Chapter 3
Experimental Designs

3.1 Introduction

This chapter covers the fundamentals of experimental design as applied to wildlife studies. Milliken and Johnson (1984) defined *experimental design* as the combination of a design structure, treatment structure, and the method of randomization. We discuss most of the common design and treatment structures currently used in wildlife science from the relatively simple to the more complex. While we touch on sampling (randomization) plans because they are an integral part of experimental design, we delay detailed discussion of sampling until Chap. 4. Data analysis also is integral to study design but we leave this discussion to Chap. 5.

3.2 Principles

The relatively large geographic areas of interest, the amount of natural variability (noise) in the environment, the difficulty of identifying the target population, the difficulty of randomization, and the paucity of good controls make wildlife studies challenging. Wildlife studies typically focus on harvestable species and relatively scarce species of concern (e.g., threatened and endangered species) and factors that influence their abundance (e.g., death, reproduction, and use). In wildlife studies, the treatment is usually a management activity, land use change, or other perturbation contamination event potentially affecting a wildlife population. Additionally, this event could influence populations over an area much larger than the geographic area of the treatment. In most instances, quantification of the magnitude and duration of the treatment effects necessarily requires an observational study, because there usually is not a random selection of treatment and control areas. Early specification of the target population is essential in the design of a study. If investigators can define the target population, then decisions about the basic study design and sampling are much easier and the results of the study can be appropriately applied to the population of interest.

Hurlbert (1984) divided experiments into two classes: mensurative and manipulative. *Mensurative studies* involve making measurements of uncontrolled events at

M.L. Morrison et al., *Wildlife Study Design.*

one or more points in space or time with space and time being the only experimental variable or treatment. Mensurative studies are more commonly termed observational studies, a convention we adopt. *Observational studies* can include a wide range of designs including the BACI, line-transect surveys for estimating abundance, and sample surveys of resource use. The important point here is that all these studies are constrained by a specific protocol designed to answer specific questions or address hypotheses posed prior to data collection and analysis. *Manipulative studies* include much more control of the experimental conditions; there are always two or more treatments with different experimental units receiving different treatments and random application of treatments.

Eberhardt and Thomas (1991), as modified by Manly (1992) provided a useful and more detailed classification of study methods (Fig. 3.1). The major classes in their scheme are studies where the observer has control of events (manipulative experiments) and the study of uncontrolled events. Replicated and unreplicated manipulative experiments follow the classical experimental approach described in most statistics texts. Many of the designs we discuss are appropriate for these experiments. Their other category of manipulative experiment, sampling for modeling, deals with the estimation of parameters of a model hypothesized to represent the investigated process (see Chap. 4).

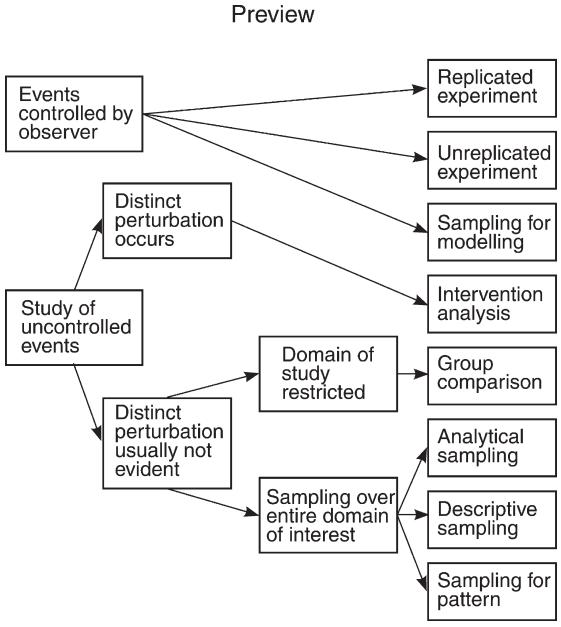

Fig. 3.1 Classification scheme of the types of research studies as proposed by Eberhardt and Thomas (1991) and modified by Manly (1992). Reproduced from Eberhardt et al. (1991) with kind permission from Springer Science + Business Media

The study of uncontrolled events can be broadly classified as observational studies. Observational studies also are referred to as "sample surveys" (Kempthorne 1966), "planned surveys" (Cox 1958), and "unplanned experiments/observational studies" (National Research Council 1985). We suggest Manly (1992) and McKinlay (1975) for additional discussion of the design and analysis of observational studies.

In dealing with observational studies, Eberhardt and Thomas (1991) distinguished between situations where some perturbation occurs and where this is not the case. The study of a perturbation is common in wildlife sciences, such as the study of some environmental contamination (e.g., the Exxon Valdez oil spill). Eberhardt and Thomas called these studies intervention analysis because they typically use time-series (Box and Tiao 1975) methods to study the effect of some distinct event. These *environmental impact studies* typically are large field studies as opposed to manipulative experiments, although manipulative experiments and smaller observational studies aid understanding of the mechanism of impact. In observational studies, data are collected by visual detection of an event in time and space. Many of the basic designs mentioned in this chapter (e.g., BACI) are covered in more detail in Chap. 6.

Eberhardt and Thomas (1991) identified four types of observational studies where no obvious perturbation exists. These studies correspond to investigations designed to develop better understanding of the biology of a system or population. Manly (1992) suggested, and we agree, that the "observational category" of Eberhardt and Thomas is really a special type of observational study where possible observations are limited to selected groups within the entire population of interest. The comparison of groups is another common form of wildlife study, often characterized by the study of representative study areas or groups of animals. The final three classes of study include the possibility of sampling the entire population or area of interest. The point of describing this scheme is that there is a variety of study types, and the design of each will determine the inferences that one can make with the resulting data (Manly 1992).

3.3 Philosophies

Scientific research is conducted under two broad and differing philosophies for making statistical inferences: *design/data-based* and *model-based*. These differing philosophies are often confused but both rely on current data to some degree and aim to provide *statistical inferences*. There is a continuum from strict design/data-based analysis (e.g., finite sampling theory [Cochran 1977] and randomization testing [Manly 1991]) to pure model-based analysis (e.g., global climate models and habitat evaluation procedures [HSI/HEP] using only historical data [USDI 1987]). A combination of these two types of analyses is often employed in wildlife studies, resulting in inferences based on a number of interrelated arguments. For more detailed discussion on design-based and model-based approaches see Chap. 4.

3.3.1 Design/Data-based Analysis

In the analysis of design/data-based studies, basic statistical inferences concerning the study areas or study populations are justified by the design of the study and data collected (Cochran 1977; Scheaffer et al. 1990). Computer-intensive statistical methods (e.g., randomization, permutation testing, etc.) are available that require no additional assumptions beyond the basic design protocol (e.g., Manly 1991). Design/data-based statistical conclusions stand on their own merits for the agreed-upon:

- Response variables
- Procedures to measure the variables
- Design protocol

Reanalysis of the data later does not mean the original statistical inferences were incorrect; instead, the original analysis stands if consensus still exists on the above study conditions.

3.3.2 Model-based Analysis

As the name implies, model-based analyses predict the outcome of experiments using models. In the extreme case of model-based analysis where no new data are available, all inferences are justified by assumption, are deductive, and are subject to counterarguments. The model-based approach usually involves the combination of new data with parameters from the literature or data from similar studies using a theoretical mathematical or statistical model. An example of this approach is the demographic modeling of wildlife populations combined with use of radio-telemetry data to estimate the influence of some perturbation on critical parameters in the model. This approach is illustrated by the telemetry studies of the golden eagle (*Aquila chrysaetos*) (Hunt 1995) in Altamont Pass, California, as described by Shenk et al. (1996).

3.3.3 Mixtures of Design/Data-based and Model-based Analyses

Inferences from wildlife studies often require mixtures of the strict design/data-based and pure model-based analyses. Examples of analyses using mixtures of study designs include:

1. Design/data-based studies conducted on a few target wildlife species
2. Manipulative tests using surrogate species to estimate the effect of exposure to some perturbation on the species of concern (Cade 1994)

3. Deductive professional judgment and model-based analyses used to quantify effects on certain components of the population or habitat in the affected area

Strict adherence to design/data-based analysis in wildlife studies may be impossible, but we recommend that the design/data-based analysis be adhered to as closely as possible. The value of indisputable design/data-based statistical inferences on at least a few response variables cannot be overemphasized in establishing confidence in the overall assessment of treatment effects. However, in some circumstances, model-based methods provide a suitable alternative to design/data-based methods. Additional discussion of the advantages, limitations, and appropriate applications of model-based methods exist in Chap. 4 and in Gilbert (1987), Johnson et al. (1989), and Gilbert and Simpson (1992).

3.4 Replication, Randomization, Control, and Blocking

Fisher (1966) defined the traditional design paradigm for the manipulative experiment in terms of the replication, randomization, control, and blocking, introduced in Chap. 2. Two additional methods are useful for increasing the precision of studies in the absence of increased replication:

1. Group randomly allocated treatments within homogeneous groups of experimental units (blocking)
2. *Use analysis of covariance* (ANCOVA) when analyzing the response to a treatment to consider the added influence of variables having a measurable influence on the dependent variable

3.4.1 Replication

Replication makes statistical inference possible by allowing the estimation of variance inherent in natural systems. Replication also reduces the likelihood that chance events will heavily influence the outcome of studies. In studies of wildlife populations, the experimental unit may be an animal, a group of animals, or all the animals within a specific geographic area. Using the wrong experimental unit can lead to errors in the identification of proper sample sizes and estimates of sample variance.

A good rule to follow when estimating the appropriate sample size in an experiment is that the analysis has only one value from each experimental unit. If five sample plots are randomly located in a study area, then statistical inferences to the area should be based on five values – regardless of the number of animals or plants that may be present and measured or counted in each plot. It becomes obvious that replication is difficult and costly in wildlife studies, particularly when the treatment

is something as unique as an environmental perturbation, such as an oil spill, new wind plant, or dam.

3.4.2 Randomization

Like replication, an unbiased set of *independent data* is essential for estimating the error variance and for most statistical tests of treatment effects. Although truly unbiased data are unlikely, particularly in wildlife studies, a randomized sampling method can help reduce bias and dependence of data and their effects on the accuracy of estimates of parameters. A systematic sample with a random start is one type of randomization (Krebs 1989).

Collecting data from *representative locations* or *typical settings* is not random sampling. If landowners preclude collecting samples from private land within a study area, then sampling is not random for the entire area. In studies conducted on representative study areas, statistical inference is limited to the protocol by which the areas are selected. If private lands cannot be sampled and public lands are sampled by some unbiased protocol, statistical inference is limited to public lands. The selection of a proper sampling plan (see Chap. 4) is a critical step in the design of a project and may be the most significant decision affecting the utility of the data when the project is completed. If the objective of the study is statistical inference to the entire area, yet the sampling is restricted to a subjectively selected portion of the area, then there is no way to meet the objective with the study design. The inference to the entire area is reduced from a statistical basis to expert opinion.

3.4.3 Control and Error Reduction

Replication can increase the *precision* of an experiment (see Chap. 2), although this increased precision can be expensive. As discussed by Cochran and Cox (1957) and Cox (1958), the precision of an experiment can also be increased through:

1. Use of experimental controls
2. Refinement of experimental techniques, including greater sampling precision within experimental units
3. Improvement of experimental designs, including stratification and measurements of nontreatment factors (covariates) potentially influencing the experiment

Good experimental design should strive to improve confidence in cause and effect conclusions from experiments through the *control (standardization) of related variables* (Krebs 1989).

ANCOVA uses information measured on related variables as an alternative to standardizing variables (Green 1979). For example, understanding differences in predator use between areas improves when considered in conjunction with factors

influencing use, such as the relative abundance of prey in each area. These factors are often referred to as *concomitant variables* or *covariates*. ANCOVA combines *analysis of variance* (ANOVA) and *regression* to assist interpretation of data when no specific experimental controls have been used (Steel and Torrie 1980). This analysis method allows adjustment of variables measured for treatment effects for differences in other independent variables also influencing the treatment response variable. ANCOVA assists in controlling error and increasing precision of experiments.

Precision can also be improved using *stratification*, or assigning treatments (or sampling effort) to homogeneous strata, or blocks, of experimental units. Stratification can occur in space (e.g., units of homogeneous vegetation) and in time (e.g., sampling by season). Strata should be small enough to maximize homogeneity, keeping in mind that smaller blocks may increase sample size requirements. For example, when stratifying an area by vegetation type, each stratum should be small enough to ensure a relatively consistent vegetation pattern within strata. Nevertheless, stratification requires some minimum sample size necessary to make estimates of treatment effects within strata. It becomes clear that stratification for a variable (e.g., vegetation type) in finer and finer detail will increase the minimum sample size requirement for the area of interest. If additional related variables are controlled for (e.g., treatment effects by season), then sample size requirements can increase rapidly. Stratification also assumes the strata will remain relatively consistent throughout the life of the study, an assumption often difficult to meet in long-term field studies.

3.5 Practical Considerations

Once the decision is made to conduct a wildlife study, several practical issues must be considered:

1. *Area of interest* (area to which statistical and deductive inferences will be made). Options include the study site(s), the region containing the study sites, the local area used by the species of concern, or the population potentially affected (in this case, population refers to the group of animals interbreeding and sharing common demographics).
2. *Time of interest*. The period of interest may be, for example, diurnal, nocturnal, seasonal, or annual.
3. *Species of interest*. The species of interest may be based on behavior, existing theories regarding species and their response to the particular perturbation, abundance, or legal/social mandate.
4. *Potentially confounding variables*. These may include landscape issues (e.g., large-scale habitat variables), biological issues (e.g., variable prey species abundance), land use issues (e.g., rapidly changing crops and pest control), weather, study area access, etc.

5. *Time available to conduct studies*. The time available to conduct studies given the level of scientific or public interest, the timing of the impact in the case of an accidental perturbation, or project development schedule in the case of a planned perturbation will often determine how studies are conducted and how much data can be collected.
6. *Budget*. Budget is always a consideration for potentially expensive studies. Budget should not determine what questions to ask but will influence how they are answered. Budget will largely determine the sample size, and thus the degree of confidence one will be able to place in the results of the studies.
7. *Magnitude of anticipated effect*. The magnitude of the perturbation or the importance of the effect to the biology of the species will often determine the level of concern and the required level of precision.

The remainder of this chapter is devoted to a discussion of some of the more common experimental designs used in biological studies. We begin with the simplest designs and progress toward the more complex while providing examples of practical applications of these designs to field studies. These applications usually take liberties with Fisher's requirements for designs of true experiments and thus we refer to them as quasiexperiments. Since the same design and statistical analysis can be used with either observational or experimental data, we draw no distinction between the two types of study. Throughout the remainder of this chapter, we refer to treatments in a general sense in that treatments may be manipulations by the experimenter or variables of interest in an observational study.

3.6 Single-factor Designs

Experiments are often classified based on the number of types of treatments that are applied to experimental units. A one-factor experiment uses one type of treatment or one classification factor in the experimental units in the study, such as all the animals in a specific area or all trees of the same species in a management unit. The treatment may be different levels of a particular substance or perturbation.

3.6.1 Paired and Unpaired

The simplest form of a biological study is the comparison of the means of two populations. An unpaired study design estimates the effect of a treatment by examining the difference in the population mean for a selected parameter in a treated and control population. In a paired study design, the study typically evaluates changes in study units paired for similarity. This may take the form of studying a population before and after a treatment is applied, or by studying two very similar study units. For example, one might study the effects of a treatment by randomly assigning

treatment and control designation to each member of several sets of twins or to the right and left side of study animals, or study the effectiveness of two measurement methods by randomly applying each method to subdivided body parts or plant materials.

Comparison of population means is common in impact assessment. For example, as a part of a study of winter habitat use of mule deer (*Odocoileus hemionus*) in an area affected by gas exploration, development, and production, Sawyer et al. (2006) conducted quadrat counts of deer using the winter range from 2001 to 2005 and estimated a 49% decline in deer density after development. As Underwood (1997) points out, this is the classic "before–after" paired comparison where density is estimated before the treatment (gas development) and then compared to density estimates after development. Even though this rather dramatic decline in deer density is of concern, and represents a valid test of the null hypothesis that density will not change after development has occurred, the attribution of the change to development is not supported because of other influences potentially acting on the population. These other potential factors are usually referred to as *confounding* influences (Underwood 1997). In this case, other plausible explanations for the decline in density might be a regional decline in deer density due to weather or a response to competition with livestock for forage. Another approach to designing a study to evaluate the impacts of gas development on this group of deer is to measure density in both a treatment and a control area, where the comparison is the density in two independent groups of deer in the same region with similar characteristics except for the presence (treatment) or absence (control) of gas development.

While there is still opportunity for confounding, and cause and effect is still strictly professional judgment since this is a mensurative study, the presence or absence of a similar decline in the both the treatment and control groups of animals adds strength to the assessment of presence or absence of impact. This example illustrates a common problem in wildlife studies; that is, there is no statistical problem with the study, and there is confidence in not accepting the null hypothesis of no change in density after development. The dilemma is that there is no straightforward way of attributing the change to the treatment of interest (i.e., gas development). Fortunately, for Sawyer et al. (2006), contemporary estimates of habitat use made before and after gas development illustrated a rather clear reduction of available habitat resulting from gas development, which provides support for the conclusion that reduced density may be at least partially explained by development.

Another example of the value of paired comparisons is taken from the Coastal Habitat Injury Assessment (CHIA) following the massive oil spill when the Exxon Valdez struck Bligh Reef in Prince William Sound, Alaska in 1989 – the Exxon Valdez oil spill (EVOS). Many studies evaluated the injury to marine resources following the spill of over 41 million liters of Alaska crude oil. Pairing of oiled and unoiled sites within the area of impact of the EVOS was a centerpiece in the study of shoreline impacts by the Oil Spill Trustees' Coastal Habitat Injury Assessment (Highsmith et al., 1993; McDonald et al., 1995; Harner et al. 1995). In this case, beaches classified in a variety of oiled categories (none, light, moderate, and heavy)

were paired based on beach substrate type (exposed bedrock, sheltered bedrock, boulder/cobble, and pebble/gravel). Measures of biological characteristics were taken at each site (e.g., barnacles per square meter, macroinvertebrates per square meter, intertidal fish, and algae per square meter) and comparisons were made between pairs of sites. The results were summarized as *p*-values (probabilities of observing differences as large as seen on the hypothesis that oiling had no effect) and *p*-values were combined using a meta-analysis approach (Manly 2001).

3.6.2 Completely Randomized Design

The simplest form of an experiment is the random application of two treatments to a group of experimental units known as the *completely randomized design*. This design is possible when experimental units are very similar (homogeneous) so blocking or other forms of partitioning of variance are of little benefit or sample sizes are large enough to be sure there is good representation of the target population in each treatment group. Allocation of treatments is by a random process such that each experimental unit has the same probability of receiving any treatment. Although it is preferable to have equal replication of each treatment across experimental units, it is not necessary.

The completely randomized design is a very flexible design. Analysis is simple and straightforward, allowing comparisons of means of different groups with the simple *t*-test or two or more treatments through ANOVA (Underwood 1997). Nonparametric equivalents of these tests are also readily available. The design maximizes the degrees of freedom (df) for estimating experimental error, increasing precision when df is <20. The loss of information due to missing data is small compared with other, more complicated designs. In addition, one can expand the design with more than two treatments without major alterations to the form of the experiment. The basic model for this design is:

Observed outcome = overall mean + treatment effect + experimental variation.

The completely randomized design is often inefficient, however, since experimental error contains all the variation among experimental units (i.e., measurement error and natural variation). The design may be acceptable for laboratory studies where experimental units are carefully controlled. In field situations, without considerable knowledge of the experimental units or a pretreatment test for differences among experimental units, there is a substantial leap of faith required to assume that experimental units are homogeneous. In the absence of homogeneous experimental units, an effect may be assumed when in reality the apparent treatment effects could actually be the result of pretreatment differences. The best way to deal with this naturally occurring heterogeneity is by true randomization of treatments (Manly 1992) and by maximization of sample size within the context of project goals and practical limitations (e.g., budget). However, as Hurlbert (1984) pointed out, we seldom encounter homogeneous experimental

units in ecological studies and spatial segregation of experimental units can lead to erroneous results resulting from naturally occurring gradients (e.g., elevation and exposure effects on plant growth). This is especially problematic with small sample sizes common in field studies. A systematic selection of experimental units (see Chap. 4) may reduce the effects of spatial segregation of units for a given sample size while maintaining the mathematical properties of randomness. Regardless, the *natural gradients* existing in nature make application of the completely randomized design inappropriate for most field studies.

For a hypothetical example of the completely randomized design, assume the following situation. A farmer in Wyoming is complaining about the amount of alfalfa consumed by deer in his fields. Since the wildlife agency must pay for verified claims of damage by big game, there is a need to estimate the effect of deer use on production of alfalfa in the field. The biologist decides to estimate the damage by comparing production in plots used by deer vs. control plots not used by deer and divides the farmer's largest uniform field into a grid of plots of equal size. A sample of plots is then chosen by some random sampling procedure (see Chap. 4). Deer-proof fence protects half of the randomly selected plots, while the other half is unprotected controls. The effects of deer use is the difference between estimated alfalfa production in the control and protected plots, as measured either by comparing the two sample means by a simple *t*-test or the overall variation between the grazed and ungrazed plots by ANOVA (Mead et al. 1993).

An astute biologist who wanted to pay only for alfalfa consumed by deer could add an additional treatment to the experiment. That is, a portion of the plots could be fenced to allow deer use but exclude rabbits and other small herbivores that are not covered by Wyoming's damage law, without altering the design of the experiment. The analysis and interpretation of this expanded experiment also remains relatively simple (Mead et al. 1993).

In a real world example, Stoner et al. (2006) evaluated the effect of cougar (*Puma concolor*) exploitation levels in Utah. This study used a two-way factorial ANOVA in a completely randomized design with unequal variances to test for age differences among treatment groups (site and sex combinations) for demographic structure, population recovery, and metapopulation dynamics.

3.6.3 Randomized Complete Block Design

While the simplicity of the completely randomized design is appealing, the lack of any restriction in allocation of treatments even when differences in groups of experimental units are known seems illogical. In ecological experiments and even most controlled experiments in a laboratory, it is usually desirable to take advantage of blocking or stratification (see Chap. 4 for discussion) as a form of error control. In the deer example discussed earlier, suppose the biologist realizes there is a gradient of deer use with distance from cover. This variation could potentially bias estimates of deer damage, favoring the farmer if by chance a majority of the plots is near cover or favoring the wildlife agency if a majority of the plots is toward the

center of the field. Dividing the field into strata or blocks and estimating deer use in each may improve the study. For example, the biologist might divide the field into two strata, one including all potential plots within 50 m of the field edge and one including the remaining plots. This stratification of the field into two blocks restricts randomization by applying treatments to groups of experimental units that are more similar and results in better estimates of the effect of deer use, resulting in an equitable damage settlement.

In the experiment where blocking is used and each treatment is randomly assigned within each block, the resulting design is called a *randomized complete block design* (Table 3.1). Blocking can be based on a large number of factors potentially affecting experimental variation. In animal studies, examples of blocks include things such as expected abundance, territoriality, individual animal weights, vegetation, and topographical features. Plant studies block on soil fertility, slope gradient, exposure to sunlight, individual plant parts, or past management. In ecological studies, it is common to block on habitat and across time. This form of grouping is referred to as local control (Mead et al. 1993). The typical analysis of randomized block designs is by ANOVA following the linear additive model

$$\text{Observed outcome} = \text{overall mean} + \text{block effect} + \text{treatment effect} + \text{residual variation} + \text{block} \times \text{treatment interaction}$$

with the block × treatment interaction serving as the error estimate for hypothesis tests.

With proper blocking, no single treatment gains or loses advantage when compared with another because of the characteristics of the units receiving the treatment. If the units within blocks are homogeneous compared to units within other blocks, the blocking reduces the effects of random variation among blocks on the errors involved in comparing treatments. Notwithstanding, poorly designed blocking creates more problems than it solves (see Chap. 4 for a discussion of problems associated with stratification).

Volesky et al. (2005) provide an example of the randomized complete block design to determine the use and herbage production (of cool-season graminoids) in response to spring livestock grazing date and stocking rate in the Nebraska Sandhills. The study used spring grazing date as the main plot, stocking rate as the split plot (see Sect. 3.8.2), with a nongrazed control and grazing rate and stocking rate were factor combinations of treatments. The analysis combined treatments across years with years as fixed effects and blocks as random effects.

Table 3.1 A randomized complete block experiment with four blocks and three treatments (A, B, and C) applied to three plots in each block

Block	Treatment		
1	A	B	C
2	A	C	B
3	B	A	C
4	C	B	A

Reproduced from Morrison et al. (2001), with kind permissions from Springer Science + Business Media

Bates et al. (2005) also used the randomized complete block design in a long-term study of the successional trends following western juniper cutting. This study established four blocks with each block divided into two plots and one plot within each block randomly assigned the cutting treatment (CUT) and the remaining plot left as woodland (WOODLAND). ANOVA was used to test for treatment effect on herbaceous standing crop (functional group and total herbaceous), cover (species and functional group), and density (species and functional group). Cover and density of shrubs and juniper were analyzed by species with response variables analyzed as a randomized complete blocks across time. The final model included blocks (four blocks, df = 3), years (1991–1997 and 2003, df = 7), treatments (CUT, WOODLAND, df = 1), and year by treatment interaction (df = 7; with the error term df = 45).

3.6.4 Incomplete Block Design

A characteristic of the randomized block design discussed earlier was that each treatment was included in each block. In some situations, blocks or budgets may not be large enough to allow all treatments to be applied in all blocks. The *incomplete block design* results when each block has less than a full complement of treatments. In a balanced incomplete block experiment (Table 3.2), all treatment effects and their differences are estimated with the same precision, as long as every pair of treatments occurs together the same number of times (Manly 1992). However, analysis of incomplete block designs is considerably more complicated than complete block designs. It is important to understand the analysis procedures before implementing an incomplete block design. Example design and analysis methods are discussed in Mead et al. (1993).

3.6.5 Latin Squares Design

The randomized block design is useful when one source of local variation exists. When additional sources of variation exist, then the randomized block design can

Table 3.2 A balanced incomplete block experiment with four blocks and four treatments (A, B, C, and D) applied to three plots in each block

Block	Treatment		
1	A	B	C
2	A	B	D
3	A	C	D
4	B	C	D

Note that each treatment pair (i.e., AB, AC, BC, and CD) occurs the same number of times. Reproduced from Morrison et al. (2001), with kind permissions from Springer Science + Business Media

Table 3.3 A Latin square experiment with two blocking factors (X and Y) each with four blocks and four treatments (A, B, C, D)

	Blocking factor (Y)			
Blocking factor (X)	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Reproduced from Morrison et al. (2001), with kind permissions from Springer Science + Business Media

be extended to form a *Latin square* (Table 3.3). For example, in a study of the effectiveness of some treatment, variation may be expected among plots, seasons, species, etc. In a Latin square, symmetry is required so that each row and column in the square is a unique block. The basic model for the Latin square design is as follows:

Observed outcome = row effect + column effect + treatment effect + random unit variation.

The Latin square design allows separation of variation from multiple sources at the expense of df, potentially reducing the ability of the experiment to detect effect. The Latin square design is useful when multiple causes of variation are suspected but unknown. However, caution should be exercised when adopting this design. As an example of the cost of the design, a 3 × 3 Latin square must reduce the mean square error by approximately 40% of the randomized block design of the same experiment to detect a treatment effect of a given size.

While the Latin square design is not a common study design in wildlife studies it can be useful in some situations. For example, with the aid of George Baxter and Lyman McDonald, both professors at the University of Wyoming, the Wyoming Game and Fish Department used the Latin square design on a commercial fisheries project involving carp (*Ctenopharyngodon idella*) in Wyoming. The Department wanted to determine the cause and frequency of "large" year classes and estimated abundance of young fish by different methods at beach sites to help answer this question. The study used three sites, three sampling periods separated by some time to let the fish settle down, and three types of gear (minnow seining, wing traps, and minnow traps). The design was set up in a balanced 3 × 3 Latin square and analysis was by ANOVA. The Latin square takes the form of rows as sites, columns as times, and three gear types with a response variable of p where p = proportion of young of the year fish caught. The Latin square is completed, where each gear type occurs once in each site and time. In addition to estimating the abundance of young fish, the Department was interested in correcting seining data collected elsewhere for biases relative to the "best" sampling method or the pooled proportions if there were significant differences.

3.6.6 Summary

Obviously the different levels of a single treatment in these designs are assumed to be independent and the treatment response assumed to be unaffected by interactions among treatment levels or between the treatment and the blocking factor. This might not present a problem if interaction is 0, an unlikely situation in ecological experiments. Heterogeneity in experimental units and strata (e.g., variation in weather, vegetation, and soil fertility) is common in the real world and results in the confounding of experimental error and interaction of block with treatment effects (Underwood 1997). This potential lack of independence with a corresponding lack of true replication can make interpretation of experiments very difficult, increasing the effect size necessary for significance (increase in Type II error).

3.7 Multiple-factor Designs

3.7.1 Factorial Designs

The preceding designs reduced the confounding effects of variance by blocking under the assumption that the different treatments of a single factor were unique and acted independently. In ecological studies, this independence of treatment effects is seldom encountered. Furthermore, studies usually deal with more than one factor or class of experimental units. Examples of factors include different treatments, such as temperature, diet, habitat, and water quality, or classifications of experimental units, such as season, time of day, sex, age, etc. *Factorial experiments* are more complex experiments where all possible combinations of factors of interest are tested and these tests are possibly replicated a number of times (Manly 1992) and with the resulting data typically analyzed with ANOVA (Underwood 1997).

3.7.2 Two-factor Designs

In a single-factor experiment, there is only one class of treatment. For example, a biologist is interested in the effects of a particular nutrient on the physical condition of deer. The biologist has 24 captive adult deer available for the study. By dividing the deer into three groups of eight deer each and feeding each group a diet with different amounts of the nutrient, the biologist has a single-factor experiment. This study becomes a *two-factor experiment* if the adult deer are divided into six groups of four deer each and a second class of treatment such as two different amounts of forage is added to the experiment. The deer could also be grouped by sex, e.g., three groups of four females and three groups of four males. The three levels of nutrient in the diet

Table 3.4 A 2 × 3 factorial experiment where factor A has three levels (a_1, a_2, and a_3) and factor B has two levels (b_1 and b_2)

		Factor A		
	Level	a_1	a_2	a_3
Factor	b_1	a_1b_1	a_2b_1	a_3b_1
B	b_2	a_1b_2	a_2b_2	a_3b_2

Reproduced from Morrison et al. (2001), with kind permissions from Springer Science + Business Media

and the two amounts of total forage (treatment factors) in the first example and the grouping by sex (classification factor) combined with the three levels of nutrients in the second example both result in a 2 × 3 factorial experiment (Table 3.4).

3.7.3 Multiple-factor Designs

Multiple-factor designs occur when one or more classes of treatments are combined with one or more classifications of experimental units. Continuing the deer feeding experiment, a multiple-factor experiment might include both classes of treatment and the classification of deer by sex resulting in a 2 × 2 × 3 factorial experiment (Table 3.5).

Classification factors, such as sex and age, are not random variables but are fixed in the population of interest and cannot be manipulated by the experimenter. On the other hand, the experimenter can manipulate treatment factors, usually the main point of an experiment (Manly 1992). It is not appropriate to think in terms of a random sample of treatments, but it is important to avoid bias by randomizing the application of treatments to the experimental units available in the different classes of factors. In the example above, a probabilistic sample of female deer selected from all females available for study receive different levels of the treatment.

In the relatively simple experiments with unreplicated single-factor designs, the experimenter dealt with treatment effects as if they were independent. In the real world, one would expect that different factors often interact. The ANOVA of factorial experiments allows the biologist to consider the effect of one factor on another. In the deer example, it is reasonable to expect that lactating females might react differently to a given level of a nutrient, such as calcium, than would male deer. Thus, in the overall analysis of the effect of calcium in the diet, it would be instructive to separate the effects of calcium and sex on body condition (main effects) from the effects of the interaction of sex and calcium. The linear model for the factorial experiment allows the subdivision of treatment effects into main effects and interactions, allowing the investigation of potentially interdependent factors. The linear model can be characterized as follows:

Observed outcome = main effect variable *A* + main effect variable *B*+(*A*)(*B*) interaction + Random unit variation

Table 3.5 An example of a 2 × 2 × 3 factorial experiment where the three levels of a mircronutrient (factor A) are applied to experimental deer grouped by sex (factor B), half of which are fed a different amount of forage (factor C)

Factor B, sex	Factor C, forage level	Factor A, micronutrient		
B_1	c_1	$a_1b_1c_1$	$a_2b_1c_1$	$a_3b_1c_1$
	c_2	$a_1b_1c_2$	$a_2b_1c_2$	$a_3b_1c_2$
B_2	c_1	$a_1b_2c_1$	$a_2b_2c_1$	$a_3b_2c_1$
	c_2	$a_1b_2c_2$	$a_2b_2c_2$	$a_3b_2c_2$

Reproduced from Morrison et al. (2001), with kind permissions from Springer Science + Business Media

Mead et al. (1993) considered this characteristic one of the major statistical contributions from factorial designs.

When interactions appear negligible, factorial designs have a second major benefit referred to as "hidden replication" by Mead et al. (1993). Hidden replication allows the use of all experimental units involved in the experiment in comparisons of the main effects of different levels of a treatment when there is no significant interaction. Mead et al. (1993) illustrated this increase in efficiency with a series of examples showing the replication possible when examining three factors, A, B, and C, each with two levels of treatment:

1. In the case of three independent comparisons, $(a_0b_0c_0)$ with $(a_1b_0c_0)$, $(a_0b_1c_0)$, and $(a_0b_0c_1)$ with four replications for each was possible, involving 24 experimental units. The variance of the estimate of the difference between the two levels of A (or B or C) is $2\sigma^2/4$, where σ^2 is the variance per plot.
2. Some efficiency is gained by reducing the use of treatment $(a_0b_0c_0)$ by combining the four treatments $(a_0b_0c_0)$, $(a_1b_0c_0)$, $(a_0b_1c_0)$, and $(a_0b_0c_1)$ into an experiment with six replications each. Thus, the variance of the estimate of the difference between any two levels is $2\sigma^2/6$, reducing the variance by two-thirds.
3. There are eight factorial treatments possible from combinations of the three factors with their two levels. When these treatments are combined with three replications, each comparison of two levels of a factor includes 12 replicates. All 24 experimental units are involved with each comparison of a factor's two levels. Thus, in the absence of interaction, the factorial experiment can be more economical, more precise, or both, than experiments looking at a single factor at a time.

There is more at stake than simply an increase in efficiency when deciding whether to select a factorial design over independent comparisons. The counterargument for case 1 above is that the analysis becomes conditional on the initial test of interaction, with the result that main effect tests of significance levels may be biased. Perhaps the only situation where example 1 might be desirable is in a study where sample sizes are extremely limited.

Multiple-factor designs can become quite complicated, and interactions are the norm. Although there may be no theoretical limit to the number of factors that can be included in an experiment, it is obvious that sample size requirements increase dramatically as experimental factors with interactions increase. This increases the cost of

experiments and makes larger factorial experiments impractical. Also, the more complicated the experiment is, the more difficulty one has in interpreting the results.

Factorial designs are reasonably common in ecology studies. Mieres and Fitzgerald (2006) used both two-factor and three-factor models in studying the monitoring and management of the harvest of tegu lizards (*Tupinambis* spp.) in Paraguay. The study applied general linear models (two-factor and three-factor ANOVA) to test the null hypothesis of no significant differences in mean size of males and females of each species among years and among check stations. To analyze data from tanneries, they used separate two-factor ANOVAs, with interaction (year and sex as factors), for each species to test the hypothesis that body size varied by year and sex. To test for size variation in tegu skins sampled in the field, the study used three-factor ANOVAs, with interaction (year, sex, and check station as factors), to test the hypothesis that body size varied by year, sex, and check station.

In a study of bandwidth selection for fixed-kernel analysis of animal utilization distributions, Gitzen et al. (2006) used mixtures of bivariate normal distributions to model animal location patterns. The study varied the degree of clumping of simulated locations to create distribution types that would approximate a range of real utilization distributions. Simulations followed a $4 \times 3 \times 3$ factorial design, with factors of distribution type (general, partially clumped, all clumped, nest tree), number of component normals (2, 4, 16), and sample size (20, 50, 150)

3.7.4 Higher Order Designs

The desire to include a large number of factors in an experiment has led to the development of complex experimental designs. For an illustration of the many options for complex designs, the biologist should consult textbooks with details on the subject (e.g., Montgomery 1991; Milliken and Johnson 1984; Mead et al. 1993; Underwood 1997). The object of these more complex designs is to allow the study of as many factors as possible while conserving observations. One such design is a form of the *incomplete block design* known as confounding. Mead et al. (1993) described confounding as the allocation of the more important treatments in a randomized block design so that differences between blocks cancel out the same way they do for comparisons between treatments in a randomized block design. The remaining factors of secondary interest, including those assumed to have negligible interactions are included as treatments in each block, allowing the estimate of their main effects while sacrificing the ability to include their effects on interactions. Thus, block effects are confounded with the effects of interactions. The resulting allocation of treatments becomes an incomplete block with a corresponding reduction in the number of treatment comparisons the experimenter must deal with. Mead et al. (1993) provided two examples that help describe the rather complicated blocking procedure. These complicated designs should not be attempted without consulting a statistician and unless the experimenter is confident about the lack of significant interaction in the factors of secondary interest.

3.8 Hierarchical Designs

3.8.1 Nested Designs

A *nested experimental design* is one that uses replication of experimental units in at least two levels of a hierarchy (Underwood 1997). Nested designs are also known as hierarchical designs and are common in biological studies. Milliken and Johnson (1984) lumped some *nested designs*, *split-plot designs*, and *repeated measures designs* into a category of designs "having several sizes of experimental units." In the earlier discussion of incomplete block experiments, the effects of confounding were dismissed because the experimenter presumably knew that the confounding effects of the interactions of some treatments were negligible. Unfortunately, as we have pointed out, the confounding effects of other variables are all too common in wildlife studies, making the estimation of treatment effects very difficult. Nested studies are a way to use replication to increase one's confidence that differences seen when comparing treatments are real and not just random chance or the effects of some other factor. Nested designs result in data from replicated samples taken from replicated plots receiving each treatment of interest. The only difference in the ANOVA of a nested design from the one-factor ANOVA is that total variation is identified as variation among sample replicates, variation among units (plots) within each treatment, and variation among treatments.

Berenbaum and Zangerl (2006) used a nested study design to study parsnip webworms (*Depressaria pastinacella*) and host plants at a continental scale by evaluating trophic complexity in a geographic mosaic and their role in coevolution. The study used a mixed/nested model (procedure UNIANOVA, SPSS 1999) to compare outcomes of the interaction between wild parsnip (*Pastinaca sativa*) in its indigenous area, Europe, to its area of introduction, the Midwestern United States. The study tested the hypothesis that increasing trophic complexity, represented by alternate host plants or the presence of natural enemies, reduces the selective impact of parsnip webworms and hence diminishes linkage between host plant chemistry and webworms that would be expected in coevolutionary hotspots (areas where webworms were common). The wild parsnip produces a phototoxic compound (furanocoumarins) that crosslink DNA and interfere with transcription in the webworm. Of interest in this study was the concentration of furanocoumarin in parsnip seeds as a function of continent and interaction of temperature and the density of webworms. The study treats the chemical characteristic of parsnip as a random factor nested within both continent and webworm density, and continent and webworm densities as fixed effects.

3.8.2 Split-plot Designs

Split-plot designs are a form of nested factorial design commonly used in agricultural and biological experiments. The study area is divided into blocks following

Table 3.6 An illustration of a two-factor split-plot experiment where factor A is considered at four levels in three blocks of a randomized complete block design and a second factor, B, is considered at two levels within each block

Block 1				Block 2				Block 3			
a_4b_2	a_1b_2	a_2b_1	a_3b_2	a_2b_1	A_1b_2	a_4b_1	a_3b_1	a_1b_1	a_2b_2	a_4b_2	a_3b_1
a_4b_1	a_1b_1	a_2b_2	a_3b_1	a_2b_2	A_1b_1	a_4b_2	a_3b_2	a_1b_2	a_2b_1	a_4b_1	a_3b_2

Note that each unit of factor A is divided into two subunits and randomization occurs for both factor A and factor B. Reproduced from Steel and Torrie (1980), with kind permission from The McGraw-Hill Company

Source: Steel and Torrie (1980)

the principles for blocking discussed earlier. The blocks are subdivided into relatively large plots called main plots, which are then subdivided into smaller plots called split plots, resulting in an incomplete block treatment structure. In a two-factor design, one factor is randomly allocated to the main plots within each block. The second factor is then randomly allocated to each split plot within each main plot. The design allows some control of the randomization process within a legitimate randomization procedure.

Table 3.6 illustrates a simple two-factor split-plot experiment. In this example, four levels of factor A are allocated as if the experiment were a single-factor completely randomized design. The three levels of factor B are then randomly applied to each level of factor A. It is possible to expand the split-plot design to include multiple factors and to generalize the design by subdividing split plots, limited only by the minimal practical size of units for measurements (Manly 1992).

The ANOVA of the split-plot experiment also occurs at multiple levels. At the main plot level, the analysis is equivalent to a randomized block experiment. At the split-plot level, variation is divided into variation among split-plot treatments, interaction of split-plot treatments with main effects, and a second error term for split plots (Mead et al. 1993). A thorough discussion of the analysis of split-plot experiments is presented in Milliken and Johnson (1984). It should be recognized that in the split-plot analysis, the overall precision of the experiment is the same as the basic design.

The split-plot design is useful in experiments when the application of one or more factors requires a much larger experimental unit than for others. For example, in comparing the suitability of different species of grass for revegetation of clear-cuts, the grass plots can be much smaller, e.g., a few square meters, as compared with the clear-cuts that might need to be several acres to be practical. The design can also be used when variation is known to be greater with one treatment vs. another, with the potential for using less material and consequently saving money. The design can be useful in animal and plant studies where litters of animals or closely associated groups of individual plants can be used as main plots and the individual animals and plants used as split plots.

Manly (1992) listed two reasons to use the split-plot design. First, it may be convenient or necessary to apply some treatments to whole plots at the same time. Second, the design allows good comparisons between the levels of the factor that is

applied at the subplot level at the expense of the comparisons between the main plots, since experimental error should be reduced within main plots. However, Mead et al. (1993) pointed out that there is actually a greater loss of precision at the main plot level than is gained at the level of split-plot comparisons. They also indicate that there is a loss of replication in many of the comparisons of combinations of main plot and split treatments resulting in a loss of precision. These authors recommend against the split-plot design except where practically necessary. Underwood (1997) also warned against this lack of replication and the potential lack of independence among treatments and replicates. This lack of independence results because, in most layouts of split-plot designs, main plots and split plots tend to be spatially very close.

Barrett and Stiling (2006) used a split-plot design in a study of Key deer (*Odocoileus virginianus clavium*) impacts on hardwood hammocks near urban areas in the Florida Keys. The study used a split-plot ANOVA model to test each response variable (total basal area of large trees and percentage of canopy cover) with deer density (low and high) and distance (urban and exurban) as factors with island (Big Pine, No Name, Cudjoe, Sugarloaf) nested within levels of deer density. The study found evidence that deer density interacted with distance indicating differences in responses between urban and exurban hammock stands.

3.8.3 Repeated Measures Designs

Experiments where several comparable measurements are taken on each experimental unit are referred to as *repeated measures designs*. Repeated measures experiments are usually associated with nested and split-plot designs. However, repeated measures experiments may occur with the simple completely randomized design or the more complex designs involving blocking, Latin squares, incomplete blocks, or split blocks (Mead et al. 1993). The experimental design structure for the allocation of experimental units, upon which multiple measurements are recorded, must be clearly defined and understood. The significance of the basic design is that it defines the form the analysis must take. In every case, the analysis of repeated measures must consider the lack of independence of the multiple measures taken on the same unit.

Repeated measures usually occur because the experimenter concludes that multiple measurements on an experimental unit increases the knowledge gained in the study. Repeated measures experiments involve a step or steps where there is no randomization of treatment levels. The most common form of repeated measures experiment is the *longitudinal experiment* where observations are taken in the same order on each experimental unit (Manly 1992). Ordering can be a function of any condition that has an order that cannot be changed, but is usually a function of time or location.

A repeated measures experiment where each experimental unit is subjected to several treatments with the order varied for groups of units is called a *changeover*

experiment (Manly 1992). In addition to the lack of independence of repeated measurements, the changeover experiment should consider the carryover effects of treatments on subsequent treatment effects. When possible, the ordering of treatments should be assigned at random to experimental units (Milliken and Johnson 1984).

Longitudinal studies are so common that the term repeated measures often refers only to this type of study (Manly 1992). Longitudinal studies are common in wildlife telemetry studies, environmental impact studies, habitat use and selection studies, studies of blood chemistry, and many other forms of wildlife research. Typically, in these studies, the logistics leads to a repeated measures experiment. For example, after going to the trouble and expense of capturing and radio-tagging an elk, deer, or golden eagle, a biologist correctly takes the position that taking repeated measurements of the marked animal improves the study. Nevertheless, it must be recognized that repeated measures on the same animal may improve the understanding of that animal's behavior but do not represent true replication leading to a better understanding of all animals. The appropriateness of repeated measures experiments is determined by the goal of the study. The biologist must guard against treating repeated measures as true replications and thus leading to what Hurlbert (1984) described as pseudoreplication.

The analysis of data from longitudinal experiments usually follows the same form as analysis of split-plot experiments. The only apparent difference between the longitudinal experiment and the split-plot design is that time, typically a split-plot factor, is beyond the control of the experimenter. Mead et al. (1993) provided a description of the more common approaches to the analysis of these studies. Manly (1992) listed the following analysis methods:

1. ANOVA of a summary variable, such as the mean of repeated measures among units or the difference between the first and last observations (Mead et al. 1993)
2. ANOVA of a response function fitted to observations on experimental units to see how they vary with the factors that describe the groups (Mead et al. 1993)
3. Multivariate ANOVA taking into account the lack of independence of observations taken at different times (Winer 1971)
4. ANOVA in the same way as with a split-plot experiment (Manly 1992)

Mead et al. (1993) pointed out that the form of the model chosen for analysis must always be biologically reasonable, recognizing that models will always require simplification of reality. When the form of analysis follows the split-plot design, there is a general assumption that observations on the same unit at different times will tend to be more similar than observations on different units. There is also the assumption that differences in observations on the same unit are independent of the time when the observations are made. That is, the first and last observations should have the same degree of similarity as consecutive observations, a seldom-valid assumption. This assumption of uniform similarity, also known as compound symmetry may be justified by the random allocation of treatment levels to split plots but should be formally tested (Manly 1992). Milliken and Johnson (1984) provided a detailed discussion of alternative models for analysis of repeated measures experiments

when the assumption of compound symmetry is appropriate. Several statistical software programs (e.g., SAS Institute, Inc.) automatically test for compound symmetry and provide alternate models when the assumption is not met.

Mead et al. (1993) suggested that the split-plot analysis oversimplifies the true situation in its assumption of uniform similarity between times and fails to use the ordering of time. They point out that multivariate ANOVA makes no assumptions about patterns of similarity of observations at different times. Multivariate ANOVA estimates the relationships between times wholly from the data, ignoring the order of times (Crowder and Hand 1990). However, Underwood (1997) maintained that the *multivariate analysis* deals with only one of the problems with the analysis, namely the nonindependence among times of sampling. The other problem of lack of replication leading to unverifiable assumptions of no interaction between times of sampling and replicated plots in each treatment is not addressed. Underwood (1997) took the relatively hard line that proper independent replicates should be the design of choice unless interactions that must be assumed to be 0 are realistically likely to be 0. Obviously, repeated measures are likely to continue in wildlife studies.

Our best advice is to consider the implications of the nonindependence of the data when interpreting the meaning of the studies. The experimenter should consult a good reference, such as Crowder and Hand (1990), when considering repeated measures experiments or, better yet, consult with a statistician experienced in dealing with the design and analysis of such studies.

The repeated measures study design is one of the most common designs in wildlife studies, particularly in the evaluation of the impacts of management or environmental perturbations. An example of a repeated measures study design is provided by Martin and Wisley (2006) in their assessment of grassland restoration success as influenced by seed additions and native ungulate activities. The study used a randomized complete block split-plot design with unequal replication, with grazing or enclosures applied to main plots and seed addition treatments applied to subplots. The statistical analysis used randomized split-plot ANOVAs, with planting as a random block term; all grazing effects were tested with the main plot error term. A repeated measures ANCOVA was used to compare grazed and ungrazed plots for existing vegetation and resource variables, with time 0 data (measurements taken before exclosures were constructed) as a covariate and used repeated-measures ANOVA of corresponding data to analyze grazing effects on seedling enhancement over time. The study analyzed the exotic seedling and seedling diversity variables with repeated-measures ANOVA for the first seed addition, and with regular ANOVA for the second seed addition.

3.9 Analysis of Covariance

ANCOVA uses the concepts of *ANOVA* and *regression* (Huitema 1980; Winer et al. 1991; Underwood 1997) to improve studies by separating treatment effects on the response variable from the effects of *confounding variables* (covariates). ANCOVA

can also be used to adjust response variables and summary statistics (e.g., treatment means), to assist in the interpretation of data, and to estimate missing data (Steel and Torrie 1980). It is appropriate to use ANCOVA in conjunction with most of the previously discussed designs.

Earlier in this chapter, we introduced the concept of increasing the precision of studies by the use of ANCOVA when analyzing the response to a treatment by considering the added influence of variables having a measurable influence on the dependent variable. For example, in the study of fatalities associated with different wind turbines, Anderson et al. (1999) recommended measuring bird use and the rotor-swept area as covariates. It seems logical that the more birds use the area around turbines and the larger the area covered by the turbine rotor, the more likely that bird collisions might occur. Figure 3.2 provides an illustration of a hypothetical example of how analysis of bird fatalities associated with two turbine types can be improved by the use of covariates. In the example, the average number of fatalities per turbine is much higher in the area with turbine type A vs. turbine type B. However, when the fatalities are adjusted for differences in bird use, the ratio of fatalities per unit of bird use is the same for both turbine types, suggesting no true difference in risk to birds from the different turbines. Normally, in error control, multiple regression is used to assess the difference between the experimental and control groups resulting from the treatment after allowing for the effects of the covariate (Manly 1992).

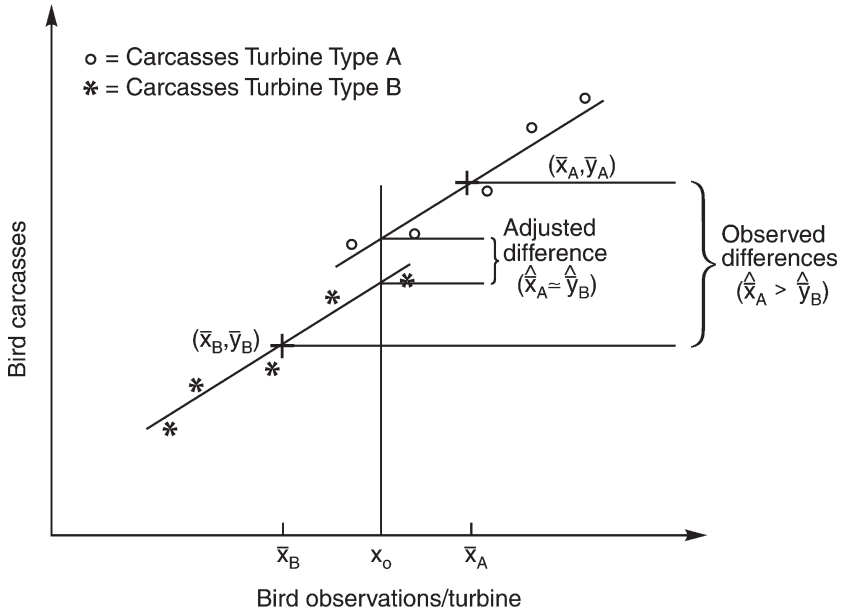

Fig. 3.2 Illustration of hypothetical example of bird fatalities associated with two turbine types (A and B) where the mean fatalities are adjusted for differences in bird use. The average number of fatalities per turbine is much higher associated with turbine type A vs. turbine type B, while the ratio of fatalities per unit of bird use is the same for both turbine types. Reproduced from Morrison et al. (2001) with kind permission from Springer Science + Business Media

ANCOVA adjusts estimates of response variables, such as treatment means. For example, when wildlife surveys record animals by habitat or behavior, these covariates adjust counts to estimate animal numbers more accurately. Strickland et al. (1994) used ANCOVA and logistic regression to adjust aerial counts of Dall's sheep in Alaska. To the authors' surprise, habitat had no effect on sightability but group size was quite important, resulting in significant upward adjustments of counts of individuals and small groups. Surveys of other large mammals (e.g., Gasaway et al. 1985; Samual et al. 1987) suggested that habitat and group size might influence the sightability of sheep. Normally, when using ANCOVA to control error and adjust parameter estimates, the experimenter measures covariates uninfluenced by treatments, such as environmental influences. When covariates are affected by treatments, then their interpretation can be misleading. For example, if one is interested in the effect on animal use in an area by the presence of wind turbines built in different habitats, the study is confounded somewhat because erecting turbines may change habitat characteristics. If this effect is relatively small or the data exist for its estimation, then ANCOVA is still preferable over ignoring the effects of the confounding variables. For example, if the tower pads and roads in the above example are the same size or are carefully measured in all habitats, their effect on bird use can be ignored or accounted for. Although measurements of covariates will have residual error, a violation of one of the necessary assumptions for ANCOVA, Glass et al. (1972) concluded that this is not a serious problem unless residual errors are large.

Manly (1992) also urged caution when using regression adjustment in ANCOVA. He points out that the linear model may be too simple, and a biased estimate of treatment effect may result or important confounding variables may not be measured. As an example, in the wildlife surveys example discussed earlier, we mentioned the propensity for surveys to include environmental covariates, such as habitat and animal behavior. However, it is our experience that variables associated with experimental methods, e.g., the type of aircraft, the experience of the observer, etc., may be far more important in determining the quality of the survey. As with repeated measures, the assumptions inherent in the basic design significantly influence ANCOVA, and good design principles (e.g., randomization and replication) are necessary even with a regression adjustment.

ANCOVA is useful in estimating missing values (Steel and Torrie 1980), and recently, in a model-based analysis of spatial data (e.g., kreiging) discussed in more detail in Chap. 4. The latter application uses the correlations between neighboring sampling units to estimate the variable of interest at points not sampled. Generally, these studies adopt a completely randomized design using a systematic grid of sampling points with a random starting point. Confidence intervals can be calculated for estimates of variables of interest indicating where increased precision is desirable. In environmental contamination studies, these initial samples may be used retrospectively for blocking or stratifying the area of interest so that additional samples can be taken where more precision is desired. We suggest more extensive reading if this form of study is of interest, starting with the summary discussion provided by Borgman et al. (1996).

Flemming et al. (2006) used ANCOVA in their study to test for the effects of embedded lead shot on body condition of common eiders (*Somateria mollissima*).

The assumptions of normality required a log-transformation (using the Andersen–Darling test) of the dependent variable total carcass lipids (TCL) and homogeneity of variances (using Bartlett's test).

Herring and Collazo (2006) used ANCOVA in the study of lesser scaup (*Aythya affinis*) winter foraging and nutrient reserve acquisition in east-central Florida. The study used ANCOVA to examine the effects of season on each of the response variables (CBM, protein, lipids, minerals) for each sex and year separately; in the models, winter period was the treatment and PC1 (first principal component) the covariate to adjust contrasts between season by size of birds.

3.10 Multivariate Analyses

To this point, we dealt with designs that are concerned with the effect of a treatment on one response variable (*univariate methods*). The point of *multivariate analysis* is to consider several related random variables simultaneously, each one being considered equally important at the start of the analysis (Manly 1986). There is a great deal of interest in the simultaneous analysis of multiple indicators (multivariate analysis) to explain complex relationships among many different kinds of response variables over space and time. This is particularly important in studying the impact of a perturbation on the species composition and community structure of plants and animals (Page et al. 1993; Stekoll et al. 1993). Multivariate techniques include multidimensional scaling and ordination analysis by methods such as principal component analysis and detrended canonical correspondence analysis (Gordon 1981; Dillon and Goldstein 1984; Green 1984; Seber 1984; Pielou 1984; Manly 1986; Ludwig and Reynolds 1988; James and McCulloch 1990; Page et al. 1993). If sampling units are selected with equal probability by simple random sampling or by systematic sampling (see Chap. 4) from treatment and control areas, and no quasiexperimental design is involved (e.g., no pairing), then the multivariate procedures are applicable.

It is unlikely that multivariate techniques will directly yield indicators of effect (i.e., combinations of the original indicators) that meet the criteria for determination of effect. Nevertheless, the techniques certainly can help explain and corroborate impact if analyzed properly within the study design. Data from many recommended study designs are not easily analyzed by those multivariate techniques, because, for example,

- In stratified random sampling, units from different strata are selected with unequal weights (unequal probability).
- In matched pair designs, the inherent precision created by the pairing is lost if that pair bond is broken.

A complete description of multivariate techniques is beyond the scope of this book and is adequately described in the sources referenced earlier. Multivariate analysis has intuitive appeal to wildlife biologists and ecologists because it deals simultaneously with variables, which is the way the real world works (see Morrison et al.

2006). However, complexity is not always best when trying to understand natural systems. We think it is worth repeating Manly's (1986) precautions:

1. Use common sense when deciding how to analyze data and remember that the primary objective of the analysis is to answer the questions of interest.
2. The complexity of multivariate analysis usually means that answers that are produced are seldom straightforward because the relationship between the observed variables may not be explained by the model selected.
3. As with any method of analysis, a few extreme observations (outliers) may dominate the analysis, especially with a small sample size.
4. Finally, missing values can cause more problems with multivariate data than with univariate data.

The following are examples of multivariate designs in wildlife studies. Miles et al. (2006) used multivariate models to study the multiscale roost site selection by evening bats on pine-dominated landscapes in southwest Georgia. The study developed 16 a priori multivariate models to describe day-roost selection by evening bats, with pooling data across gender and age classes. Model sets included all possible additive combinations of categories that described tree, plot, stand, and landscape scales. The study used logistic regression to create models and the second-order Akaike's Information Criteria (AIC_c) to identify the most parsimonious model and to predict variable importance. Kristina et al. (2006) evaluated habitat use by sympatric mule and white-tailed deer in Texas using *multivariate analysis of variance* (MANOVA) to test for differences and interactions in habitat composition of home ranges, core areas, among years, and between species for males, and among years, seasons, and species for females. Cox et al. (2006) evaluated Florida panther habitat use using a MANOVA to test the hypothesis that overall habitat selection did not differ from random with sex as a main effect and individual panthers as the experimental unit. The study used the same procedure to test for differences in habitat selection between Florida panthers and introduced Texas cougars. Lanszki et al. (2006) evaluated feeding habits and trophic niche overlap between sympatric golden jackal (*Canis aureus*) and red fox (*Vulpes vulpes*) in the Pannonian ecoregion (Hungary). They used a MANOVA to compare the canids in consumption of fresh biomass of prey based on the prey's body mass as the dependent variable, carnivore species as the fixed factor, and seasons and mass categories as covariates.

3.11 Other Designs

3.11.1 Sequential Designs

It is always desirable to use research dollars and time as efficiently as possible. In the study designs covered so far there is an a priori decision on the number of samples taken and there are two potential statistical inferences, accept or reject the null hypothesis. Sequential designs have been proposed as a way of optimizing

research dollars. Sequential designs are unique in that the sample size is not fixed before the study begins and there are now three potential statistical inferences, accept, reject, or uncertainty (more data are needed). After each sampling event, the available data are analyzed to determine if conclusions can be reached without additional sampling. The obvious advantage to this approach is the potential savings in dollars and time necessary to conclude a study.

Sequential sampling can be very useful when data are essentially nonexistent on a study population and a priori sample size estimation is essentially a guess. As an example, suppose in a regulatory setting the standard for water quality below a waste treatment facility is survival time for a particular fish species (e.g., fathead minnow). The null hypothesis is that mean survival time is less than the regulatory standard and the alternate hypothesis is greater than equal to the regulatory standard. The primary decision criterion is the acceptable risks for Type I and II errors. Typically, in a regulatory setting the emphasis is placed on reducing the Type I errors (i.e., rejecting a true null hypothesis). Sequential sampling continues until a decision regarding whether the facility is meeting the regulatory standard is possible within the acceptable risk of error.

Biological studies commonly use computer-intensive methods (see Manly 1997). Randomization tests, for example, involve the repeated sampling of a randomization distribution (say 5,000 times) to determine if a sample statistic is significant at a certain level. Manly (1997) suggests that a sequential version of a randomization test offers the possibility of reducing the number of randomizations necessary, potentially saving time and reducing the required computing power. Nevertheless, Manly (1997) advocates the use of a fixed number of randomizations to estimate the significance level rather than determining if it exceeds some prespecified level.

The above discussion of the sequential study design presumes there is comprehensive knowledge of the biology of the population of interest. That is, we know which variables are most important, the range of variables that should be studied, the proper methods and metrics to use, and potential interactions. However, the sequential study can also be thought of at a more global scale. That is, an investigation could begin with a moderately sized experiment followed by reassessment after the first set of results is obtained. The obvious advantage to this approach is that the a priori decisions made regarding the biology of populations and the resulting initial study design are modified based on new information. Adaptive resource management (Walters 1986; see Chap. 2) is popularizing this method of scientific study. Box et al. (1978) advocate "the 25% rule," that is not more than one quarter of the experimental effort (budget) should be invested in a first design. The bottom-line is that when there is a great deal of uncertainty regarding any of the necessary components of the study one should not put all of the proverbial eggs (budget and time) into one basket (study).

3.11.2 Crossover Designs

The crossover design is a close relative of the Latin square and in some instances the analysis is identical (Montgomery 1991). Simply put, crossover designs involve

the random assignment of two or more treatments to a study population during the first study period and then the treatments are switched during subsequent study periods so that all study units receive all treatments in sequence. Contrast this with the above designs where treatments are assigned in parallel groups where some subjects get the first treatment and different subjects get the second treatment. The crossover design is typically implemented with a single treatment and control, and represents a special situation where there is not a separate comparison group. In effect, each study unit serves as its own control. In addition, since the same study unit receives both treatments, there is no possibility of covariate imbalance. That is, by assigning all treatments to each of the units crossover designs eliminate effects of variation between experimental units (Williams et al. 2002).

The crossover design can be quite effective when spatially separated controls are unavailable but temporal segregation of treatments is a possibility. However, a key requirement is that the treatments must not have a lasting effect on the study units such that the response in the second allocation of treatments is influenced by the first. This potential for a carry-over effect limits to some extent the type of treatments and study units that can be used in crossover experiments. Typically study units are given some time for recovery (i.e., overcome any potential effects of the first treatment application) before the second treatment phase begins. Williams et al. (2002) describes an analysis procedure that includes a treatment effect, time effect, carry-over effect, and two random terms, one for replication and one that accounts for the sequencing of treatments.

Wolfe et al. (2004) provide a straightforward example of the application of the crossover design in the study of the immobilization of mule deer with the drug Thiafentanil (A-3080). This study utilized a balanced crossover design where each deer was randomly assigned one of two Thiafentanil dose treatments. One treatment was the existing study protocol dose (0.1 mg kg^{-1}), and the other treatment was 2× the protocol dose (0.2 mg kg^{-1}). Treatment assignments were switched for the second half of the experiment so that each animal eventually received both treatments. The first half of the crossover experiment occurred on day 0 of the study and the second half occurred 14 days later to allow the mule deer to recover from the application of the first treatment dose. As another example, a study currently being implemented at the Altamont Pass Wind Resource Area in central California, where a high (>40 per year) number of golden eagles are being killed by wind turbines. The study uses a crossover design to determine if a seasonal shutdown of turbines can be effective in reducing eagle fatalities. A set of turbines are operated during the first half of the winter season while another set is shut down and eagle fatalities are quantified; the on–off turbines are reversed for the second half of the season; and, the same protocol is followed for a second year. The objectives are to see if the overall fatalities in the area decline because of a winter shutdown, to see if winter fatalities decline due to partial shutdown, and to see if variation in fatalities occurs within seasons of operation. Thus, the treatment has been "crossed-over" to the other elements. Power remains low in such experiments, and the experimenter draws conclusions using a weight of evidence approach (where "weight of evidence" simply means you see a pattern in the response).

3.11.3 Quasiexperiments

To this point, we have concentrated on designs that closely follow the principles Fisher (1966) developed for agricultural experiments where the observer can control the events. These principles are the basis for most introductory statistics courses and textbooks. In such courses, there is the implication that the researcher will have a great deal of latitude in the control of experiments. The implication is that experimental controls are often possible and blocking for the partitioning of sources of variance can commonly be used, and the designs of experiments often become quite complicated. The principles provide an excellent foundation for the study of uncontrolled events that include most wildlife studies. However, when wildlife students begin life in the real world, they quickly learn that it is far messier than their statistics professors led them to believe.

Wildlife studies are usually observational with few opportunities for the conduct of replicated manipulative experiments. Studies usually focus on the impact of a perturbation on a population or ecosystem, and fall into the category classified by Eberhardt and Thomas (1991) as studies of uncontrolled events (see Fig. 3.1). The perturbation may be a management method or decision with some control possible or an environmental pollutant with no real potential for control. Even when some control is possible, the ability to make statistical inference to a population is limited. The normal circumstance is for the biologist to create relatively simple models of the real world, exercise all the experimental controls possible, and then, based on the model-based experiments, make subjective conjecture (Eberhardt and Thomas 1991) to the real world.

Regardless of the approach, most of the fundamental statistical principles still apply, but the real world adds some major difficulties, increasing rather than diminishing the need for careful planning. Designing observational studies require the same care as the design of manipulative experiments (Eberhardt and Thomas 1991). Biologists should seek situations in which variables thought to be influential can be manipulated and results carefully monitored (Underwood 1997). When combined with observational studies of intact ecosystems, the results of these experiments increase our understanding of how the systems work. The usefulness of the information resulting from research is paramount in the design of studies and, if ecologists are to be taken seriously by decision-makers, they must provide information useful for deciding on a course of action, as opposed to addressing purely academic questions (Johnson 1995).

The need for quasiexperiments is illustrated by using the existing controversy over the impact of wind power development on birds (Anderson et al. 1999). There is a national desire by consumers for more environmentally friendly sources of energy from so-called “Green Power.” Some industry analysts suggest that as much as 20% of the energy needs in the United States could be met by electricity produced by wind plants. As with most technology development, power from wind apparently comes with a cost to the environment. Early studies of the first large wind resource areas in the Altamont Pass and Solano County areas of California by

the California Energy Commission (Orloff and Flannery 1992) found unexpectedly high levels of bird fatalities. The resulting questions about the significance of these fatalities to the impacted populations were predictable and led to independent research on wind/bird interactions at these two sites and other wind plants throughout the country (Strickland et al. 1998a,b; Anderson et al. 1996; Howell 1995; Hunt 1995; Orloff and Flannery 1992; Erickson et al. 2002). While these studies look at project-specific impacts, the larger question is what these studies can tell us about potential impacts to birds as this technology expands. The study of the impact of wind power on birds is a classic example of the problems associated with study of uncontrolled events.

First, the distribution of wind plants is nonrandom with respect to bird populations and windy sites. Four conditions are necessary for a wind project to be feasible. There must be a wind resource capable of producing power at rates attractive to potential customers. There must be access to the wind. There must be a market for the power, usually in the form of a contract. Finally, there must be distribution lines associated with a power grid in close proximity. Thence, randomization of the treatment is not possible. Wind plants are large and expensive, and sites with favorable wind are widely dispersed. As a result, replication and contemporary controls are difficult to achieve. Nevertheless, public concern will not allow the industry, its regulators, or the scientific community to ignore the problem simply because Fisher's principles of experimental design are difficult to implement.

A second and more academic example of a quasiexperiment is illustrated by Bystrom et al. (1998) in their whole-lake study of interspecific competition among young predators and their prey. Before their study, most research on the issue occurred on a much smaller scale in enclosures or ponds. Bystrom et al. sought to evaluate the effect of competition from a prey fish (roach, *Rutilus rutilus*) on the recruitment of a predatory fish (perch, *Perca fluviatilis*). The study introduced roach to two of four small, adjacent unproductive lakes inhabited by natural populations of perch. After the introduction, the investigators collected data on diet, growth, and survival of the newborn cohorts of perch during a 13-month period. Several complications were encountered, including the incomplete removal of a second and larger predator (pike, *Esox lucius*) in two of the four lakes and an unfortunate die-off of adult perch in the roach-treatment lakes. A second unreplicated enclosure experiment was conducted in one of the lakes to evaluate intraspecific vs. interspecific competition.

Bystrom et al. (1998) attempted to follow good experimental design principles in their study. The problems they encountered illustrate how difficult experiments in nature really are. They were able to replicate both treatment and control environments and blocked treatment lakes. However, the experiment was conducted with a bare minimum of two experimental units for each treatment. They attempted to control for the effects of the pike remaining after the control efforts by blocking. They also attempted to control for intraspecific competition, but with a separate unreplicated study. It could be argued that a better study would have included replications of the enclosure study in some form of nested design or a design that considered the density of perch as a covariate in their blocked experiment. In spite

of a gallant effort, they are left with a study utilizing four subjectively selected lakes from what is likely a very large population of oligotrophic lakes in Sweden and somewhat arbitrary densities of prey and other natural predators. In addition, the two "control" lakes were not true experimental controls and some of the differences seen between the control and treatment conditions no doubt resulted from preexisting differences. It is doubtful that a sample size of two is sufficient replication to dismiss the possibility that differences attributed to the treatment could have occurred by chance. Any extrapolation of the results of this study to other lakes and other populations of perch is strictly a professional judgment; it is subject to the protocols and unique environmental conditions of the original study and is not an exercise of statistical inference.

3.11.3.1 Characteristics of Quasiexperimental Designs

In observational studies of treatment effects, conclusions concerning cause-and-effect relationships are limited. Practically speaking, identical control areas seldom exist and similar *reference* areas must be used instead. Moreover, there is seldom random assignment of treatment, and replication is usually impossible. Oil spills only occur along shipping lanes and power plant sites tend to be unique topographically, geographically, and biologically, and no one would duplicate an oil spill for the sake of science. In the case of an industrial development, where most of the potential construction sites are known, the decision regarding where to locate a new facility never includes a random element in the process. The expense of a new facility or the potential damage caused by a contaminant spill makes replication impractical. Thus, one does not have a true experiment.

Wildlife investigators usually design studies to learn something about some treatment that leads to the prediction of outcomes at unstudied contemporary or future sites with the same or similar treatment (see Sect. 1.2.3.2). For example, from data generated from a probabilistic sample of study plots throughout all oiled areas resulting from an oil spill, the biologist can make statistical (*inductive*) inference to the entire oiled area. The practice of extending the conclusions of wildlife studies beyond the specific study areas to unstudied areas is acceptable, as long as study assumptions are specified and it is clear that the extrapolation is based on expert opinion (*deductive inference*). For example, one can make deductive predictions of the impact of future oil spills in similar areas based on the data from a study of an oil spill. When the extrapolation is presented as an extension of statistical conclusions, it is an improper form of data analysis. In the wind power example, deductive inferences that extend beyond the specific study areas to draw general conclusions about cause-and-effect aspects of operating a wind plant may be possible if enough independent studies of different wind plants identify similar effects. However, *statistical inferences* beyond the study areas are not possible; nor should this be the primary objective of *quasiexperiments*, given the unique aspects of any particular development or ecological inquiry.

3.11.3.2 Examples of Quasiexperimental Designs

The following discussion deals primarily with the study of a distinct treatment or perturbation. These designs fall into the category of *intervention analysis* in Eberhardt and Thomas's (1991) classification scheme. Because these designs typically result in data collected repeatedly over time they are also called an *interrupted time series* (Manly 1992). We do not specifically discuss designs for studies when no distinct treatment or perturbation exists, as these depend on sampling and may be characterized by the way samples are allocated over the area of interest. Sampling plans are covered in detail in Chap. 4.

There are several alternative methods of observational study when estimating the impact of environmental perturbations or the effects of a treatment. The following is a brief description of the preferred designs, approximately in order of reliability for sustaining confidence in the scientific conclusions. A more complete description of these designs can be found in Chap. 6 under the discussion of impact studies and in Manly (1992) and Anderson et al. (1999).

3.11.3.3 Before–After/Control-Impact Design

The *before–after/control-impact* (*BACI*) design is a common design reported in the ecological literature (e.g., Stewart-Oaten 1986), and has been called the "optimal impact study design" by Green (1979). The term *BACI* is so common that we retain the letter *C* in the name, even though we use the term *reference area* rather than *control area*, as true control areas rarely exist. In the BACI design, experimental units are randomly allocated to both treatment and reference areas and populations before the treatment is applied.

The BACI design is desirable for studies of impact or treatment effects because it addresses two major quasiexperimental design problems:

1. Response variables, such as the abundance of organisms, vary naturally through time, so any change observed in a study area between the pretreatment and posttreatment periods could conceivably be unrelated to the treatment (e.g., the construction and operation of a wind plant). Large natural changes are expected during an extended study period.
2. There are always differences in the random variables between any two areas. Observing a difference between treatment and reference areas following the treatment does not necessarily mean that the treatment was the cause of the difference. The difference may have been present prior to treatment. Conversely, one would miss a treatment effect if the levels of the response variable on the reference and treatment areas were the same after the treatment, even though they were different before the treatment.

By collecting data at both reference and treatment areas using exactly the same protocol during both pretreatment and posttreatment periods one can ask the question:

Did the average difference in abundance between the reference area(s) and the treatment area change after the treatment?

The BACI design is not always practical or possible. Adequate reference areas are difficult to locate, the perturbation does not always allow enough time for study before the impact, and multiple times and study areas increase the cost of study. Additionally, alterations in land use or disturbance occurring before and after treatment complicate the analysis of study results. We advise caution when employing this method in areas where potential reference areas are likely to undergo significant changes that potentially influence the response variable of interest. If advanced knowledge of a study area exists, the area of interest is somewhat varied, and the response variable of interest is wide ranging, then the BACI design is preferred for observational studies for treatment effect.

3.11.3.4 Matched Pairs in the BACI Design

Matched pairs of study sites from treatment and reference areas often are subjectively selected to reduce the natural variation in impact indicators (Skalski and Robson 1992; Stewart-Oaten et al. 1986). Statistical analysis of this form of quasiexperiment is dependent on the sampling procedures used for site selection and the amount of information collected on concomitant site-specific variables. For example, sites may be randomly selected from an assessment area and each subjectively matched with a site from a reference area.

When matched pairs are used in the BACI design to study a nonrandom treatment (perturbation), the extent of statistical inferences is limited to the assessment area, and the reference pairs simply act as an indicator of baseline conditions. Inferences also are limited to the protocol by which the matched pairs are selected. If the protocol for selection of matched pairs is unbiased, then statistical inferences comparing the assessment and reference areas are valid and repeatable. For example, McDonald et al. (1995) used this design to evaluate the impacts of the *Exxon Valdez* oil spill on the intertidal communities in Prince William Sound, Alaska. Since the assessment study units were a random sample of oiled units, statistical inferences were possible for all oiled units. However, since the reference units were subjectively selected to match the oiled units, no statistical inferences were possible or attempted to nonoiled units. The selection of matched pairs for extended study contains the risk that sites may change before the study is completed, making the matching inappropriate (see discussion of stratification in Chap. 4). The presumption is that, with the exception of the treatment, the pairs remain very similar – a risky proposition in long-term studies.

3.11.3.5 Impact-Reference Design

The *impact-reference design* quantifies treatment effects through comparison of response variables measured on a treatment area with measurements from one or more reference areas. Studies of the effect of environmental perturbations fre-

quently lack "before" baseline data from the assessment area and/or a reference area requiring an alternative to the BACI, such as the impact-reference design. Assessment and reference areas are censused or randomly subsampled by an appropriate sampling design. Design and analysis of treatment effects in the absence of preimpact data follow Skalski and Robson's (1992) (see Chap. 6) recommendations for accident assessment studies.

Differences between assessment and reference areas measured only after the treatment might be unrelated to the treatment, because site-specific factors differ. For this reason, differences in natural factors between assessment and reference areas should be avoided as much as possible. Although the design avoids the added cost of collecting preimpact data, reliable quantification of treatment effects must include as much temporal and spatial replication as possible. Additional study components, such as the measurement of other environmental covariates that might influence response variables, may help limit or explain variation and the confounding effects of these differences. ANCOVA may be of value to adjust the analysis of a random variable to allow for the effect of another variable.

3.11.3.6 Response-Gradient Design

The *response-gradient design* is useful for quantifying treatment effects in relatively small study areas with homogeneous environments. If the distribution of experimental units is relatively restricted (e.g., small home ranges of passerines) and a response is expected to vary relative to the distance or time from the application of the treatment (gradient of response), this design is an excellent choice for observational studies. When this design is appropriate, treatment effects can usually be estimated with more confidence and associated costs should be less than for those designs requiring baseline data and/or reference areas.

Analysis of the response-gradient design considers the relationship between the response variable and the gradient of treatment levels. For example, in the analysis of an environmental impact, the analysis considers the relationship between the impact indicator and distance from the hypothesized impact source. In effect, the study area includes the treatment area with a reference area on its perimeter. This design does not require that the perimeter of the treatment area be free of effect, only that the level of effect be different. If a gradient of biological response(s) is identified, the magnitude of differences can be presumed to represent at least a minimum estimate of the amount of effect. This response-gradient design would be analogous to a laboratory toxicity test conducted along a gradient of toxicant concentrations. An example might be an increasing rate of fledgling success in active raptor nests or a decrease in passerine mortality as a function of distance to a wind plant.

As in any field study, treatment effects will likely be confounded by the effect of naturally varying factors on response variables. Thus, it is important to have supporting measurements of covariates to help interpret the observed gradient of response. In the example of decreased mortality in passerines associated with a wind plant, an obvious covariate to consider would be level of use of the species of interest.

If one discovers a gradient of response is absent but a portion of the study area meets the requirements of a reference area, data analysis compares the response variables measured in the treatment and control portions of the study area. The impact-gradient design can be used in conjunction with BACI, impact reference, and before–after designs.

3.11.3.7 Before–After Design

The *before–after design* is a relatively weak design, which is appropriate when measurements on the study area before the treatment are compared with measurements on the same area following the treatment. Wildlife managers use long-term monitoring programs to track resources within an area and periodically analyze the resulting data as a before–after designed study. However, observed changes might be unrelated to the treatment, because confounding factors also change with time (see the earlier discussion of the BACI design). Reliable quantification of treatment effects usually include additional study components to limit variation and the confounding effects of natural factors that may change with time.

Because of the difficulty in relating posttreatment differences to treatment effects in the absence of data from reference areas, indirect indicators of treatment effect can be particularly useful in detecting impacts using the before–after design. The correlation of exposure to toxic substances and a physiological response in wildlife has been documented well enough for some substances to allow the use of the physiological response as a *biomarker* for evidence of effect. Examples of biomarkers used in impact studies include the use of blood plasma dehydratase in the study of lead exposure, acetylcholinesterase levels in blood plasma in the study of organophosphates, and the effect of many organic compounds on the microsomal mixed-function oxidase system in liver (Peterle 1991).

Costs associated with conducting the before–after design should be less than those for designs requiring reference areas. Statistical analysis procedures include the time-series method of intervention analysis (Box and Tiao 1975). An abrupt change in the response variable at the time of the treatment may indicate that the response is due to the treatment (e.g., an oil spill) and confidence in this interpretation increases if the response variables return to baseline conditions through time after removal of the treatment. Interpretation of this type of response without reference areas or multiple treatments is difficult and more subjective than the other designs discussed. This type of design is most appropriate for study of short-term perturbations rather than for long-term and ongoing perturbations, such as an industrial development or the study of some persistent contaminant.

3.11.4 Improving Reliability of Study Designs

When studies using reference areas are possible, the use of more than one reference area increases the reliability of conclusions concerning quantification of a treatment

response in all designs (Underwood 1994). Multiple reference areas help deal with the frequently heard criticism that the reference area is not appropriate for the treatment area. Consistent relationships among several reference areas and the treatment area will generate far more scientific confidence than if a single reference area is used. In fact, scientific confidence is likely increased more than would be expected given the increase in number of reference areas. This confidence comes from the *replication in space* of the baseline condition. Multiple reference areas also reduce the impact on the study if one reference area is lost, e.g., due to a change in land use affecting response variables.

Collection of data on study areas for several time periods before and/or after the treatment also will enhance reliability of results. This *replication in time* allows the detection of convergence and divergence in the response variables among reference and treatment areas. The data can be tested for interaction among study sites, time, and the primary indicator of effect (e.g., mortality), assuming the data meet the assumptions necessary for ANOVA of repeated measures. The specific test used depends on the response variable of interest (e.g., count data, percentage data, continuous data, categorical data) and the subsampling plan used (e.g., point counts, transect counts, vegetation collection methods, GIS [Geographic Information System] data available, radio-tracking data, capture–recapture data). Often, classic ANOVA procedures will be inappropriate and nonparametric, Bayesian, or other computer-intensive methods will be required.

3.11.5 Model-based Analysis and Use of Site-Specific Covariates

The conditions of the study may not allow a pure design/data-based analysis, particularly in impact studies. For example, animal abundance in an area might be estimated on matched pairs of impacted and reference study sites. However, carefully the matching is conducted, uncontrolled factors always remain that may introduce too much variation in the system to allow one to statistically detect important differences between the assessment and reference areas. In a field study, there likely will be naturally varying factors whose effects on the impact indicators are confounded with the effects of the incident. Data for easily obtainable random variables that are correlated with the impact indicators (covariates) will help interpret the gradient of response observed in the field study. These variables ordinarily will not satisfy the criteria for determining impact, but are often useful in model-based analyses for the prediction of impact (Page et al. 1993; Smith 1979). For example, in the study of bird use on the Wyoming wind plant site, Western Ecosystems Technology, Inc. (1995) developed indices to prey abundance (e.g., prairie dogs [*Cynomys*], ground squirrels [*Spermophilus*], and rabbits [*Lagomorpha*]). These ancillary variables are used in model-based analyses to refine comparisons of avian predator use in assessment and reference areas. Land use also is an obvious covariate that could provide important information when evaluating differences in animal use among assessment and reference areas and time.

Indicators of degree of exposure to the treatment also should be measured on sampling units. As in the response-gradient design, a clear effect–response

relationship between response variables and level of treatment will provide corroborating evidence of effect. These indicators are also useful with other concomitant variables in model-based analyses to help explain the "noise" in data from natural systems. For example, in evaluating the effect of an oil spill, the location of the site with regard to prevailing winds and currents or substrate of the oiled site are useful indicators of the degree of oil exposure.

3.12 Meta-analyses

A common practice when embarking on a new investigation is to review the literature on the subject and subjectively assess knowledge about the research question of interest. Typically, in the wildlife research field, one finds numerous independent quasiexperiments. For example, if one is interested in the impact of antler restrictions on deer populations, hunting effects on prairie grouse, or herbicide effects on sagebrush, it might be possible to find studies conducted in several states, or even several studies within states. The resulting review of the literature usually produces a subjective evaluation of what all these independent studies mean, and in a sense is a form of *meta-analysis*. Alternatively, the investigator could compare these independent studies statistically in a quantitative meta-analysis.

A number of procedures exist for statistical meta-analysis. Manly (2001) describes two methods for comparing studies by combining the *p*-values from several independent studies (Fisher 1970; Folks 1984) using a chi-square analysis for tests of significance. Fisher's approach is simple and provides a test of whether the null hypothesis is false for any of the studies. However, other methods are more appropriate when addressing the more interesting question usually asked by wildlife scientists; that is, is the null hypothesis generally supported when considering all the studies. One common concern when conducting meta-analysis is the potential variation in studies related to the methods and metrics used, independent of the treatment effects (i.e., are we comparing apples and oranges).

An alternative form of meta-analysis used in medical research involves a statistical analysis of data pooled from independent studies on the response to a particular management action. This approach is appealing, but is most appropriate when study methods and metrics are similar among the studies included in the analysis. In both forms of meta-analysis, the rules for deciding to include or exclude studies are of paramount importance.

Conducting meta-analysis on observational studies, the common form of wildlife study, can be useful, but also controversial because of the inherent variability among studies.

Egger et al. (1998) suggest that while formal meta-analysis of observational studies can be misleading if insufficient attention is not given to heterogeneity, it is

a desirable alternative to writing highly subjective narrative reviews. They make the logical recommendation that meta-analysis of observational studies should follow many of the principles of systematic reviews: a study protocol should be written in advance, complete literature searches carried out, and studies selected and data extracted in a reproducible and objective fashion. Following this systematic approach exposes both differences and similarities of the studies, allows the explicit formulation and testing of hypotheses, and allows the identification of the need for future studies. Particular with observational studies, meta-analysis should carefully consider the differences among studies and stratify the analysis to account for these differences and for known biases.

Erickson et al. (2002) provide a nice example of a meta-analysis using pooled data from a relatively large group of independent observational studies of the impacts of wind power facilities on birds and bats. The meta-analysis evaluated data on mortality, avian use, and raptor nesting for the purpose of predicting direct impacts of wind facilities on avian resources, including the amount of study necessary for those predictions. The authors considered approximately 30 available studies in their analysis of avian fatalities. In the end, they restricted the fatality and use components of the meta-analysis to the 14 studies that were conducted consistent with recommendations by Anderson et al. (1999). They also restricted their analysis to raptors and waterfowl/waterbird groups because the methods for estimating use appeared most appropriate for the larger birds.

Based on correlation analyses, the authors found that overall impact prediction for all raptors combined would typically be similar after collection of one season of raptor use data compared to a full year of data collection. The authors cautioned that this was primarily the case in agricultural landscapes where use estimates were relatively low, did not vary much among seasons, and mortality data at new wind projects indicated absent to very low raptor mortality. Furthermore, the authors recommended more than one season of data if a site appears to have relatively high raptor use and in landscapes not yet adequately studied.

Miller et al. (2003) reviewed results of 56 papers and subjectively concluded that current data (on roosting and foraging ecology of temperate insectivorous bats) were unreliable due to small sample sizes, short-term nature of studies, pseudoreplication, inferences beyond scale of data, study design, and limitations of bat detectors and statistical analyses. To illustrate the value of a quantitative meta-analysis, Kalcounis-Ruppell et al. (2005) used a series of meta-analyses on the same set of 56 studies to assess whether data in this literature suggested general patterns in roost tree selection and stand characteristics. The authors also repeated their analyses with more recent data, and used a third and fourth series of meta-analyses to separate the studies done on bat species that roost in cavities from those that roost in foliage. The quantitative meta-analysis by Kalcounis-Ruppell et al. (2005) provided a much more thorough and useful analysis of the available literature compared to the more subjective analysis completed by Miller et al. (2003).

3.13 Power and Sample Size Analyses

Traditionally in the analysis of an experiment, a null hypothesis (H_0) is the *straw man* that must be rejected to infer statistically that a response variable has changed or that a cause-and-effect relationship exists. The typical H_0 is that there is no difference in the value of a response variable between control areas and assessment areas or that there is a zero correlation between two response variables along their gradients. In the regulatory setting and in impact studies, this approach usually places the burden of scientific proof of impact on regulators.

The classical use of a H_0 protects only against the probability of a Type I error (also called α, concluding that impact exists when it really does not, i.e., a false positive). By convention the significance level is set at $\alpha = 0.05$ before the conclusion of effect is considered to be valid, although there is nothing magic about 0.05. The probability of a Type II error (also called β, concluding no effect when in fact effect does exist, i.e., a false negative) is almost always unknown, commonly ignored and is often much larger than 0.05. At a given α-level, the risk of a Type II error can be decreased by increasing sample size, reducing sampling error, or, in some situations, through use of better experimental design and/or more powerful types of analysis. In general, the power of a statistical test of some hypothesis is the probability that it rejects the H_0 when it is false $1 - \beta$. An experiment is said to be very powerful if the probability of a Type II error is very small.

As Underwood (1997) points out, it makes intuitive sense to design a study to make equal the probability of making either a Type I or II error. However, he introduces the precautionary principle that the willingness to accept a type of error will depend on the nature of the study. For example, in testing drugs or in environmental monitoring it may be more acceptable to commit a Type I error much more often than a type Type II error. Thus, one would want to design a more powerful study to decrease the probability of concluding no effect when one actually exists.

In summary, four interrelated factors determine statistical power: power increases as sample size, α-level, and effect size increase; power decreases as variance increases. Understanding statistical power requires an understanding of Type I and Type II error, and the relationship of these errors to null and alternative hypotheses. It is important to understand the concept of power when designing a research project, primarily because such understanding grounds decisions about how to design the project, including methods for data collection, the sampling plan, and sample size. To calculate power the researcher must have established a hypothesis to test, understand the expected variability in the data to be collected, decide on an acceptable α-level, and most importantly, a biologically relevant response level.

3.13.1 Effect Size

Effect size is the difference between the null and alternative hypotheses. That is, if a particular management action is expected to cause a change in abundance of an

organism by 10%, then the effect size is 10%. Effect size is important in designing experiments for obvious reasons. At a given α-level and sample size, the power of an experiment increases with effect size and, conversely, the sample size necessary to detect an effect typically increases with a decreasing effect size.

Given that detectable effect size decreases with increasing sample size, there comes a condition in most studies that a finding of a statistically significant difference has no biological meaning (for example, a difference in canopy cover of 5% over a sampling range of 30–80%; see Sect. 1.5.3). As such, setting a biologically meaningful effect size is the most difficult and challenging aspect of power analysis and this "magnitude of biological effect" is a hypothetical value based on the researcher's biological knowledge. This point is important in designing a meaningful research project. Nevertheless, the choice of effect size is important and is an absolute necessity before it is possible to determine the power of an experiment or to design an experiment to have a predetermined power (Underwood 1997).

3.13.2 Simple Effects

When the question of interest can be reduced to a single parameter (e.g., differences between two population means or the difference between a single population and a fixed value), establishing effect size is in its simplest form. There are three basic types of simple effects:

- *Absolute effect size* is set when the values are in the same units; for example, looking for a 10 mm difference in wing length between males and females of some species.
- *Relative effect size* is used when temporal or spatial control measures are used and effects are expressed as the difference between in response variable due. As expected, relative effect sizes are expressed as percentages (e.g., the percent increase in a population due to a treatment relative to the control).
- S*tandardized effect sizes* are measures of absolute effect size scaled by variance and therefore combine these two components of hypothesis testing (i.e., effect size and variance). Standardized effect sizes are unit-less and are thus comparable across studies. They are, however, difficult to interpret biologically and it is thus usually preferable to use absolute or relative measures of effect size and consider the variance component separately.

3.13.3 Complex Effects

Setting an effect size when dealing with multiple factors or multiple levels of a single factor is a complex procedure involving and examination of the absolute effect size based on the variance of the population means:

$$\sigma^2 = 1/k \sum (\mu_i - \mu_{\text{mean}})^2.$$

Steidl and Thomas (2001) outlined four approaches for establishing effect size in complex situations:

- *Approach 1*. Specify all cell means. In an experiment with three treatments and a control, you might state that you are interested in examining power given a control value of 10 g and treatment yields of 15, 20, and 25 g. Although these statements are easy to interpret, they are also difficult to assign.
- *Approach 2*. Delineate a measure of effect size based on the population variances through experimenting with different values of the means. That is, you experiment with different values of the response variable and reach a conclusion based on what a meaningful variance would be.
- *Approach 3*. Simplify the problem to one of comparing only two parameters. For example, in a one-factor ANOVA you would define a measure of absolute effect size ($\mu_{\text{max}} - \mu_{\text{min}}$), which places upper and lower bounds on power, each of which can be calculated.
- *Approach 4*. Assess power at prespecified levels of standardized effect size for a range of tests. In the absence of other guidance, it is possible to calculate power at three levels as implied by the adjectives small, medium, and large. This approach is seldom applied in ecological research and is mentioned here briefly only for completeness.

In sum, power and sample size analyses are important aspects of study design, but only so that we can obtain a reliable picture of the underlying distribution of the biological parameters of interest. The statistical analyses that follow provide additional guidance for making conclusions. By setting effect size or just your expectation regarding results (e.g., in an observational study) a priori, the biology drives the process rather than the statistics. That is, the proper procedure is to use statistics to first help guide study design, and later to compliment interpretations. The all too common practice of collecting data, applying a statistical analysis, and then interpreting the outcome misses the needed biological guidance necessary for an adequate study. What you are doing, essentially, is agreeing a priori to accept whatever guidance the statistical analyses provide and then trying to force a biological explanation into that framework. Even in situations where you are doing survey work to develop a list of species occupying a particular location, stating a priori what you expect to find, and the relative order by abundance, provides a biological framework for later interpretation (and tends to reduce the fishing expedition mentality).

Sensitivity analysis can be used to help establish an appropriate effect size. For example, you can use the best available demographic information – even if it is from surrogate species – to determine what magnitude of change in, say, reproductive success will force λ (population rate of increase) above or below 1.0. This value then sets the effect size for prospective power analysis or for use in guiding an observational study (i.e., what difference in nest success for a species would be of interest when studying reproduction along an elevation gradient?). For a primarily

observational study, there will usually be information – sometimes qualitative – on the likely distribution and relative abundance of the element of interest (e.g., previous studies, field guides and natural history reports, expert opinion).

3.14 Prospective Power Analysis

A primary defense against weak tests of hypotheses is to perform a prospective power analysis at the start of the research, hopefully following a pilot study (Garton et al. 2005). The first step in the prospective power analysis is to decide on the null hypothesis, alternate hypothesis, and significance level before beginning the investigation (Zar 1998). Power analysis can be used to help make a decision regarding the necessary sample size, or at least inform the investigator of the chances of detecting the anticipated effect size with the resources available. Zar (1998) is a useful reference for methods for estimating the required sample size for most common sampling and experimental designs.

Prospective power analysis is used to:

- Determine the number of replicates or samples necessary to achieve a specified power given the specified effect size, alpha, and variance (scenario 1)
- The power of a test likely to result when the maximum number of replicates or samples that you think can be obtained are gathered (scenario 2)
- The minimum effect size that can be detected given a target power, alpha, variance, and sample size (scenario 3)

Below, we discuss each of these topics as applied to ecological field research:

1. *Scenario 1*. In this scenario you are able to specify the effect size, set α (an easy task relative to setting effect size), and estimate the population variance. We have previously discussed how to establish effect size. Estimating the population variance can be accomplished either through previous work on the element of interest (pilot test or existing literature), or by using estimates from a similar element (e.g., congeneric species). Remember that power analysis is used to provide a starting point for research and is not intended to set a final sample size. Thus, using a range (min, max) of estimates for population variance provides you with a method to estimate what your sample size should be, given the effect size and alpha you have selected.

   If you determine you cannot achieve the desired number of samples using α and effect size you initially selected, then your primary option is to change α and effect size; variance can seldom be modified. Many papers are available that discuss selection of α; we do not review them here. Effect size can be modified, but remember that you must be able to justify the effect size that you set in the publication that follows the research.
2. *Scenario 2*. Here you are asking what you can achieve given the available sampling situation. This is often the situation encountered in wildlife research where

a funding entity (e.g., agency) has developed a request for a study (i.e., Request for Proposal or RFP) that includes a specific sampling location(s), sampling conditions, and a limit to the amount of funding available. By accepting such funding, you are in essence accepting what the resulting power and effect size. You often have the ability, however, to adjust the sampling protocol to ensure that you can address at least part of the study objectives with appropriate rigor. In this scenario you conduct power analysis in an iterative manner using different effect sizes and α-levels to determine what you can achieve with the sample size limits in place (Fig. 3.3).

3. *Scenario 3*. Here you are determining what effect size you can achieve given a target power, α-level, variance, and sample size. As discussed earlier, α can be changed within some reasonable bounds (i.e., a case can usually be made for ≤ 0.15) and variance is set. Here you also are attempting to determine what role sample size has in determining effect size.

In summary, the advantage of prospective power analysis is the insight you gain regarding the design of your study. Moreover, even if you must conduct a study given inflexible design constraints, power analysis provides you with knowledge of the likely rigor of your results.

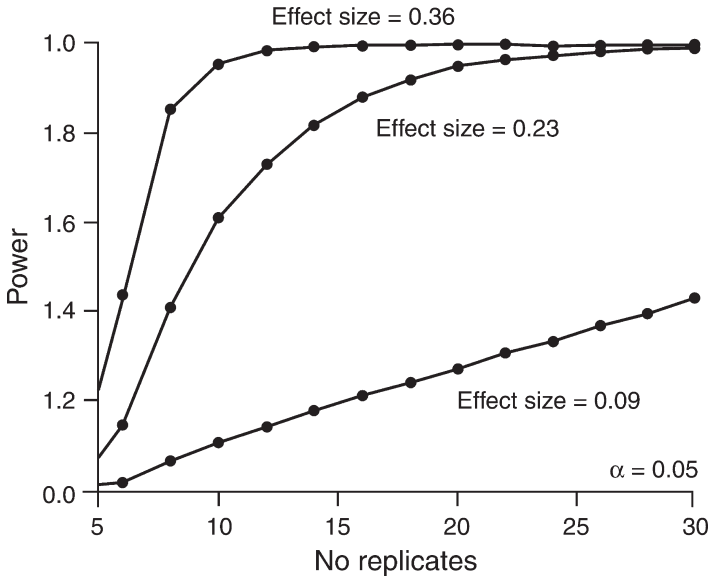

Fig. 3.3 The influence of number of replicates on statistical power to detect small (0.09), medium (0.23), and large (0.36) effect sizes (differences in the probability of predation) between six large and six small trout using a Wilcoxon signed-ranks test. Power was estimated using a Monte Carlo simulation. Reproduced from Steidel et al. (2001) with kind permission from Springer Science + Business Media

3.15 Retrospective Power Analysis

As the names implies, retrospective power analysis is conducted after the study is completed, the data have been collected and analyzed, and the outcome is known. Statisticians typically dismiss retrospective power analysis as being uninformative and perhaps inappropriate and its application is controversial (Gerard et al. 1998). However, in some situations retrospective power analysis can be useful. For example, if a hypothesis was tested and not rejected you might want to know the probability that a Type II error was committed (i.e., did the test have low power?). As summarized by Steidl and Thomas (2001), retrospective power analysis is useful in distinguishing between two reasons for failing to reject the null hypothesis:

- The true effect size was not biologically significant.
- The true effect size was biologically significant but you failed to reject the null hypothesis (i.e., you committed a Type II error).

To make this distinction, you calculate the power to detect a minimally biologically significant effect size given the sample size, α, and variance used in the study. If the resulting power at this effect size is large, then the magnitude of the minimum biologically significant effect would likely lead to statistically significant results. Given that the test was actually not significant, you can infer that the true effect size is likely not this large. If, however, power was small at this effect size, you can infer that the true effect size could be large or small and that your results are inconclusive.

Despite the controversy, retrospective power analysis can be a useful tool in management and conservation. Nevertheless, retrospective power analysis should never be used when power is calculated using the observed effect size. In such cases, the resulting value for power is simply a reexpression of the *p*-value, where low *p*-values lead to high power and vice versa.

3.16 Power Analysis and Wildlife Studies

In practice, observational studies generally have low statistical power. In the case of environmental impact monitoring, the H_0 will usually be that there is no impact to the variable of interest. Accepting a "no impact" result when an experiment has low statistical power may give regulators and the public a false sense of security. The α-level of the experiment is usually set by convention and the magnitude of the effect in an observational study is certainly not controllable. In the case of a regulatory study, the regulation may establish the α-level. Thus, sample size and estimates of variance usually determine the power of observational studies. Many of the methods discussed in this chapter are directed toward reducing variance in observational studies. In properly designed observational studies, the ultimate determinant of statistical power is sample size.

The lack of sufficient sample size necessary to have reasonable power to detect differences between treatment and reference (control) populations is a common

problem in observational studies. For example, reasonably precise estimates of direct animal death from a given environmental perturbation may be made through carcass searches. However, tests of other parameters indicating indirect effects for any given impact (e.g., avoidance of a particular portion of their range by a species) may have relatively little power to detect an effect on the species of concern. Most field studies will result in data that must be analyzed with an emphasis on detection of biological significance when statistical significance is marginal. For a more complete study of statistical power, see Cohen (1973), Peterman (1989), Fairweather (1991), Dallal (1992), and Gerard et al. (1998).

The trend of differences between reference and treatment areas for several important variables may detect effects, even when tests of statistical significance on individual variables have marginal confidence. This deductive, model-based approach is illustrated by the following discussion. The evaluation of effects from wind energy development includes effects on individual birds (e.g., reduction or increase in use of the area occupied by the turbines) and population effects such as mortality (e.g., death due to collision with a turbine). Several outcomes are possible from impact studies. For example, a decline in bird use on a new wind plant without a similar decline on the reference area(s) may be interpreted as evidence of an effect of wind energy development on individual birds. The presence of a greater number of carcasses of the same species near turbines than in the reference plots increases the weight of evidence that an effect can be attributed to the wind plant. However, a decline in use of both the reference area(s) and the development area (i.e., area with wind turbines) in the absence of large numbers of carcasses suggests a response unrelated to the wind plant. Data on covariates (e.g., prey) for the assessment and reference area(s) could be used to further clarify this interpretation.

The level at which effects are considered biologically significant is subjective and will depend on the species/resource involved and the research question of interest. Additionally, we note that a biologically significant effect, although not statistically significant, can have population level implications (see Sect. 1.5.3). In the case of bird fatalities, even a small number of carcasses of a rare species associated with the perturbation may be considered significant, particularly during the breeding season. A substantial number of carcasses associated with a decline in use relative to the reference area, particularly late in the breeding season during the dispersal of young, may be interpreted as a possible population effect. The suggestion of a population effect may lead to additional, more intensive studies.

3.17 Sequential Sample Size Analysis

Sequential sample size analysis is primarily a graphical method of evaluating sample size as data are collected, and attempting to justify the sample size collected after the study is completed. While a study is ongoing, you can easily plot the values of any variable of interest as the sample size increases. For example, one might calculate means and variance as every ten vegetation (or habitat use) plots are gathered

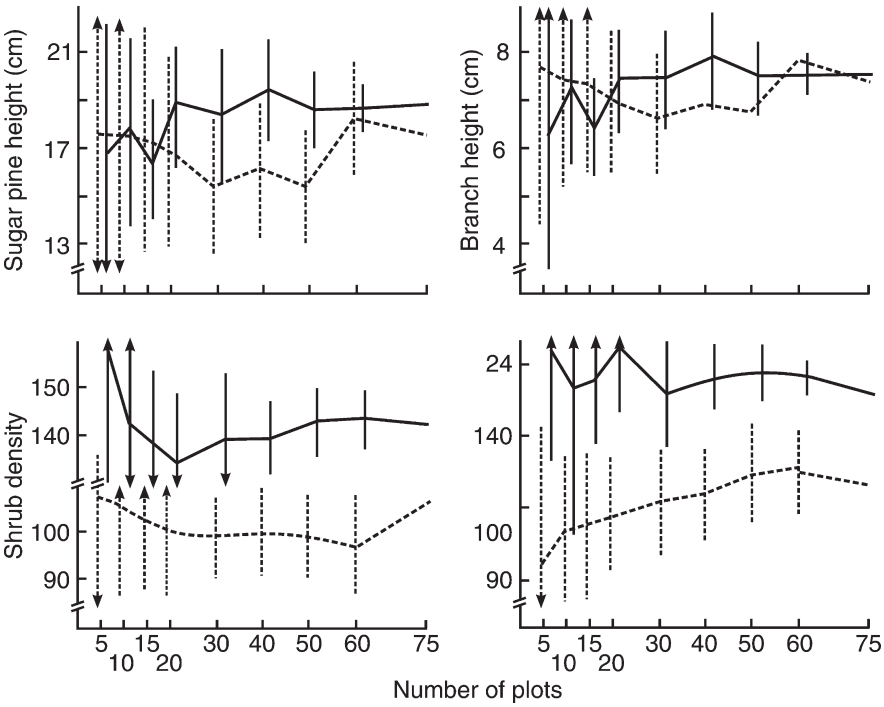

Fig. 3.4 An illustration of how means and variance stabilize with additional sampling. Note that in the all four examples the means (*horizontal solid* and *dashed lines*) and variance (*vertical solid* and *dashed lines*) stabilize with 20–30 plots. Knowledge of the behavior of means and variance influences the amount of sampling in field studies

for a species of interest. You can justify ceasing sampling when the means and variance stabilize (i.e., asymptote; see Fig. 3.4). In a similar fashion, you can take increasingly large random subsamples from a completed data set, calculate the mean and variance, and determine if the values reached an asymptote.

3.18 Bioequivalence Testing

Much has been written criticizing null hypothesis significance testing including applications to wildlife study (see Sect. 1.4.1; Johnson 1999; Steidl and Thomas 2001). McDonald and Erickson (1994), and Erickson and McDonald (1995) describe an alternative approach often referred to as bioequivalence testing. Bioequivalence testing reverses the burden of proof so that a treatment is considered biologically significant until evidence suggests otherwise; thus the role of the null and alternative hypotheses are switched. As summarized by Steidl and Thomas (2001), a minimum effect size that is considered biologically significant is defined.

Then, the alternative hypothesis is stated such that the true effect size is greater than or equal to the minimum effect size that was initially selected. Lastly, the alternative hypothesis is that the true effect size is less than the initial minimum effect size. Thus, Type I error occurs when the researcher concludes incorrectly that no biologically significant difference exists when one does. Recall that this is the type of error addressed by power analysis within the standard hypothesis-testing framework. Bioequivalence testing controls this error rate a priori by setting the α-level of the test. Type II error, however, does remain within this framework when the researcher concludes incorrectly that an important difference exists when one does not.

For a real world example of the significance of value of this alternative approach, consider testing for compliance with a regulatory standard for water quality. In the case of the classic hypothesis testing, poor laboratory procedure resulting in wide confidence intervals could easily lead to a failure to reject the null hypothesis that a water quality standard had been exceeded. Conversely, bioequivalence testing protects against this potentiality and is consistent with the precautionary principle. While this approach appears to have merit, it is not currently in widespread use in wildlife science.

3.19 Effect Size and Confidence Intervals

As discussed earlier, null hypothesis significance testing is problematic because any two samples will usually, show a statistical difference if examined finely enough, such as through increasing sample size (see Sect. 1.4.1). Conversely, no statistical significance will be evident if the sample size was too small or the variance in the data is too great even when differences are biologically important (see Sect. 1.5.3). These scenarios can be distinguished by reporting an estimate of the effect size and its associated confidence interval, thus providing far more biological information than available through a *p*-value.

Confidence intervals (CI) may be used to test a null hypothesis. When estimated with the data for an observed effect size, a CI represents the likely range of numbers that cannot be excluded as possible values of the true effect size if the study were repeated infinitely into the future with probability $1 - \alpha$. If the $100(1 - \alpha)\%$ CI 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.. Additionally, CIs provide more information than a hypothesis test because they establish approximate bounds on the likely value of the true effect size. Figure 3.5 (from Steidl and Thomas 2001) presents the possible various hypothetical observed effects and their associated $100(1 - \alpha)\%$ CI. Note that when the vertical CI line crosses the solid horizontal line (zero effect), no statistically significant effect has occurred.

Case A – the CI for the estimated effect excludes 0 effect and includes only biologically significant effects; the study is both statistically and biologically significant.

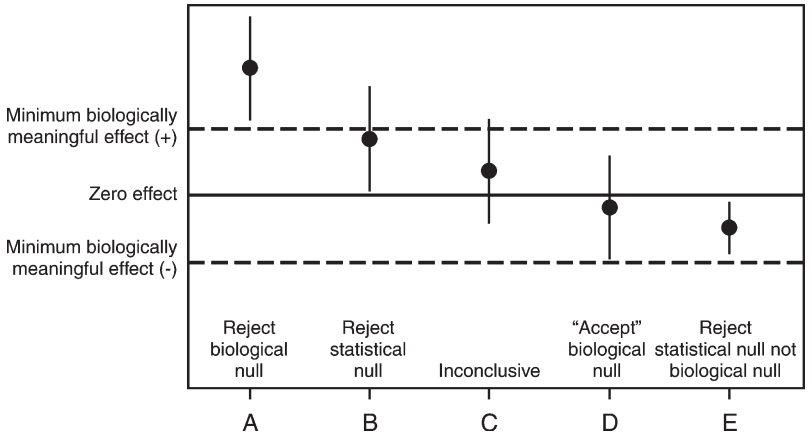

Fig. 3.5 Hypothetical observed effects (*circles*) and their associated $100(1-\alpha)\%$ confidence intervals. The *solid line* represents zero effect, and *dashed lines* represent minimum biologically important effects. In case A, the confidence interval for the estimated effect excludes zero effect 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. Reproduced from Steidel et al. (2001) with kind permission from Springer Science + Business Media

Case B – the CI excludes 0 so it is statistically significant, but includes values that are below that thought to be biologically significant; the study is thus inconclusive biologically.

Case C – the CI includes 0 effect and biologically significant effects, so it is inconclusive statistically.

Case D – the CI includes 0 effect but excludes all effects considered biologically significant; thus the null hypothesis of no biologically significant effect cannot be rejected.

Case E – the CI excludes 0 effect but does not include effects considered biologically significant; the study is statistically but not biologically significant. This situation often occurs when you have very large sample sizes – note now the CI has narrowed.

In CI estimation, the focus is on establishing plausible bounds on the true effect size and determining whether biologically significant effect sizes are contained within those bounds. In retrospective power analysis, however, the focus is on the probability of obtaining a statistically significant result if the effect sizes were truly biologically significant. Steidl and Thomas (2001) concluded that the CI approach

was preferable because interpretation of results is relatively straightforward, more informative, and viewed from a biological rather than a probabilistic context.

3.20 Making Adjustments When Things Go Wrong

As in much of life, things can and often do go wrong in the best-designed studies. The following are a few case studies that illustrate adjustments that salvage a study when problems occur.

Case 1 – As previously discussed, Sawyer (2006) conducted a study to determine the impact of gas development on habitat use and demographics of mule deer in southwestern Wyoming. Although the study of habitat use clearly demonstrated a decline in use of otherwise suitable habitat, the lack of a suitable control hampered identification of the relationship of this impact to population demographics. Sawyer (2006) established a reference area early in the study based on historical data supplemented by aerial surveys during a pilot study period. While the impact area boundary remained suitable over the course of the 4-year study, the boundary around the control area turned out to be inadequate. That is, each year the deer distribution was different, resulting in the need for continually expanding the area being surveyed as a control. Thus, even though the numbers of deer remained relatively unchanged in the reference area, the fact that the boundaries continued to change made a comparison of abundance and other demographic characteristics between the control and impact area problematic. Demographic data for the deer within the impact area did show declines in reproductive rate and survival, although the reductions were not statistically different from 0. Additionally, emigration rates did not satisfactorily explain the decline in deer numbers in the impact area. Finally, simulations using the range of reproduction and survival measured in the impact area suggested that those declines, while not statistically significant could, when combined with emigration rates explain the decline in deer numbers. While there is still opportunity for confounding and cause and effect is still strictly professional judgment, the weight of evidence suggests that the loss in effective habitat caused by the gas development may have resulted in a decline in deer abundance and supports a closer look at the impact of gas development on mule deer in this area.

Case 2 – McDonald (2004) surveyed statisticians and biologists, and reported successes and failures in attempts to study rare populations. One of the survey respondents, Lowell Diller (Senior Biologist, Green Diamond Resource Company, Korbel, California, USA) suggested that "A rare population is one where it is difficult to find individuals, utilizing known sampling techniques, either because of small numbers, secretive and/or nocturnal behavior, or because of clumped distribution over large ranges, i.e., a lot of zeros occur in the data. Therefore, a rare population is often conditional on the sampling techniques available." Lowell provided an illustration of his point by describing surveys conducted for snakes during the mid-1970s on the Snake River Birds of Prey Area in southern Idaho. Surveys were being conducted for night snakes (*Hypsiglena torquata*), which were thought to be one of the

rarest snakes in Idaho with only four known records for the state. His initial surveys, using standard collecting techniques for the time (turning rocks and such along some transect, or driving roads at night), confirmed that night snakes were very rare. In the second year of his study, however, he experimented with drift fences and funnel traps and suddenly began capturing numerous night snakes. They turned out to be the most common snakes in certain habitats and were the third most commonly captured snake within the entire study area. This case study illustrates two points, unsuccessful surveys may be the result of "common wisdom" being incorrect, and/or standard techniques may be ineffective for some organisms and/or situations.

Case 3 – The Coastal Habitat Injury Assessment started immediately after the EVOS in 1989 with the selection of heavily oiled sites for determining the rate of recovery. To allow an estimate of injury, the entire oiled area was divided into 16 strata based on the substrate type (exposed bedrock, sheltered bedrock, boulder/cobble, and pebble/gravel) and degree of oiling (none, light, moderate, and heavy). Four sites were then selected from each of the 16 strata for sampling to estimate the abundance of more than a thousand species of animals and plants. The stratification and site selection were all based on the information in a geographical information system (GIS). Unfortunately, some sites were excluded from sampling because of their proximity to active eagle nests, and more importantly, many of the oiling levels were misclassified and some of the unoiled sites were under the influence of freshwater dramatically reducing densities of marine species. So many sites were misclassified by the GIS system that the initial study design was abandoned in 1990. Alternatively, researchers matched each of the moderately and heavily oiled sites sampled in 1989 with a comparable unoiled control site based on physical characteristics, resulting in a paired comparison design. The Trustees of Natural Resources Damage Assessment, the state of Alaska and the US Government, estimated injury by determining the magnitude of difference between the paired oiled and unoiled sites (Highsmith et al. 1993; McDonald et al. 1995; Harner et al. 1995). Manley (2001) provides a detailed description of the rather unusual analysis of the resulting data.

McDonald (2004) concluded that the most important characteristics of successful studies are (1) they trusted in random sampling, systematic sampling with a random start, or some other probabilistic sampling procedure to spread the initial sampling effort over the entire study area and (2) they used appropriate field procedures to increase detection and estimate the probability of detection of individuals on sampled units. It seems clear that including good study design principles in the initial study as described in this chapter increases the chances of salvaging a study when things go wrong.

3.21 Retrospective Studies

As the name implies, a retrospective study is an observational study that looks backward in time. Retrospective studies can be an analysis of existing data or a study of events that have already occurred. For example, we find data on bird fatalities from

several independent surveys of communications towers and we figure out why they died. Similarly, we design a study to determine the cause of fatalities in an area that has been exposed to an oil spill. A retrospective study can address specific statistical hypotheses relatively rapidly, because data are readily available or already in hand; all we need to do is analyze the data and look for apparent treatment effects and correlations. In the first case, the birds are already dead; we just have to tabulate all the results and look at the information available for each communications tower. Numerous mensurative experiments used to test hypotheses are retrospective in nature (See Sinclair 1991; Nichols 1991); and, medical research on human diseases is usually a retrospective study. Retrospective studies are opposed to prospective studies, designed studies based on a priori hypotheses about events that have not yet occurred.

Retrospective studies are common in ecology and are the only option in most post hoc impact assessments. Williams et al. (2002) offer two important caveats to the interpretation of retrospective studies. First, inferences from retrospective studies are weak, primarily because response variables may be influenced by unrecognized and unmeasured covariates. Second, patterns found through mining the data collected during a retrospective study are often used to formulate a hypothesis that is then tested with the same data. This second caveat brings to mind two comments Lyman McDonald heard Wayne Fuller make at a lecture at Iowa State University. The paraphrased comments are that "the good old data are not so good" and "more will be expected from the data than originally designed." In general, data mining should be avoided or used as to develop hypotheses that are tested with newly obtained empirical data. Moreover, all the above study design principles apply to retrospective studies.

3.22 Summary

Wildlife studies may include manipulative experiments, quasiexperiments, or mensurative or observational studies. With manipulative experiments there is much more control of the experimental conditions; there are always two or more different experimental units receiving different treatments; and there is a random application of treatments. Observational studies involve making measurements of uncontrolled events at one or more points in space or time with space and time being the only experimental variable or treatment. Quasiexperiments are observational studies where some control and randomization may be possible. The important point here is that all these studies are constrained by a specific protocol designed to answer specific questions or address hypotheses posed prior to data collection and analysis.

Once a decision is made to conduct research there are a number of practical considers including the area of interest, time of interest, species of interest, potentially confounding variables, time available to conduct studies, budget, and the magnitude of the anticipated effect.

Single-factor designs are the simplest and include both paired and unpaired experiments of two treatments or a treatment and control. Adding blocking, including randomized block, incomplete block, and Latin squares designs further complicates the completely randomized design. Multiple designs include factorial experiments, two-factor experiments and multifactor experiments. Higher order designs result from the desire to include a large number of factors in an experiment. The object of these more complex designs is to allow the study of as many factors as possible while conserving observations. Hierarchical designs as the name implies increases complexity by having nested experimental units, for example split-plot and repeated measures designs. The price of increased complexity is a reduction in effective sample size for individual factors in the experiment.

ANCOVA uses the concepts of ANOVA and regression to improve studies by separating treatment effects on the response variable from the effects of covariates. ANCOVA can also be used to adjust response variables and summary statistics (e.g., treatment means), to assist in the interpretation of data, and to estimate missing data.

Multivariate analysis considers several related random variables simultaneously, each one considered equally important at the start of the analysis. This is particularly important in studying the impact of a perturbation on the species composition and community structure of plants and animals. Multivariate techniques include multidimensional scaling and ordination analysis by methods such as principal component analysis and detrended canonical correspondence analysis.

Other designs frequently used to increase efficiency, particularly in the face of scarce financial resources, or when manipulative experiments are impractical include sequential designs, crossover designs, and quasiexperiments. Quasiexperiments are designed studies conducted when control and randomization opportunities are possible, but limited. The lack of randomization limits statistical inference to the study protocol and inference beyond the study protocol is usually expert opinion. The BACI study design is usually the optimum approach to quasiexperiments. Meta-analysis of a relatively large number of independent studies improves the confidence in making extrapolations from quasiexperiments.

An experiment is statistically very powerful if the probability of concluding no effect when in fact effect does exist is very small. Four interrelated factors determine statistical power: power increases as sample size, α-level, and effect size increase; power decreases as variance increases. Understanding statistical power requires an understanding of Type I and Type II error, and the relationship of these errors to null and alternative hypotheses. It is important to understand the concept of power when designing a research project, primarily because such understanding grounds decisions about how to design the project, including methods for data collection, the sampling plan, and sample size. To calculate power the researcher must have established a hypothesis to test, understand the expected variability in the data to be collected, decide on an acceptable α-level, and most importantly, a biologically relevant response level. Retrospective power analysis occurs after the study is completed, the data have been collected and analyzed, and with a known outcome. Statisticians typically dismiss retrospective power analysis as being uninformative

and perhaps inappropriate and its application is controversial, although it can be useful in some situations.

Bioequivalence testing, an alternative to the classic null hypothesis significance testing reverses the burden of proof and considers the treatment biologically significant until evidence suggests otherwise; thus switching the role of the null and alternative hypotheses. The use of estimation and confidence intervals to examine treatment differences is also an effective alternative to null hypothesis testing and often provides more information about the biological significance of a treatment.

Regardless of the care taken, the best-designed experiments can and many will go awry. The most important characteristics of successful studies include (1) they trusted in random sampling, systematic sampling with a random start, or some other probabilistic sampling procedure to spread the initial sampling effort over the entire study area and (2) they used an appropriate field procedures to increase detection and estimate the probability of detection of individuals on sampled units. It seems clear that including good study design principles in the initial study as described in this chapter increases the chances of salvaging a study when things go wrong.

Study designs must be study-specific. The feasibility of different study designs will be strongly influenced by characteristics of the different designs and by the available opportunities for applying the treatment (i.e., available treatment structures). Other, more practical considerations include characteristics of study subjects, study sites, the time available for the study, the time period of interest, the existence of confounding variables, budget, and the level of interest in the outcome of the study by others. Regardless of the study environment, all protocols should follow good scientific methods. Even with the best of intentions, though, study results will seldom lead to clear-cut statistical inferences.

There is no single combination of design and treatment structures appropriate for all situations. Our advice is to seek assistance from a statistician and let common sense be your guide.

References

Anderson, R. L., J. Tom, N. Neumann, and J. A. Cleckler. 1996. Avian Monitoring and Risk Assessment at Tehachapi Pass Wind Resource Area, California. Staff Report to California Energy Commission, Sacramento, CA, November, 1996.

Anderson, R. L., M. L. Morrison, K. Sinclair, and M. D. Strickland. 1999. Studying Wind Energy/Bird Interactions: A Guidance Document. Avian Subcommittee of the National Wind Coordinating Committee, Washington, DC.

Barrett, M. A., and P. Stiling. 2006. Key deer impacts on hardwood hammocks near urban areas. J. Wildl. Manage. 70(6): 1574–1579.

Bates, J. D., R. F. Miller, and T. Svejcar. 2005. Long-term successional trends following western juniper cutting. Rangeland Ecol. Manage. 58(5): 533–541.

Berenbaum, M. R., and A. R. Zangerl. 2006. Parsnip webworms and host plants at home and abroad: Trophic complexity in a geographic mosaic. Ecology 87(12): 3070–3081.

Borgman, L. E., J. W. Kern, R. Anderson-Sprecher, and G. T. Flatman. 1996. The sampling theory of Pierre Gy: Comparisons, implementation, and applications for environmental sampling, in

L. H. Lawrence, Ed. Principles of Environmental Sampling, 2nd Edition, pp. 203–221. ACS Professional Reference Book, American Chemical Society, Washington, DC.
Box, G. E. P., and B. C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc. 70: 70–79.
Box, G. E. P., W. G. Hunter, and J. S. Hunter. 1978. Statistics for Experimenters, an Introduction to Design, Data Analysis, and Model Building. Wiley, New York.
Bystrom, P., L. Persson, and E. Wahlstrom. 1998. Competing predators and prey: Juvenile bottlenecks in whole-lake experiments. Ecology 79(6): 2153–2167.
Cade, T. J. 1994. Industry research: Kenetech wind power. In Proceedings of National Avian-Wind Power Planning Meeting, Denver, CO, 20–21 July 1994, pp. 36–39. Rpt. DE95-004090. Avian Subcommittee of the National Wind Coordinating Committee, % RESOLVE Inc., Washington, DC, and LGL Ltd, King City, Ontario.
Cochran, W. G. 1977. Sampling Techniques, 3rd Edition. Wiley, New York.
Cochran, W. G., and G. Cox. 1957. Experimental Designs, 2nd Edition. Wiley, New York.
Cohen, J. 1973. Statistical power analysis and research results. Am. Educ. Res. J. 10: 225–229.
Cox, D. R. 1958. Planning of Experiments (Wiley Classics Library Edition published 1992). Wiley, New York.
Cox, J. J., D. S. Maehr, and J. L. Larkin. 2006. Florida panther habitat use: New approach to an old problem. J. Wildl. Manage. 70(6): 1778–1785.
Crowder, M. J., and D. J. Hand. 1990. Analysis of Repeated Measures. Chapman and Hall, London.
Dallal, G. E. 1992. The 17/10 rule for sample-size determinations (letter to the editor). Am. Stat. 46: 70.
Dillon, W. R., and M. Goldstein. 1984. Multivariate analysis methods and applications. Wiley, New York.
Eberhardt, L. L., and J. M. Thomas. 1991. Designing environmental field studies. Ecol. Monogr. 61: 53–73.
Egger, M., M. Schneider, and D. Smith. 1998. Meta-analysis spurious precision? Meta-analysis of observations studies. Br. Med. J. 316: 140–144.
Erickson, W. P., and L. L. McDonald. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. Environ. Toxicol. Chem. 14: 1247–1256.
Erickson, W., G. Johnson, D. Young, D. Strickland, R. Good, M. Bourassa, K. Bay, and K. Sernka. 2002. Synthesis and Comparison of Baseline Avian and Bat Use, Raptor Nesting and Mortality Information from Proposed and Existing Wind Developments. Prepared by Western EcoSystems Technology, Inc., Cheyenne, WY, for Bonneville Power Administration, Portland, OR. December 2002 [online]. Available: http://www.bpa.gov/Power/pgc/wind/Avian_and_Bat_Study_12-2002.pdf
Fairweather, P. G. 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Freshwat. Res. 42: 555–567.
Fisher, R. A. 1966. The Design of Experiments, 8th Edition. Hafner, New York.
Fisher, R. A. 1970. Statistical Methods for Research Workers, 14th Edition. Oliver and Boyd, Edinburgh.
Flemming, R. M., K. Falk, and S. E. Jamieson. 2006. Effect of embedded lead shot on body condition of common eiders. J. Wildl. Manage. 70(6): 1644–1649.
Folks, J. L. 1984. Combination of independent tests, in P. R. Krishnaiah and P. K. Sen, Eds. Handbook of Statistics 4, Nonparametric Methods, pp. 113–121. North-Holland, Amsterdam.
Garton, E. O., J. T. Ratti, and J. H. Giudice. 2005. Research and experimental design, in C. E. Braun, Ed. Techniques for Wildlife Investigation and Management, 6th Edition, pp. 43–71. The Wildlife Society, Bethesda, Maryland, USA.
Gasaway, W. C., S. D. Dubois, and S. J. Harbo. 1985. Biases in aerial transect surveys for moose during May and June. J. Wildl. Manage. 49: 777–784.
Gerard, P. D., D. R. Smith, and G. Weerakkody. 1998. Limits of retrospective power analysis. J. Wildl. Manage. 62: 801–807.

Gilbert, R. O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

Gilbert, R. O., and J. C. Simpson. 1992. Statistical methods for evaluating the attainment of cleanup standards. Vol. 3, Reference-Based Standards for Soils and Solid Media. Prepared by Pacific Northwest Laboratory, Battelle Memorial Institute, Richland, WA, for U.S. Environmental Protection Agency under a Related Services Agreement with U.S. Department of Energy, Washington, DC. PNL-7409 Vol. 3, Rev. 1/UC-600.

Gitzen, R. A., J. J. Millspaugh, and B. J. Kernohan. 2006. Bandwidth selection for fixed-kernel analysis of animal utilization distributions. J. Wildl. Manage. 70(5): 1334–1344.

Glass, G. V., P. D. Peckham, and J. R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237–288.

Gordon, A. D. 1981. Classification. Chapman and Hall, London.

Green, R. H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley, New York.

Green, R. H. 1984. Some guidelines for the design of biological monitoring programs in the marine environment, in H. H. White, Ed. Concepts in Marine Pollution Measurements, pp. 647–655. University of Maryland, College Park. MD.

Harner, E. J., E. S. Gilfillan, and J. E. O'Reilly. 1995. A comparison of the design and analysis strategies used in assessing the ecological consequences of the Exxon Valdez. Paper presented at the International Environmetrics Conference, Kuala Lumpur, December 1995.

Herring, G., and J. A. Collazo. 2006. Lesser scaup winter foraging and nutrient reserve acquisition in east-central Florida. J Wildl. Manage. 70(6): 1682–1689.

Highsmith, R. C., M. S. Stekoll, W. E. Barber, L. Deysher, L. McDonald, D. Strickland, and W. P. Erickson. 1993. Comprehensive assessment of coastal habitat, final status report. Vol. I, Coastal Habitat Study No. 1A. School of Fisheries and Ocean Sciences, University of Fairbanks, AK.

Howell, J. A. 1995. Avian Mortality at Rotor Swept Area Equivalents. Altamont Pass and Montezuma Hills, California. Prepared for Kenetech Windpower (formerly U.S. Windpower, Inc.), San Francisco, CA.

Huitema, B. E. 1980. The Analysis of Covariance and Alternatives. Wiley, New York.

Hunt, G. 1995. A Pilot Golden Eagle population study in the Altamont Pass Wind Resource Area, California. Prepared by Predatory Bird Research Group, University of California, Santa Cruz CA, for National Renewable Energy Laboratory, Golden, CO. Rpt. TP-441-7821.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54: 187–211.

James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? Annu. Rev. Ecol. Syst. 21: 129–166.

Johnson, D. H. 1995. Statistical sirens: The allure of nonparametrics. Ecology 76: 1998–2000.

Johnson, D. H. 1999. The insignificance of statistical significance testing. J. Wildl. Manage. 63(3): 763–772.

Johnson, B., J. Rogers, A. Chu, P. Flyer, and R. Dorrier. 1989. Methods for Evaluating the Attainment of Cleanup Standards. Vol. 1, Soils and Solid Media. Prepared by WESTAT Research, Inc., Rockville, MD, for U.S. Environmental Protection Agency, Washington, DC. EPA 230/02-89-042.

Kalcounis-Ruppell, M. C., J. M. Psyllakis, and R. M. Brigham. 2005. Tree roost selection by bats: An empirical synthesis using meta-analysis. Wildl. Soc. Bull. 33(3): 1123–1132.

Kempthorne, O. 1966. The Design and Analysis of Experiments. Wiley, New York.

Krebs, C. J. 1989. Ecological Methodology. Harper and Row, New York.

Kristina, J. B., B. B. Warren, M. H. Humphrey, F. Harwell, N. E., Mcintyre, P. R. Krausman, and M. C. Wallace. 2006. Habitat use by sympatric mule and white-tailed deer in Texas. J. Wildl. Manage. 70(5): 1351–1359.

Lanszki, J. M., M. Heltai, and L. Szabo. 2006. Feeding habits and trophic niche overlap between sympatric golden jackal (Canis aureus) and red fox (Vulpes vulpes) in the Pannonian ecoregion (Hungary). Can. J. Zool. 84: 1647–1656.

Ludwig, J. A., and J. F. Reynolds. 1988. Statistical Ecology: A Primer on Methods and Computing. Wiley, New York.
Manly, B. F. J. 1986. Multivariate Statistical Methods: A Primer. Chapman and Hall, London.
Manly, B. F. J. 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London.
Manly, B. F. J. 1992. The Design and Analysis of Research Studies. Cambridge University Press, Cambridge.
Manly, B. F. J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd Edition, 300 pp. Chapman and Hall, London (1st edition 1991, 2nd edition 1997).
Manly, B. F. J. 2001. Statistics for environmental science and management. Chapman and Hall/CRC, London.
Martin, L. M., and B. J. Wisley. 2006. Assessing grassland restoration success: Relative roles of seed additions and native ungulate activities. J. Appl. Ecol. 43: 1098–1109.
McDonald, L. L. 2004. Sampling rare populations, in W. L. Thompson, Ed. Sampling Rare or Elusive Species, pp. 11–42. Island Press, Washington, DC.
McDonald, L. L., and W. P. Erickson. 1994. Testing for bioequivalence in field studies: Has a disturbed site been adequately reclaimed?, in D. J. Fletcher and B. F. J. Manly, Eds. Statistics in Ecology and Environmental Monitoring, pp. 183–197. Otago Conference Series 2, Univ. Otago Pr., Dunedin, New Zealand.
McDonald, L. L., W. P. Erickson, and M. D. Strickland. 1995. Survey design, statistical analysis, and basis for statistical inferences in Coastal Habitat Injury Assessment: *Exxon Valdez* Oil Spill, in P. G. Wells, J. N. Butler, and J. S. Hughes, Eds. *Exxon Valdez* Oil Spill: Fate and Effects in Alaskan Waters. ASTM STP 1219. American Society for Testing and Materials, Philadelphia, PA.
McKinlay, S. M. 1975. The design and analysis of the observational study – A review. J. Am. Stat. Assoc. 70: 503–518.
Mead, R., R. N. Curnow, and A. M. Hasted. 1993. Statistical Methods in Agriculture and Experimental Biology, 2nd Edition. Chapman and Hall, London.
Mieres, M. M., and L. A. Fitzgerald. 2006. Monitoring and managing the harvest of tegu lizards in Paraguay. J. Wildl. Manage. 70(6): 1723–1734.
Miles, A. C., S. B. Castleberry, D. A. Miller, and L. M. Conner. 2006. Multi-scale roost site selection by evening bats on pine-dominated landscapes in southwest Georgia. J. Wildl. Manage. 70(5): 1191–1199.
Miller, D. A., E. B. Arnett, and M. J. Lacki. 2003. Habitat management for forest-roosting bats of North America: A critical review of habitat studies. Wildl. Soc. Bull. 31: 30–44.
Milliken, G. A., and D. E. Johnson. 1984. Analysis of Messy Data. Van Nostrand Reinhold, New York.
Montgomery, D. C. 1991. Design and Analysis of Experiments, 2nd Edition. Wiley, New York.
Morrison, M. L., G. G. Marcot, and R. W. Mannan. 2006. Wildlife–Habitat Relationships: Concepts and Applications, 2nd Edition. University of Wisconsin Press, Madison, WI.
National Research Council. 1985. Oil in the Sea: Inputs, Fates, and Effects. National Academy, Washington, DC.
Nichols, J. D. 1991. Extensive monitoring programs viewed as long-term population studies: The case of North American waterfowl. Ibis 133(Suppl. 1): 89–98.
Orloff, S., and A. Flannery. 1992. Wind Turbine Effects on Avian Activity, Habitat Use, and Mortality in Altamont Pass and Solano County Wind Resource Areas. Prepared by Biosystems Analysis, Inc., Tiburon, CA, for California Energy Commission, Sacramento, CA.
Page, D. S., E. S. Gilfillan, P. D. Boehm, and E. J. Harner. 1993. Shoreline ecology program for Prince William Sound, Alaska, following the *Exxon Valdez* oil spill: Part 1 – Study design and methods [Draft]. Third Symposium on Environmental Toxicology and Risk: Aquatic, Plant, and Terrestrial. American Society for Testing and Materials, Philadelphia, PA.
Peterle, T. J. 1991. Wildlife Toxicology. Van Nostrand Reinhold, New York.
Peterman, R. M. 1989. Application of statistical power analysis on the Oregon coho salmon problem. Can. J. Fish. Aquat. Sci. 46: 1183–1187.

Pielou, E. C. 1984. The Interpretation of Ecological Data: A Primer on Classification and Ordination. Wiley, New York.

Samual, M. D., E. O. Garton, M. W. Schlegel, and R. G. Carson. 1987. Visibility bias during aerial surveys of elk in northcentral Idaho. J. Wildl. Manage. 51: 622–630.

Sawyer, H., R. M. Nielson, F. Lindzey, and L. L. McDonald. 2006. Winter habitat selection of mule deer before and during development of a natural gas field. J. Wildl. Manage. 70: 396–403.

Scheaffer, R. L., W. Mendenhall, and L. Ott. 1990. Elementary Survey Sampling. PWS-Kent, Boston.

Seber, G. A. F. 1984. Multivariate Observations. Wiley, New York.

Shenk, T. M., A. B. Franklin, and K. R. Wilson. 1996. A model to estimate the annual rate of golden eagle population change at the Altamont Pass Wind Resource Area. In Proceedings of National Avian-Wind Power Planning Meeting II, Palm Springs, California, 20–22 September 1995, pp. 47–54. Proceedings prepared for the Avian Subcommittee of the National Wind Coordinating Committee Washington, DC, by LGL Ltd, King City, Ontario.

Sinclair, A. R. E. 1991. Science and the practice of wildlife management. J. Wildl. Manage. 55: 767–773.

Skalski, J. R., and D. S. Robson. 1992. Techniques for Wildlife Investigations: Design and Analysis of Capture Data. Academic, San Diego, CA.

Smith, W. 1979. An oil spill sampling strategy, in R. M. Cormack, G. P. Patil, and D. S. Robson, Eds. Sampling Biological Populations, pp. 355–363. International Co-operative Publishing House, Fairland, MD.

Steel, R. G. D., and J. H. Torrie. 1980. Principles and Procedures of Statistics: A Biometrical Approach, 2nd Edition. McGraw-Hill, New York.

Steidl, R. J. and L. Thomas. 2001. Power analysis and experimental design, in Scheiner, S. M. and J. Gurevitch, Eds. Design and Analysis of Ecological Experiments, 2nd Edition, pp 14–36. Oxford University Press, New York.

Stekoll, M. S., L. Deysher, R. C. Highsmith, S. M. Saupe, Z. Guo, W. P. Erickson, L. McDonald, and D. Strickland. 1993. Coastal Habitat Injury Assessment: Intertidal communities and the *Exxon Valdez* oil spill. Presented at the *Exxon Valdez* Oil Spill Symposium, February 2–5, 1993, Anchorage, AK.

Stewart-Oaten, A. 1986. The Before–After/Control-Impact-Pairs Design-for Environmental Impact. Prepared for Marine Review Committee, Inc., Encinitas, CA.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "Pseudoreplication" in time? Ecology 67: 929–940.

Stoner, D. C., M. L. Wolfe, and D. M. Choate. 2006. Cougar exploitation levels in Utah: Implications for demographic structure, population recovery, and metapopulation dynamics. J. Wildl. Manage. 70(6): 1588–1600.

Strickland, M. D., L. McDonald, J. W. Kern, T. Spraker, and A. Loranger. 1994. Analysis of 1992 Dall's sheep and mountain goat survey data, Kenai National Wildlife Refuge. Bienn. Symp. Northern Wild Sheep and Mountain Goat Council.

Strickland, M. D., G. D. Johnson, W. P. Erickson, S. A. Sarappo, and R. M. Halet. 1998a. Avian use, flight behavior and mortality on the Buffalo Ridge, Minnesota, Wind Resource Area. In Proceedings of National Avian-Wind Power Planning Meeting III. Avian Subcommittee of the National Wind Coordinating Committee, % RESOLVE, Inc., Washington, DC.

Strickland, M. D., D. P. Young, Jr., G. D. Johnson, W. P. Erickson, and C. E. Derby. 1998b. Wildlife monitoring studies for the SeaWest Windpower Plant, Carbon County, Wyoming. In Proceedings of National Avian-Wind Power Planning Meeting III. Avian Subcommittee of the National Wind Coordinating Committee, % RESOLVE, Inc., Washington, DC.

Underwood, A. J. 1994. On beyond BACI: Sampling designs that might reliably detect environmental disturbances. Ecol. Appl. 4: 3–15.

Underwood, A. J. 1997. Experiments in Ecology. Cambridge University Press, Cambridge.

United States Department of the Interior [USDI]. 1987. Type B Technical Information Document: Guidance on Use of Habitat Evaluation Procedures and Suitability Index Models for CERCLA

Application. PB88-100151. U.S. Department of the Interior, CERCLA 301 Project, Washington, DC.

Volesky, J. D., W. H. Schacht, P. E. Reece, and T. J. Vaughn. 2005. Spring growth and use of cool-season graminoids in the Nebraska Sandhills. Rangeland Ecol. Manage. 58(4): 385–392.

Walters, C. 1986. Adaptive Management of Renewable Resources. Macmillan, New York.

Western Ecosystems Technology, Inc. 1995. Draft General Design, Wyoming Windpower Monitoring Proposal. Appendix B in Draft Kenetech/PacifiCorp Windpower Project Environmental Impact Statement. FES-95-29. Prepared by U.S. Department of the Interior, Bureau of Land Management, Great Divide Resource Area, Rawlins, WY, and Mariah Associates, Inc., Laramie, WY.

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and Management of Animal Populations, Modeling, Estimation, and Decision Making. Academic, New York.

Winer, B. J. 1971. Statistical Principles in Experimental Design, 2nd Edition. McGraw-Hill, New York.

Winer, B. J., D. R. Brown, and K. M. Michels. 1991. Statistical Principles in Experimental Design, 3rd Edition. McGraw-Hill, New York.

Wolfe, L. L., W. R. Lance, and M. W. Miller. 2004. Immobilization of mule deer with Thiafentanil (A-3080) or Thiafentanil plus Xylazine. J. Wildl. Dis. 40(2): 282–287.

Zar, J. H. 1998. Biostatistical analysis, 2nd Edition. Prentice-Hall, Englewood Cliffs, NJ.