power analysis is one such tool.

In this chapter, we focus on the use of power analysis in research design, called prospective (or a priori) power analysis. We review some basic theory and discuss the practical details of doing prospective power analyses. We also consider the usefulness of calculating power after data have been collected and analyzed, called retrospective (or a posteriori or post hoc) power analysis. Power analysis is most appropriate when data are to be analyzed using formal hypothe-

sis-testing procedures. However, parameter estimation is often a more appropriate and informative approach by which to make inferences, so we discuss related techniques when estimation is the main goal of the study. Our discussion stays within the frequentist statistical paradigm; issues within the likelihood and Bayesian frameworks are considered elsewhere (chapter 17; Berger 1985; Royall 1997; Burnham and Anderson 1998; Barnett 1999).

Power analysis is increasing in popularity, as evidenced by the spate of introductory articles recently published in the biological literature (e.g., Hayes 1987; Peterman 1990; Muller and Benignus 1992; Taylor and Gerrodette 1993; Searcy-Bernal 1994; Thomas and Juanes 1996; Steidl et al. 1997). These all provide somewhat different perspectives and, in some cases, different background material than we present here. Unfortunately, power is given only cursory treatment in many biometry textbooks (e.g., Sokal and Rohlf 1995; Steel et al. 1996), although this has been changing to some extent (e.g., Rao 1998; Zar 1996). In addition, some specialized texts (Kraemer and Thiemann 1987; Cohen 1988; Lipsey 1990) and an excellent introductory monograph (Nemac 1991) focus on implementing power analysis using SAS. We provide other selected references throughout this chapter.

2.2 Statistical Issues

2.2.1 Statistical Hypothesis Testing

The theory of power analysis falls within the larger framework of statistical hypothesis testing (the so-called Neyman–Pearson approach; Neyman and Pearson 1928; Barnett 1999). In this framework, a research question is phrased as a pair of complementary statistical hypotheses, the null (H_0) and alternative (H_a) hypotheses. The finding that would be of interest to the researcher is stated typically as the alternative hypothesis, and a negative finding is stated as the null hypothesis. For example, suppose we were interested in assessing whether the average amount of plant biomass harvested per plot differs between control and treatment plots subjected to some manipulation. Typically, the null and alternative hypotheses of interest (in this case as two-tailed hypotheses) would be phrased as

H_0: $\mu_T = \mu_C$, which represents the case of equal population means

H_a: $\mu_T \neq \mu_C$, which represents the case of unequal population means

Imagine that we have collected data from 20 plots, 10 treatment and 10 control. We can use these data to calculate a test statistic that provides a measure of evidence against the null hypothesis of equal means. If we make a few assumptions about the distribution of this statistic, we can calculate the probability of finding a test statistic at least as extreme as the one observed, if the null hypothesis is true. This probability is often called the *P*-value or significance level of the test. Lower *P*-values suggest that the test statistic we calculated would be an unlikely result if the null hypothesis were indeed true, whereas higher *P*-values

suggest that the test statistic would not be an unlikely result if the null hypothesis were true.

In this example, imagine that the amount of plant biomass averaged 113 kg/ha on treatment plots and 103 kg/ha on control plots (i.e., estimates of true population means μ_T and μ_C are $\bar{y}_T = 113$ and $\bar{y}_C = 103$, respectively). From these data, we also determined $s = 15$, which is an estimate of the pooled population standard deviation, σ. With this information and presuming the data approximate the necessary assumptions, we can use a two-sample *t*-test to generate a test statistic for the null hypothesis of $\mu_T = \mu_C$:

$$t = \frac{\bar{y}_T - \bar{y}_C}{\frac{s}{\sqrt{n}}} = \frac{113 - 103}{\frac{15}{\sqrt{20}}} = 2.98$$

This test statistic, based on a sample size of 20 (and therefore 18 degrees of freedom for this test) is associated with a two-tailed $P = 0.008$, indicating that the probability of obtaining a test statistic at least as extreme as the one we observed (2.98) if the null hypothesis of equal means is true is about 8 in 1,000—a reasonably unlikely occurrence.

If we apply the hypothesis-testing framework rigorously (which we do not advocate, but which is necessary for this discussion), we would use the value of the test statistic as the basis for a dichotomous decision about whether to reject the null hypothesis in favor of the alternative hypothesis. If the test statistic exceeds an arbitrary threshold value, called a *critical value*, then we conclude that the null hypothesis is false because evidence provided by the data suggests that attaining a test statistic as extreme as the one we observed is unlikely to occur by chance. The critical value is the value of the test statistic that yields $P = \alpha$, where α is the Type I error rate established by the researcher before the experiment is performed (see subsequent discussion). In this example, if we had chosen a Type I error rate of $\alpha = 0.05$, then $t_{crit} = 2.10$. Because the observed *t*-value is greater than the critical value, we reject the null hypothesis, which is a "statistically significant" result at the given α-level.

There is always a chance, however, that no matter how unlikely the test statistic (and therefore, how low the *P*-value), the null hypothesis may still be true. Therefore, each time a decision is made to reject or not reject a null hypothesis, there are two types of errors that can be made (table 2.1). First, a null hypothesis

Table 2.1 Possible outcomes of statistical hypothesis tests[a]

	Decision and result	
Reality	Do not reject null hypothesis	Reject null hypothesis
Null hypothesis is true	Correct $(1 - \alpha)$	Type I error (α)
Null hypothesis is false	Type II error (β)	Correct $(1 - \beta)$

[a]Probabilities associated with each decision are given in parentheses.

that is actually true might be rejected incorrectly (a Type I error; a false positive). As in the previous example, the rate at which a Type I error will be accepted is the α-level and is established by the researcher. Second, a null hypothesis that is actually false might not be rejected (a Type II error; a false negative). The probability of a Type II error is denoted as β. Statistical power is equal to $1 - \beta$ and is defined as the probability of correctly rejecting the null hypothesis, given that the alternative hypothesis is true (figure 2.1).

The statistical power of a test is determined by four factors in the following ways: power increases as sample size, α-level, and effect size (difference between the null and alternative hypothesis) increase; power decreases as variance increases. Some measures of effect size incorporate variance, leaving only three components. Effect size is the component of power least familiar to many researchers; we discuss this in detail in the next section.

2.2.2 Measures of Effect Size

In the context of power analysis; effect size is defined broadly as the difference between the null hypothesis and a specific alternative hypothesis. The null hypothesis is often one of no effect, and in these cases effect size is the same as the alternative hypothesis. For example, in the plant biomass experiment, the null hypothesis is no difference in mean biomass between treatment and control plots. One specific alternative hypothesis states that a 20 kg/ha difference between treatment and control plots exists. Effect size, in this case, is $(20 - 0) = 20$ kg/ha. However, other measures of effect size could have been used.

Choosing a meaningful effect size (or range of effect sizes) to use in experimental planning is usually the most challenging aspect of power analysis. In gen-

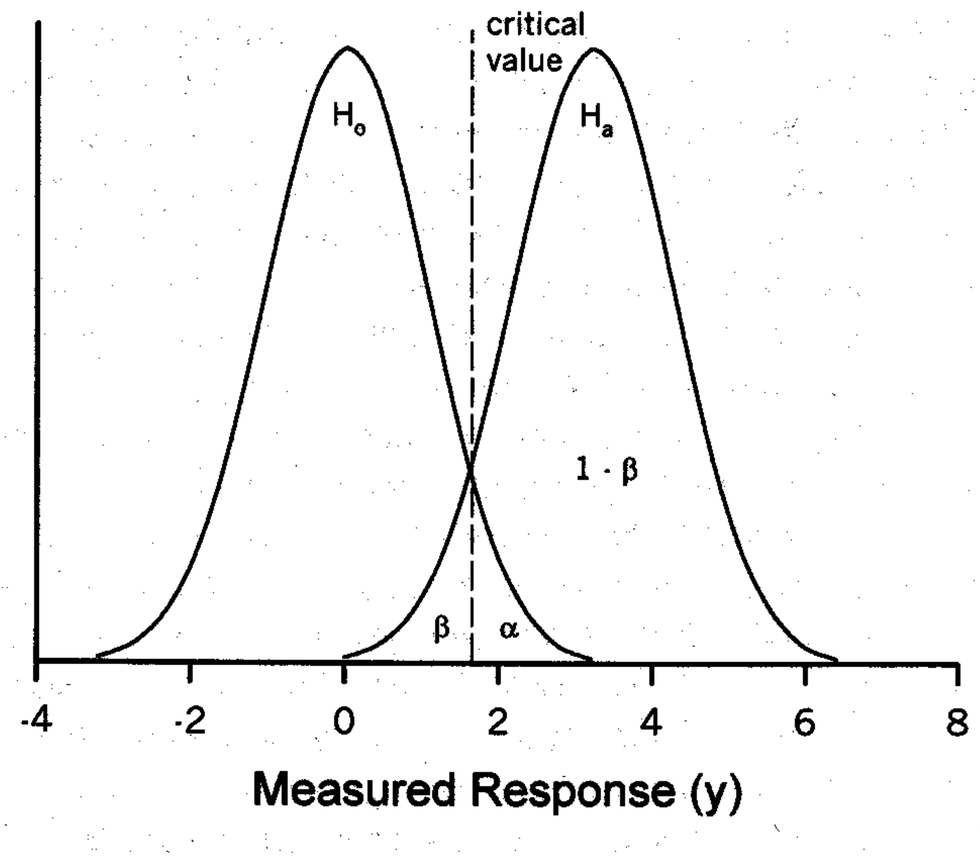

Figure 2.1 Graphical illustration of α, β, power $(1 - \beta)$, and the critical value for a statistical test of a null (H_0) versus alternative (H_a) hypothesis.

eral, we want to use effect sizes that are biologically important in the context of the study. In the plant biomass experiment, a difference of 1 kg/ha is not likely to be biologically important, but a difference of 20 kg/ha may be, depending on the goals of the study.

The effect size concept can also be used to quantify the results of experiments (chapter 18). In these cases, effect size is then defined broadly as the treatment response (for manipulative experiments) or the degree to which the phenomenon of interest is present in the population of interest (for observational studies). Effect size used in power analysis is not a population parameter; rather it is a hypothetical value that is determined by the null and alternative hypotheses as specified by the researcher. This point is critical and causes a great deal of confusion when power is examined retrospectively (section 2.3.2).

When discussing results of power analyses, the particular measure of effect size used must be specified explicitly because there usually are several measures of effect size available for a given statistical test (Richardson 1996). We introduce several measures of effect size and comment on their use.

Simple effects. When the question of interest can be reduced to one about a single parameter, such as questions about the difference between two population means (or any other parameter) or the difference between a single population mean and a fixed value, then establishing a meaningful measure of effect size is straightforward. Most apparent are measures of *absolute effect size* (or raw effect size), which are stated as departures from the null hypothesis and have the same units as the parameters of interest. For example, in a two-sample setting comparing two population means, the null hypothesis typically would be stated as H_0: $\mu_1 - \mu_2 = 0$. A useful measure for specifying absolute effect size, therefore, is the difference between population means (or equivalently, the difference between the null and alternative hypotheses): $|\mu_1 - \mu_2|$. We used this measure of effect size when establishing the effect size of 20 kg/ha in the previous example. Similarly, in simple linear regression, one measure of absolute effect size is the difference between the slope of the regression line and a slope of zero (or any other fixed, meaningful value, such as the annual rate of change in a monitored population that would trigger management action). In logistic regression, a measure of absolute effect size is the deviation from an odds ratio of 1 (chapter 11). Because absolute effect sizes are related directly to measurements made by researchers, they are the easiest to specify and interpret.

In research studies with a temporal or spatial control, measures of *relative effect size* are useful because they represent the change in the response variable due to a treatment relative to the control $(\mu_T - \mu_C)/\mu_C$. Relative effect sizes are usually expressed as percentages, for example, the percentage increase in population size due to treatment. In the plant biomass example, we could specify that we are interested in a 20% increase in yield. This would correspond to a yield of 120 kg/ha if the true average harvest in the control plot were 100 kg/ha ((120 − 100)/100 = 20%). Finally, *standardized effect sizes* are measures of absolute effect size scaled by variance (or a related estimate of variation) and therefore combine these two components of hypothesis testing. In the two-sample setting, a standardized measure of effect size is $|\mu_1 - \mu_2|/\sigma$, where σ is the pooled within-population

standard deviation. In the plant biomass example, if the population standard deviation were 15 and the true yield from control plots were 100 kg/ha, then an absolute effect size of 20 kg/ha would correspond to a standardized effect size of $|120 - 100|/15 = 1.33$. Standardized measures are unitless and therefore comparable across studies. They can be useful in research planning as a way to specify effect size when no previous data exist (for example, when there is no information about σ). However, they may be more difficult to interpret in terms of biological importance, so we prefer specifying absolute or relative measures where possible and considering the variance component of power analysis separately.

Complex effects. Establishing a meaningful effect size when an experiment includes multiple factors or multiple levels of a single factor is considerably more challenging. For example, if the plant biomass experiment were extended to include additional treatment levels, one possible null hypothesis for this new experiment would be H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$, where μ_1 is the control yield and μ_2 through μ_k are the yield of the $(k - 1)$ treatment levels (a one-factor fixed-effect ANOVA). In this context, a useful absolute effect size can be based on the variance of the population means:

$$\sigma_\mu^2 = \frac{1}{k}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2$$

Subsequently, $(\sigma_\mu^2)^{1/2}$ or σ_μ provides an absolute measure of effect size in the same units as the original data, much like the two-sample measure $|\mu_1 - \mu_2|$ discussed previously. Unlike the two-sample case, however, biological interpretation of σ_μ with more than two groups can be challenging.

Four approaches have been used to establish effect sizes for these more complex situations. The first approach is to specify all of the cell means (the μ_i's). In an experiment with three treatments and a control, for example, we might specify that we are interested in examining power given a control yield of 100 kg/ha and treatment yields of 120, 130, and 140 kg/ha. This approach requires researchers to make an explicit statement in terms of the experimental effects they consider to be biologically important. Although this exercise is challenging, these statements are easily interpretable. The second approach is to use measures of effect size such as σ_μ, but to seek to understand their meaning by experimenting with different values of the μ_i's. For example, yields of 100, 120, 130, and 140 kg/ha, correspond to a σ_μ of 7.4. After some experimentation with different yields, we may conclude that $\sigma_\mu \geq 7$ represents a biologically important effect. The third approach is to simplify the problem to one of comparing only two parameters. For example, in a one-factor ANOVA, we could define a measure of absolute effect size as $(\mu_{max} - \mu_{min})$, which places upper and lower bounds on power, each of which can be calculated (Cohen 1988; Nemac 1991). The fourth approach is to assess power at prespecified levels of standardized effect sizes (e.g., σ_μ/σ for the previous ANOVA example or $|\mu_1 - \mu_2|/\sigma$ for a two-sample *t*-test) that have been suggested for a range of tests (Cohen 1988). In the absence of other guidance, power can be calculated at three levels as implied by the adjectives small, medium, and large (Cohen 1988). These conventions are used widely in psychol-

ogy and other disciplines, where a medium standardized effect size may correspond to median effect sizes used in psychological research (Sedlmeier and Gigerenzer 1989). There is no guarantee, however, that these standardized effect sizes have any meaning in ecological research, so we recommend this approach only as a last resort.

2.3 Types of Power Analyses

2.3.1 Prospective Power Analyses

Prospective power analyses are performed when planning a study. They are exploratory in nature and provide the opportunity to investigate—individually or in some combination—how changes in study design and the components of power (sample size, α, effect size, and within-population variance) influence the ability to achieve study goals. Most commonly, prospective power analyses are used to determine (1) the number of replicates or samples (n) necessary to ensure a specified level of power for tests of the null hypotheses, given specified effect sizes, α, and variance, (2) the power of tests of the null hypothesis likely to result when the maximum number of replicates possible is constrained by cost or logistics, given the effect sizes, α, and variance, and (3) the minimum effect size that can be detected, given a target level of power, α, variance, and sample size.

Example 1. Sample sizes necessary to achieve a specified level of power, where population variance is known. Imagine that we are planning a new plant biomass experiment. Assume from previous work that field plots yielded an average of 103 kg/ha under control conditions and that the population standard deviation, σ, was 16. In this new experiment, we decide to consider the treatment effective if it increases or decreases plant biomass on plots by an average of 20% (i.e., we will use a two-tailed test). The relative effect size of interest, therefore, is 20%, and the absolute effect size is 20% of control plots or 103 kg/ha $\times$ 0.20 = 20.6 kg/ha. After some consideration of the relative consequences of Type I and Type II errors in this experiment (section 2.5.4), we establish $\alpha = \beta = 0.1$, so the target power is $1 - \beta = 0.9$. Because the population standard deviation is known, we use a Z-test for analysis. We then calculate that 22 samples are required (11 controls and 11 treatments) to meet a power of 0.9 for 20% effect size and $\sigma =$ 16 (see http://www.oup-usa.org/sc/0195131878/).

In addition to the challenges involved in choosing biologically meaningful effect sizes (section 2.3.2), this example illustrates similar challenges establishing the relative importance of Type I and Type II errors (α and β, respectively, table 2.1) in power analyses, which we explore in section 2.5.4.

The previous example is somewhat unrealistic because we assumed that the population variance was known in advance, a rare scenario. In real-world prospective analyses, we need methods to estimate variance. Undoubtedly, the preferred method for obtaining a priori estimates of variance for power analysis is to conduct a pilot study. Pilot studies offer a number of other advantages, including the opportunity to test field methods and to train observers. We recommend using

not just the variance estimated from the pilot study, but also the upper and lower confidence limits of the variance estimate to assess the sensitivity of the results and to provide approximate "best case" and "worst case" scenarios (see example 2). A second method for obtaining variances is to use those from similar studies performed previously on the same or similar systems, which can be gathered from colleagues or from the literature. Again, it is useful to repeat the power analyses using the highest and lowest variances available (preferably based on confidence intervals) or values somewhat higher and lower than any single estimate gathered. If variance estimates are obtained from systems other than the study system, the intervals used should be correspondingly wider. Finally, if no previously collected data are available, then the only choice is to perform power analyses using a range of plausible values of variance and hope that these encompass the true value.

Example 2. Sample sizes necessary to achieve a specified level of power, where a previous estimate of population variance is used. In almost all cases, the population variance used for prospective analyses will not be known. In example 1, assume the estimated standard deviation is still 16 but was based on a previous study where the sample size was 20. Because we are not assuming that the variance is known, we use a *t*-test for analysis. This means that the sample size required given a population standard deviation of 16 will be slightly higher than in example 1—in this case 24 rather than 22 (see http://www.oup-usa.org/sc/0195131878/).

To assess the sensitivity of this result, we can recalculate the required sample size using, for example, 90% confidence limits, 12.63 and 22.15, on the estimate of the standard deviation (see the appendix). These lead to required sample sizes of 14 and 44, respectively. If the population variance in the new experiment is the same as in the previous study, then the probability of obtaining a variance larger than the upper 90% confidence limit is $(1 - 0.9)/2 = 0.05$. Therefore, if we are conservative and use the larger sample size, we have a 95% chance of obtaining the level of power desired. Substituting confidence limits on variance into power calculations in this way leads to exact confidence limits on power for any *t*-test or fixed effect *F*-test (Dudewicz 1972; Len Thomas, 1999, unpublished ms). If instead we are constrained to a maximum sample size of 24, we can use confidence limits on the variance estimate to calculate confidence limits on the expected power given a fixed sample size. Using this approach and $n = 24$, 90% confidence limits on power are 0.72 and 0.98. If a power of 0.72 is not acceptable, then we must use a higher α-level or perhaps reevaluate and increase treatment intensity (and therefore the likely effect size) used in the study.

Most prospective power analyses are more complex than these examples because many components of the research design can be altered, each of which can influence the resulting power of the effort (singly or in combination with other components). In addition, study goals often are not defined so sharply, and power analysis begins as an investigation of what is possible. At this stage, therefore, we recommend producing tables or graphs displaying the interactions among plausible levels of the design components and their effects on power (e.g., figure 2.2). Further, multiple goals or hypotheses are usually considered in most research

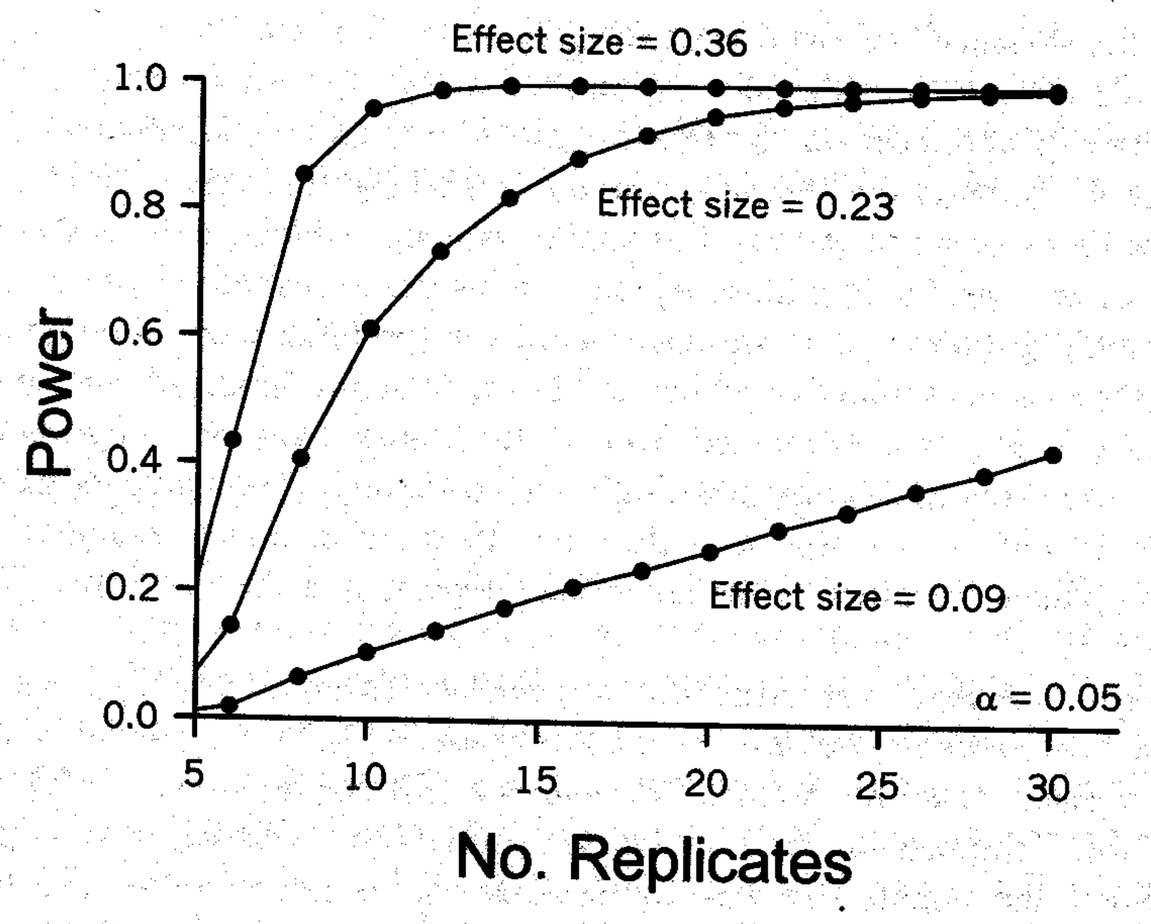

Figure 2.2 The influence of number of replicates on statistical power to detect small (0.09), medium (0.23), and large (0.36) effect sizes (differences in the probability of predation) between six large and six small trout using a Wilcoxon signed-ranks test. Power was estimated using a Monte Carlo simulation.

designs. This entails multiple, related power analyses, and the consideration of the relative importance of these different goals. Finally, study design is usually more complex than comparing samples from two populations. In these cases, considering alternative possible designs is important, as power often can be increased within a fixed budget and sample size by imaginative design decisions (e.g., Steidl et al. 1997). These include increasing the likely effect size by increasing the range or intensity of treatment levels, reducing experimental error by blocking (chapter 4) or measuring covariables, and selecting an efficient method of assigning treatments to experimental units (usually the number of replicates should be highest in treatment combinations where the variance is expected to be highest). Using a statistical model for data analysis that is consistent with the design can also have a strong influence on power (Hatfield et al. 1996; Steidl et al. 1997). These and other techniques for increasing efficiency are discussed in texts on experimental and sampling design (e.g., Thompson 1992; Kuehl 1994).

2.3.2 Retrospective Power Analyses

Retrospective power analyses are performed after a study has been completed and the data analyzed. At this point, the outcome of the statistical test is known: either the null hypothesis was rejected or it was not. If it was not rejected, we may be concerned with committing a Type II error if the statistical power of the test was low. At this point, all the information necessary to calculate power is available,

the design, sample size, and α-level used are known, and the effect size and variance observed in the sample provide estimates of the effect size and variance in the population. Whether any of this information is appropriate for estimating retrospective power is controversial (Thomas 1997). Some researchers believe that the whole concept of retrospective power is invalid within the hypothesis-testing framework (Goodman and Berlin 1994; Zumbo and Hubley 1998; Gerard et al. 1998; Hoenig and Heisey 2001). Others feel that it is informative if the effect size observed is not used in the calculations (Rotenberry and Weins 1985; Cohen 1988; Peterman 1990; Muller and Benignus 1992; Searcy-Bernal 1994; Steidl et al. 1997; Thomas 1997). Although we believe that retrospective power analyses have a place, we favor alternative approaches such as the use of confidence intervals, for reasons discussed here and in section 2.5.2.

Retrospective power is a concern most often when a statistical test has failed to provide sufficient evidence to reject the null hypothesis. In this case, we wish to distinguish between the two reasons for failing to reject the null hypothesis: (1) the true effect size was not biologically important and therefore the null hypothesis was true or nearly true, and (2) the true effect size was biologically important but we failed to reject the false null hypothesis (i.e., we committed a Type II error). To make this distinction, we can calculate the power to detect a minimum biologically important effect, given the sample size, α-level used, and variance estimated in the study. If power at this effect size is large, then true effect sizes of the magnitude of the minimum biologically important effect would likely lead to statistically significant results. Given that the test was not significant, we can infer that the true effect size is likely not this large. Using similar logic, if power at this effect size is small, we can infer that the true effect size could be large or small, so the results are inconclusive.

Example 3. Retrospective power analysis. In an attempt to explain regional differences in reproductive success of ospreys (*Pandion haliaetus*), thickness of eggshells (an indicator of organochlorine contamination) was compared between a colony with low reproduction and one with normal reproduction (Steidl et al. 1991). A two-tailed, two-sample *t*-test comparing mean thickness between colonies yielded $t_{49} = 1.32$, $P = 0.19$, which is not statistically significant at any reasonable α-level (low reproduction: $\bar{y} = 0.459$ mm, $n = 10$; normal reproduction: $\bar{y} = 0.481$ mm, $n = 41$; pooled $s = 0.0480$).

In this case, failing to detect a biologically important difference in eggshell thickness could lead to incorrect conservation decisions. One way to address the possibility that a biologically important difference existed but was not detected by the statistical test is through retrospective power analysis. This raises the question of what comprises a biologically important difference. In this case, assume that previous research has suggested that a 5% reduction is eggshell thickness would likely impair reproduction. This would translate into an absolute difference of 0.024 mm (0.481×0.05), which gives an estimated power of 0.29 for a two-tailed *t*-test using the observed value of s and $\alpha = 0.05$, with 95% confidence limits of 0.20 to 0.39 (see http://www.oup-usa.org/sc/0195131878/). Of course, power to detect a larger 10% difference (0.048 mm) in eggshell thickness is higher at 0.80, with 95% confidence limits of 0.61 and 0.92.

Another relevant approach is to estimate the minimum detectable effect size for a given level of power, which is the minimum effect size that would have yielded $P \leq \alpha$. In this example (5% reduction) and with $\alpha = 0.05$, $s = 0.048$, and power = 0.8, the minimum detectable eggshell thickness is 0.048 mm (95% confidence limits of 0.040 and 0.060). Similarly, you could also estimate the sample size that would have been necessary to detect the observed effect size. In this example, the sample size necessary to detect the observed effect size (0.481 – 0.459 = 0.022 mm) would have been 128 (approximate 95% confidence limits of 90 and 197) (see http://www.oup-usa.org/sc/0195131878/).

Although the use of retrospective power analysis when the null hypothesis is not rejected has been recommended broadly, a number of problems are commonly ignored. First, we assume implicitly that the estimate of power at a given effect size (or effect size for a given power) can be translated into a statement of confidence about the true effect size. For example, "given the null was not rejected and that retrospective power for effect size x is $1 - \alpha$, then we have at least $(1 - \alpha)100\%$ confidence that the interval $(-x, x)$ contains the true effect size." However, such a statement has never been justified formally (Hoenig and Heisey 2001). Second, performing retrospective power calculations only when the null hypothesis is not rejected compromises these analysis. Third, confidence intervals about the estimates of power or the detectable effect size are conservative (i.e., too wide), although there are methods for correcting them (Muller and Pasour 1997). Fourth, because retrospective power calculations do not use information about the observed effect size, they are inefficient compared to the inferences that can be drawn using standard confidence intervals about the estimated effect size (section 2.5.2). Because of these problems, we believe that estimating power retrospectively is rarely useful, and instead we recommend the use of confidence intervals about estimated effect size.

One situation in which retrospective power analysis is never helpful is when power is estimated with the effect size observed in the study (sometimes called the observed power). The calculated value of power is then regarded as an estimate of the "true" power of the test, i.e., the power given the underlying population effect size. Such calculations are uninformative and potentially misleading (Steidl et al. 1997; Thomas 1997; Gerard et al. 1998). First, they do not take into account the biological significance of the effect size used. Second, the observed power estimates are simply a reexpression of the P-value: low P-values lead to high power and vice versa. Third, even as estimates of "true" power, they are biased and imprecise.

2.4 Statistical Solutions: Calculating Power

2.4.1 Power Analysis Using Standard Tables or Software

Power can be estimated for common statistical tests using tables or figures in statistics texts (e.g., Rao 1998; Zar 1996) or specialized monographs (Kraemer and Thiemann 1987; Cohen 1988; Lipsey 1990). This approach can provide an

easy way to obtain quick, approximate results but is not ideal for an in-depth study of power for two reasons. First, estimates of power and related parameters (such as minimum detectable effect size) often are inaccurate, either because they must be interpolated from tabulated values or read from a graph, or in some cases because the tabulated values are themselves based on approximations (Bradley et al. 1996). Second, an in-depth study requires calculating the desired statistics at many levels of the other parameters and graphing the results, which is laborious if done by hand.

Alternatively, a growing number of computer programs perform power analysis (http://www.oup-usa.org/sc/0195131878/). These range from "freeware" programs to large, relatively sophisticated commercial packages. Further, some general-purpose statistical and spreadsheet software packages have built-in power analysis capabilities or add-on modules or macros (e.g., the SAS module UnifyPow; O'Brien 1998).Thomas and Krebs (1997) performed a detailed review of 29 programs, comparing their breadth, ease of learning, and ease of use. Although their specific recommendations about packages will become increasingly outdated as new software is released, the criteria they used and their general comments remain relevant. They considered an ideal program to be one that (1) covers the test situations most commonly encountered by researchers; (2) is flexible enough to deal with new or unusual situations; (3) produces accurate results; (4) calculates power, sample size, and detectable effect size; (5) allows easy exploration of multiple values of input parameters; (6) accepts a wide variety of measures of effect size as input, both raw and standardized; (7) allows estimation of sampling variance from pilot data and from the sampling variability statistics commonly reported in the literature; (8) gives easy-to-interpret output; (9) produces presentation-quality tables and graphs for inclusion in reports; (10) allows easy transfer of results to other applications; and (11) is well documented. They recommended that beginner to intermediate users consider the specialized commercial power analysis programs nQuery Advisor, PASS, or Stat Power, whereas those on a budget try some of the freeware packages such as GPower and PowerPlant (see http://www.oup-usa.org/sc/0195131878/ for an up-to-date list of available software).

2.4.2 Programming Power Analysis Using General-purpose Statistical Software

Most statistical tests performed by ecologists are based on the Z-, t-, F-, or χ^2-distributions. Power analysis for these tests can be programmed in any general-purpose statistical package that contains the appropriate distribution functions (http://www.oup-usa.org/sc/0195131878/). The advantage of this approach is that power analyses can be tailored exactly to the experimental or sampling design being considered. This is particularly useful for relatively complex designs that are not supported by most dedicated power-analysis software. This approach may be most convenient for those who already own a suitable statistics package.

Programming your own power analyses for the t-, F-, and χ^2-tests requires an understanding of noncentral distributions and noncentrality parameters. The

parametric distributions commonly used for testing are known as *central* distributions, which are special cases of more general distributions called *noncentral* distributions. Whereas central distributions describe the distribution of a statistic under the assumption that the null hypothesis is true, noncentral distributions describe the distribution under any specified alternative hypothesis. Compared to central distributions, noncentral distributions contain one additional parameter, called a *noncentrality parameter*, which corresponds to the relevant measure of effect size. For example, the noncentrality parameter, δ for the noncentral t-distribution, assuming a two-sample t-test, is

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

The exact way in which software packages define noncentrality parameters can vary. As this formula illustrates, the noncentrality parameter can be considered as a measure of standardized effect size (in this case $|\mu_1 - \mu_2|/\sigma$), with an additional term that depends on the way in which sample units are allocated to treatments (O'Brien and Muller 1993). When the noncentrality parameter is zero ($\delta = 0$), noncentral distributions equal their corresponding central distributions. In general, the more false the null hypothesis, the larger the noncentrality parameter (Steiger and Fouladi 1997). Programming power analyses involves translating the measure of effect size used into a noncentrality parameter, then using this value in the appropriate noncentral distribution function. Only central distributions are required for power analysis using Z-tests or random effects F-tests (Sheffé 1959, p. 227). SAS code for the examples in this chapter and other relevant SAS probability functions are provided at http://www.oup-usa.org/sc/0195131878/.

2.4.3 Power Analysis Using Simulation

Sooner or later, we encounter a statistical test for which the previous two approaches are not appropriate. This may be because the test is not covered by tables or accessible software, or because there is no agreed-upon method of calculating power for that test. One example of the second situation is nonparametric tests, where the distribution of the processes producing the data are not fully specified so their distribution under the alternative hypothesis is unknown (see the following example). Specific examples in ecology include analyzing multisite trends (Gibbs and Melvin 1997), modeling predator functional responses (Marshal and Boutin 1999), and assessing trends in fish populations (Peterman and Bradford 1987).

In these situations, power analyses can be performed using stochastic (Monte Carlo) simulations. The approach is simple. First, write a computer routine that mimics the actual experiment, including the analysis. In the plant biomass experiment, for example, the program would use a pseudorandom number generator to create 10 biomass measurements from a normal distribution with mean μ_C and standard deviation σ for controls, and 10 measurements from a normal distribution with mean μ_T and standard deviation σ for treatments. The routine then

would analyze these data using a *t*-test. Second, for each level of the input parameters (in this case μ_C, μ_T, and σ), program the routine to run many times (see subsequent discussion), and tally whether the results were statistically significant for each run. Finally, calculate the proportion of results that were significant. If the computer model is an accurate representation of the real experiment, then the probability of getting a statistically significant result from the model is equal to the probability of getting a statistically significant result in the real experiment, in other words, the statistical power. Hence, the proportion of significant results from the simulation runs is an estimate of power.

The number of simulation runs required depends on the desired precision of the power estimate (table 2.2 and appendix). For precision to one decimal place, 1,000 runs should suffice, whereas for precision to two decimal places, 100,000 runs are necessary.

Example 4. Power analyses by simulation for a nonparametric test. In an experiment investigating the effect of prey size on predation rates, several replicate groups of six large and six small juvenile fish were exposed to a predatory fish; the number of small and large fish depredated was recorded for each group. A Wilcoxon signed-ranks test (a nonparametric equivalent of the one-sample *t*-test) was used to test the null hypothesis that the median difference in the number killed between size classes was zero. Thomas and Juanes (1996) explored the power of this experiment using simulations. They assumed that, within a group, the number of fish killed in each size class was a binomial random variable, and they varied the number of replicate groups (the sample size) and the difference in probability of predation between large and small fish. Their results (figure 2.2) suggested that at least 14 groups were necessary to achieve power of 0.8 given a difference in survival between size classes (effect size) of 0.23.

To simulate experiments analyzed using nonparametric tests, such as the previous one, we must specify fully the data-generating process. In these situations, simulations allow us to explore the power of the experiment under a range of different assumptions. In the example, the probability that a fish is depredated was assumed to be constant within each size class. We could arguably make the model more realistic by allowing probability of predation to vary within groups according to some distribution (for example, the beta distribution). However,

Table 2.2 Dependence of the precision of power estimates from Monte Carlo simulations on the number of simulation runs[a]

Number of simulations	SE ($\hat{\beta}$)	99% CI
100	0.050	0.371–0.629
1 000	0.016	0.460–0.541
10 000	0.005	0.487–0.513
100 000	0.002	0.496–0.504

[a]Calculations are performed at a true power (β) of 0.5 and therefore represent minimum levels of precision (see appendix).

there is always a trade-off between simplicity and realism, and we should be content to stop adding complexity to models when they adequately mimic the features of the experiment and subsequent data that are of particular interest. In the example, the variance of the data generated by the model was similar to that of the experimental data, providing a degree of confidence in the model.

Another related approach is the use of bootstrap resampling (chapter 14) to obtain retrospective power estimates from experimental data. In this approach, many bootstrap data sets are generated from the original data, and the same statistical test is performed on each one. The proportion yielding statistically significant results is an estimate of the power of the test for the given experiment. Unless modified in some way, this approach will estimate power at the observed effect size, which is not useful (section 2.3.2). Therefore, power must be estimated over a range of effect sizes, by adding and subtracting effects to the observed data (e.g., Hannon et al. 1993).

2.5 Related Issues and Techniques

2.5.1 Bioequivalence Testing

There have been numerous criticisms of the hypothesis-testing approach (e.g., Yoccoz 1991; Nester 1996; Johnson 1999; references in Harlow et al. 1997 and Chow 1998). One criticism is that the null hypothesis can virtually never be true and therefore is of no interest. For example, no matter how homogeneous a population, no two samples drawn from the population will be identical if measured finely enough. Consequently, bioequivalence testing was developed in part to counter this criticism (Chow and Liu 1999) and is used commonly in pharmaceutical studies and increasingly in ecological studies (e.g., Dixon and Garrett 1994).

Bioequivalence testing reverses the usual burden of proof, so that a treatment is considered biologically important until evidence suggests otherwise. This is achieved by switching the roles of the null and alternative hypotheses. First, a minimum effect size that is considered biologically important is defined (say, Δ_{crit}). Next, the null hypothesis in stated such that the true effect size is greater than or equal to Δ_{crit}. Finally, the alternative hypothesis is stated such that true effect size is less than Δ_{crit}. The plant biomass experiment discussed previously, for example, could be phrased as:

H_0: $|\mu_T - \mu_C| \geq \Delta_{crit}$, which represents the case where a biologically important effect exists

H_a: $|\mu_T - \mu_C| < \Delta_{crit}$, which represents the case where no biologically important effect exists

In this context, a Type I error occurs when the researcher concludes incorrectly that no biologically important difference exists when one does. This is the type

of error that is addressed by power analysis within the standard hypothesis-testing framework; in bioequivalence testing, this error rate is controlled a priori by setting the α-level of the test. Some have argued that this approach is preferable and eliminates the need for retrospective power analysis (Hoenig and Heisey 2001). However, Type II errors still exist within this framework when the researcher concludes incorrectly that an important difference exists when one does not. If this type of error is a concern, then a retrospective investigation will still be necessary when the null hypothesis is not rejected.

2.5.2 Estimating Effect Sizes and Confidence Intervals

Another criticism of hypothesis testing is that the statistical significance of a test does not reflect the biological importance of the result, because any two samples will differ significantly if measured finely enough. For example, a statistically significant result can be found for a biologically trivial effect size when sample sizes are large enough or variance small enough. Conversely, a statistically insignificant result can be found either because the effect is not biologically important or because the sample size is small or the variance large. These scenarios can be distinguished by reporting an estimate of the effect size and its associated confidence interval, rather than simply reporting a *P*-value.

Confidence intervals can function to test the null hypothesis. When estimated for an observed effect size, a confidence interval represents the likely range of numbers generated from the data that cannot be excluded as possible values of the true effect size with probability $1 - \alpha$. If the $100(1 - \alpha)\%$ confidence interval for the observed effect does not include the value established by the null hypothesis, you can conclude with $100(1 - \alpha)\%$ confidence that a hypothesis test would be statistically significant at level α. In addition, however, confidence intervals provide more information than hypothesis tests because they establish approximate bounds on the likely value of the true effect size. More precisely, on average, $100(1 - \alpha)\%$ confidence intervals will contain the true value of the estimated parameter $100(1 - \alpha)\%$ of the time. Therefore, in situations where the null hypothesis would not be rejected by a hypothesis test, we can use the confidence interval to assess whether a biologically important effect is plausible (figure 2.3). If the confidence interval does not include a value large enough to be considered biologically important, then we can conclude with $100(1 - \alpha)\%$ confidence that no biologically important effect occurred. Conversely, if the interval does include biologically important values, then results are inconclusive. This effectively answers the question posed by retrospective power analysis, making such analyses unnecessary (Goodman and Berlin 1994; Thomas 1997; Steidl et al. 1997; Gerard et al. 1998).

Confidence interval estimation and retrospective power analysis are related but not identical. In the estimation approach, the focus is on establishing plausible bounds on the true effect size and determining whether biologically important effect sizes are contained within these bounds. In power analysis, the focus is on the probability of obtaining a statistically significant result if the effect size were truly biologically important. Despite these differences, the conclusions drawn

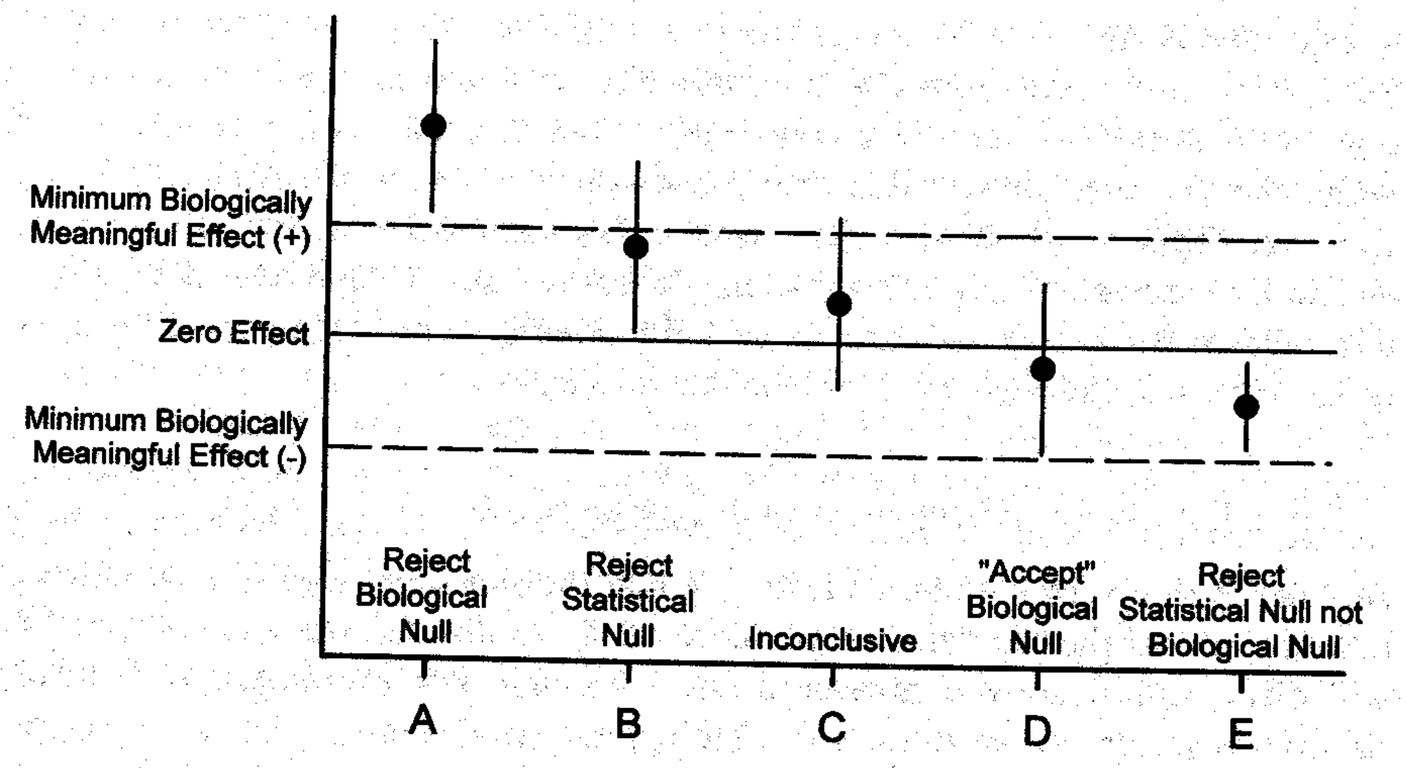

Figure 2.3 Hypothetical observed effects (circles) and their associated $100(1-\alpha)\%$ confidence intervals. The solid line represents zero effect, and dashed lines represent minimum biologically important effects. In case A, the confidence interval for the estimated effect excludes zero effect and includes only biologically important effects, so the study is both statistically and biologically important. In case B, the confidence interval excludes zero effect, so the study is statistically significant; however, the confidence interval also includes values below those thought to be biologically important, so the study is inconclusive biologically. In case C, the confidence interval includes zero effect and biologically important effects, so the study is both statistically and biologically inconclusive. In case D, the confidence interval includes zero effect but excludes all effects considered biologically important, so the "practical" null hypothesis of no biologically important effect can be accepted with $100(1-\alpha)\%$ confidence. In case E, the confidence interval excludes zero effect but does not include effects considered biologically important, so the study is statistically but not biologically important.

from both approaches are often similar. Nevertheless, we prefer the confidence interval approach because interpretation of results is straightforward, more informative, and viewed from a biological rather than probabilistic context.

Example 5. Confidence intervals in lieu of retrospective power. In the osprey eggshell study from example 3, the mean difference in eggshell thickness between regions (the observed absolute effect size) was estimated to be 0.022 mm with a standard error of 0.0169. In the hypothesis-testing approach (example 3), assume we established α at 0.05; we would then use a $100(1-\alpha)=95\%$ confidence interval. The 95% confidence interval on this observed effect size (mean difference) ranges from −0.012 to 0.056 mm. This interval contains the value of 0 predicted by the null hypothesis, so we know the statistical test would not be rejected at $\alpha=0.05$, as we showed previously ($P=0.19$). However, our conclusion about the results of this study will depend on the effect size we consider biologically important. If we consider a relative difference of 10% (0.048 mm) or greater between colonies to be important, then we can consider the results to

be inconclusive because the confidence interval includes this value (figure 2.3). If instead we consider a relative difference of ≥ 20% (0.096 mm) to be important, then we can conclude with 95% confidence that the study showed no important effect because this value is excluded by the confidence interval.

2.5.3 Design Using Confidence Intervals

If the results of a study are to be evaluated using confidence intervals about effect size, then we might design the study to achieve a specified level of precision, or equivalently, a confidence interval of specified width, rather than a desired level of power. For example, we could plan a study to attain a confidence interval that is narrow enough to exclude the null effect size if the true effect size is that which we establish as the minimum to be biologically important. The confidence interval width that we determine is a random variable, however, so there is a 50% chance that it will be wider or narrower than the planned width. Therefore, a conservative approach to design in an estimation context is important (as it is in all aspects of design), and power analysis is a useful tool for this approach (Greenland 1988; Borenstein 1994; Goodman and Berlin 1994).

As mentioned previously, when the realized $100(1-\alpha)\%$ confidence interval excludes the null effect size, this is equivalent to rejecting the null hypothesis at level α. Therefore, the probability that the confidence interval excludes the null effect size, given some specified true effect size, is equal to the power of the test. So, to have a $(1-\beta)$ probability of achieving $100(1-\alpha)\%$ confidence intervals narrow enough to exclude the null hypothesis at a specified true effect size, we must have $(1-\beta)$ power at that effect size.

Example 6. Prospective power analysis for prespecified confidence interval width. We are planning to evaluate the results of the next plant biomass experiment using confidence intervals. As in example 2, we will assume that the variance is not known. For planning purposes, we will base our calculations on a previous study where the estimated standard deviation was 16 with sample size 20. Assume that we wish to have a 90% chance of obtaining 90% confidence limits large enough to exclude zero difference should the true difference be 20% or greater (i.e., ≥ 20.6 kg/ha). Because the confidence limits are symmetric, the desired confidence interval width is therefore $2 \times 20.6 = 41.2$ kg/ha.

This scenario leads to exactly the same power analysis as in example 2: the estimated sample size required is 24, but when we incorporate our uncertainty about the variance estimate, the sample size required is between 14 and 44. Further, calculating the expected confidence interval widths, given the expected variance and sample size, is instructive. With a sample size of 24 and standard deviation of 16, the expected confidence interval width is 13.5 kg/ha. So, we can be 90% sure of achieving a confidence interval width of less than 41.2 kg/ha, but 50% sure that the width will be less than 13.5 kg/ha. As with all prospective design tools, figures displaying how these values change as other factors in the design change prove extremely useful in research planning.

2.5.4 Consequences and Considerations for Establishing α and β

Results of all prospective and retrospective power analyses depend on the levels at which α and β are established. In prospective power analyses, for example, decreasing α (say from 0.10 to 0.05) or increasing the target level of power (say, from 0.7 to 0.9) will always increase the sample sizes necessary to detect a given effect size. As with establishing meaningful effect sizes, choosing these error rates will forever be a challenge.

In general, establishing α and β requires balancing the costs and consequences of Type I and Type II errors (Shrader-Frechette and McCoy 1992; table 2.1). Traditionally, scientists have focused only on Type I errors (hence the impetus for this chapter). However, when there are considerable risks associated with decisions based on the results of hypothesis tests that are not rejected, the consequences of Type II errors often can exceed those of Type I errors (Hayes 1987; Peterman 1990; Steidl et al. 1997). Decisions resulting from hypothesis tests that were not rejected have an underlying, often unrecognized, assumption about the relative costs of Type I and Type II errors that is independent of their true costs (Toft and Shea 1983; Cohen 1988; Peterman 1990). In particular, when $\beta = \alpha$, scientists have decided, perhaps unknowingly, that the costs of Type I errors exceed those of Type II errors when their recommendations assume that a null hypothesis that was not rejected was actually true (i.e., when the null hypothesis was inappropriately accepted). Some have suggested that Type II errors be considered paramount when a decision would result in the loss of unique habitats or species (Toft and Shea 1983; Shrader-Frechette and McCoy 1992). Other approaches have been suggested to balance Type I and Type II error rates based on their relative costs (Osenberg et al. 1994).

As we discussed previously (section 2.5), hypothesis testing has been misused by scientists too often (see also Salsburg 1985, Yoccoz 1991), especially in the context of environmental decision making. Hypothesis tests assess only "statistical significance." The issue of "practical or biological importance" may be better evaluated using confidence intervals (section 2.5.2, although we must still choose the level of confidence to use). We suggest that the reliance on hypothesis testing in decision-making circumstances be decreased in favor of more informative methods that better evaluate available information, including confidence intervals (section 2.5.2), bioequivalence testing (section 2.5.1), Bayesian methods (chapter 17), and decision theory, an extension of Bayesian methods that incorporates the "cost" of making right and wrong decisions (Barnett 1999). Nearly all resource-based decisions are complex, and reducing that complexity to a dichotomous yes–no decision is naïve. Typically, the relevant issue is a not whether a particular effect or phenomenon exists, but whether the magnitude of the effect is biologically consequential. Hypothesis testing should not be the only tool used in decision making, especially when the risks associated with an incorrect decision are considerable. In these instances, knowledge of the potential risks and available evidence for each decision should guide the decision-making process.

2.6 Conclusions

We recommend that researchers evaluate their study design critically before committing to a particular scheme. Only before serious data collection has begun and considerable investment been made can research goals be evaluated freely and details of the experimental design be changed to improve efficiency. We recommend power analysis as a tool for evaluating alternatives in design. This technique forces us to explicitly state our goals (including effect sizes considered biologically important and tolerable levels of error) and make a plan for the analysis of the data—something done far too rarely in practice. In many cases, the power analysis will force us to be more realistic about our goals and perhaps convince us of the need to consult a statistician, either to help us with the power analysis or to help outline the options for study design and analyses. No matter how harsh the realism, the insights gained will save much time and effort. As Sir Ronald Fisher once said, perhaps the most a statistician can do after data have been collected is pronounce a postmortem on the effort.

Acknowledgments Thanks to John Hoenig for sharing his unpublished manuscript and to Ken Burnham, John Hayes, John Hoenig, and Eric Schauber for many insightful discussions. We also thank Sam Scheiner, Jessica Gurevitch, and two anonymous reviewers for their helpful comments on the first draft of this chapter.

Appendix

Here we present formulae and sample calculations for Examples 1–3. SAS code for these examples is available at http://www.oup-usa.org/sc/0195131878/. This appendix also contains the formula for calculating confidence limits for an estimated standard deviation. Confidence intervals on power, required sample size, or minimum detectable effect size can be calculated by substituting the upper and lower confidence limits for standard deviation into the power formulas in the SAS code supplied. Finally, we outline the method used to calculate the precision of power estimates from Monte-Carlo simulation (Table 2.2).

All of the examples in this chapter are of two-tailed two-sample tests, where the test statistic is a Z or t value. In these cases, power is calculated as the probability that the test statistic is greater than or equal to the upper critical value of the appropriate distribution, plus the probability that the test statistic is less than or equal to the lower critical value of the distribution. For a Z-test, power $(1 - \beta)$ is:

$$1 - \beta = (1 - F_z(Z_{1-\alpha/2} - Z_{hyp})) + F_Z(Z_{\alpha/2} - Z_{hyp}) \quad (1)$$

where $F_z(x)$ is the cumulative distribution function of the Z distribution at x, and Z_{hyp} is the $100p$ percentile from the standard normal distribution, calculated as:

$$Z_{hyp} = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{2}$$

where Δ is the specified difference between group means, σ is the pooled standard deviation, and n_1 and n_2 are samples sizes for each group. For a t-test, power is:

$$1 - \beta = (1 - F_t(t_{1-\alpha/2,\nu} \mid \nu,\delta)) + F_t(t_{\alpha/2,\nu} \mid \nu,\delta) \tag{3}$$

where $F_t(x \mid \nu,\delta)$ is the cumulative density function of the noncentral t distribution with ν degrees of freedom and noncentrality parameter δ, evaluated at x, and $t_{p,\nu}$ is the $100p$ percentile from the central t distribution with ν degrees of freedom. The noncentrality parameter δ is calculated as:

$$\delta = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{4}$$

Example 1. Sample sizes necessary to achieve a specified level of power, where population variance is known. Estimate sample sizes ($n = n_1 + n_2$; $n_1 = n_2$) necessary to achieve a specified level of power (1- β) to detect a minimum biologically important difference (Δ) between means of two groups, given α, and the known, pooled standard deviation, σ. Formulas (1) and (2) could be used iteratively using different values of n_1 and n_2, however Zar (1996) provides a more direct formula:

$$n_1 = n_2 = \frac{2\sigma^2(Z_{\alpha/2} + Z_\beta)^2}{\Delta^2}. \tag{5}$$

In this example, $\alpha = \beta = 0.1$, $\Delta = 20.6$ kg/ha, $\sigma = 16$ kg/ha; $Z_{0.1/2} = 1.64$ and $Z_{0.1} = 1.28$. Therefore,

$$n_1 = n_2 = \frac{2 \cdot 16^2(1.64 + 1.28)^2}{20.6^2} = 10.3$$

which indicates than 11 samples (rounding up) would be necessary for each group to meet the specified level of power, yielding a total $n = 22$.

Example 2. Sample sizes necessary to achieve a specified level of power, where a previous estimate of population variance used. Because the pooled standard deviation is estimated rather than known, the t-test rather than the Z-test is appropriate, and (5) cannot be used. Instead, we provide an initial estimate of n required in (4), substitute the calculated noncentrality (δ) into (3), and calculate power. We then continue to adjust our estimate of n until we equal or exceed the level of power specified. For example, beginning with an estimated $n_1 = n_2 = 11$ or $n = 22$ (from example 1), calculate noncentrality using equation (4):

$$\delta = \frac{20.6}{16}\sqrt{\frac{11 \cdot 11}{11 + 11}} = 3.019.$$

To estimate power using equation (3), first calculate degrees of freedom $\nu = n_1 + n_2 - 2 = 20$, $t_{0.95,\,20} = 1.725$, and $t_{0.05,\,20} = -1.725$. Power is then:

$$1-\beta = (1 - F_t(1.725\,|\,20,3.02)) + F_t(-1.725\,|\,20,3.02)$$
$$= 1 - 0.102030 + 0.000002 = 0.897$$

This is slightly less than power of 0.9 specified, so increase the estimate of n to $n_1 = n_2 = 12$, which yields $\delta = 3.154$, $\nu = 22$, $t_{0.95,\,22} = 1.717$, $t_{0.05,\,22} = -1.717$, so power is:

$$1-\beta = (1 - F_t(1.717\,|\,22,3.15)) + F_t(-1.717\,|\,22,3.15)$$
$$= 1 - 0.079325 + 0.000001 = 0.921.$$

Therefore, 12 samples from each group are necessary to meet the specified level of power, yielding a total $n = 24$.

Example 3. Retrospective power analysis. First, estimate power for a hypothesis test already conducted that was not rejected given a minimum biologically important difference in means between two groups ($\Delta = |\mu_1 - \mu_2|$), sample size, α, and an estimate (s) of the pooled standard deviation (σ). In this example, $\Delta = |\mu_1 - \mu_2| = 0.024$ mm, $s = 0.048$, $n_1 = 10$, $n_2 = 41$, and $\alpha = 0.05$.

Calculate an estimate of noncentrality appropriate for the two-tailed, two-sample t-test (4):

$$\hat{\delta} = \frac{0.024}{0.048}\sqrt{\frac{10 \cdot 41}{10 + 41}} = 1.418$$

Calculate degrees of freedom $\nu = n_1 + n_2 - 2 = 49$, $t_{0.975,\,49} = 2.010$, and $t_{0.025,\,49} = -2.010$, and estimate power using equation (3):

$$1-\beta = (1 - F_t(2.010\,|\,49,1.418)) + F_t(-2.010\,|\,49,1.418)$$
$$= 1 - 0.71567 + 0.00040 = 0.285.$$

Minimum detectable effect size (Δ_{mde}) is estimated iteratively. Begin with an arbitrary estimate of detectable effect size, calculate power, then adjust Δ_{mde} until the specified level of power is obtained. For example, begin with an estmate of $\Delta_{mde} = 0.030$ in (4):

$$\hat{\delta} = \frac{0.030}{0.048}\sqrt{\frac{10 \cdot 41}{10 + 41}} = 1.772$$

Calculate degrees of freedom $\nu = n_1 + n_2 - 2 = 49$, $t_{0.975,\,49} = 2.010$, and $t_{0.025,\,49} = -2.010$, and substitute the estimate of δ from above into (3) to calculate power:

$$1-\beta = (1 - F_t(2.010\,|\,49,1.772)) + F_t(-2.010\,|\,49,1.772)$$
$$= -0.58808 + 0.00011 = 0.412$$

which is below the power of 0.9 specified. Therefore, increase the estimate of Δ_{mde} until you determine that the minimum effect size that could have detected was 0.0484.

Finally, the sample size that would have been necessary to detect the observed effect size ($\hat{\Delta}$) at the specified level of power (0.80) is also calculated iteratively. Begin with an arbitrary estimate of sample size, calculate power, then adjust the

estimated sample size until the specified level of power is obtained. For example, begin with $n_1 = n_2 = 30$ or $n = 60$ in (4):

$$\hat{\delta} = \frac{0.024}{0.048}\sqrt{\frac{30 \cdot 30}{30 + 30}} = 1.936.$$

Calculate degrees of freedom $\nu = n_1 + n_2 - 2 = 58$, $t_{0.975,\,58} = 2.001$, and $t_{0.025,\,58} = -2.001$, and substitute the estimate of δ from above into (3) to calculate power:

$$\begin{aligned} 1 - \beta &= (1 - F_t(2.001\,|58,1.936)) + F_t(-2.001\,|58,1.936) \\ &= 1 - 0.52216 + 0.00006 = 0.478 \end{aligned}$$

which is well below the power of 0.9 specified, so we increase the estimate of n until we determine that $n = 128$ (64 per group) were necessary to detect the observed effect size at the level of power specified.

Confidence limits for population standard deviation

The $(1 - \alpha)$ confidence limits for the population standard deviation, based on an estimated standard deviation, s, are given by:

$$\sqrt{\frac{\nu s^2}{\chi^2_{(\alpha/2),\nu}}} \text{ and } \sqrt{\frac{\nu s^2}{\chi^2_{(1-\alpha/2),\nu}}}$$

where ν is the degrees of freedom ($n - 2$ for the examples in this chapter) and $\chi^2_{p,\nu}$ is the $100p$ percentile from a χ^2 distribution with ν degrees of freedom (e.g., Zar 1996, p. 113–115).

Precision of power estimates from Monte-Carlo simulations

Each simulation is assumed to be an independent Bernoulli trial with probability of success, β, equal to the true power of the test. Under these conditions, $SE(\hat{\beta}) = \beta(1 - \beta)/n$, where n is the number of simulations. $SE(\hat{\beta})$ will be at its maximum (and so precision at its minimum) when $\beta = 0.5$.