3 Sampling and Experimental Design

Michael J. Hansen, T. Douglas Beard Jr., and Daniel B. Hayes

3.1 INTRODUCTION

Even though many different methods are used to sample fish populations, their habitats, and anglers, sampling plans often share common traits. The appropriate gear and protocol to collect data are important, but the sampling design and the characteristics of the population determine the statistical properties of the estimates obtained. Similarly, when experiments are conducted to determine the response of fishes or their habitats to treatment, the overall experimental design and underlying variability determine the power of the experiment and can limit the questions or hypotheses that can be addressed. The goal of this chapter is to describe some of the most common sampling and experimental designs used in fisheries science. Our principal intent is not to teach the theory underlying these topics but rather to illustrate common data analysis approaches based on that statistical theory.

3.1.1 Populations and Samples

Fisheries scientists take samples from populations because data or information from all individuals in the population typically cannot be obtained. Fundamental to the idea of sampling is that a population of sampling units exists from which samples are taken. Ideally, all sampling units in the population can be sampled, but in many field sampling programs the sample frame, or the set of sample units that are actually available to be sampled, may be only a subset of the entire target population. In general, whenever the sample units in the sample frame differ from the units in the target population, the design may provide results that reflect the sample frame but not the target population (termed bias; see section 3.1.2). The degree of bias due to this mismatch is generally case specific and is virtually impossible to determine. Throughout this chapter, we assume that the sample units in the sampling frame match the units in the target population and that all units are sampled with equal efficiency.

Definition of the sampling unit is not always straightforward and often depends on the objectives of the study. For example, individual fish are sampling units in a

telemetry study of fish home range if the investigator wishes to know how individual fish in a single population use available habitat. If the telemetry study is conducted in several lakes, each lake may be viewed as a sampling unit, with individual fish as secondary sampling units. In both of these cases, the sampling units are naturally defined units. In contrast, consider a situation where sampling units are defined as possible seining sites (Figure 3.1). Seining-site boundaries are defined by the investigator, not by natural boundaries. The critical concept underlying this example is that after the size of a seine site is defined, and an arbitrary starting point is determined, a finite population of nonoverlapping sampling units is defined. This example also illustrates a case in which some sample units are not part of the sample frame because they cannot be sampled with the gear used. Whenever some sites that are part of the target population are not part of the

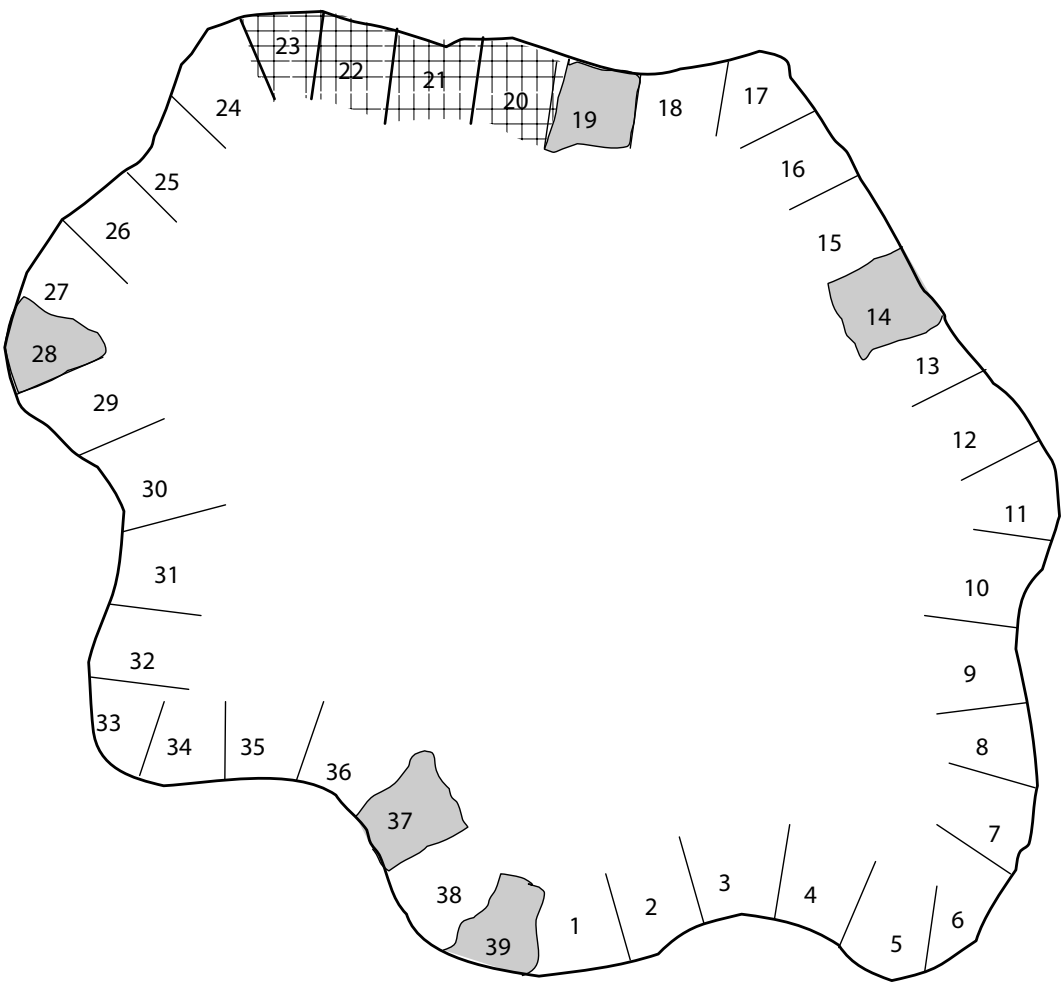


Figure 3.1 Example of a population of seining sites in a lake. Sites selected for sampling are shaded in gray, whereas sites that could not be sampled because of obstructions or soft bottom are crosshatched and were not considered part of the sample frame.

sampling frame, the attribute being estimated will reflect the sample frame but not the target population, so results of sampling will be biased in relation to the true value for the target population.

3.1.2 Bias

The goal of a sampling program is to provide estimates about the characteristics of a population. When used in the context of statistical sampling, bias is generally defined as the difference between the true value of the population attribute and the expected value (i.e., the mean across all possible samples) of the estimator (Cochran 1977). As indicated earlier, a significant source of bias can occur when units in the sample frame differ from units in the entire population or when sampling units within the population are not all sampled with equal efficiency. Therefore, the investigator must define the target population in a way that reflects this mismatch or consider using gears and protocols that produce samples that more accurately characterize the target population. Appropriate sampling gear and protocols are prerequisite for applying the methods covered in this chapter.

In addition to the potential biases described above, estimators such as ratio estimators (section 3.2.2.3) may also result in biased estimates. Although a biased estimator sounds like something to be avoided, for estimators such as ratio estimators some degree of bias is unavoidable (see section 3.2.2.3). However, these biases differ from biases due to sampling frame problems in that the amount of bias can be estimated (remembering that this is based on an average), and a decision can be made whether the bias is acceptable. In some cases, the amount of bias introduced is negligible and is more than offset by gains in precision. Bias is generally evaluated in combination with precision (described below), and their combination is expressed as mean square error (MSE, in squared units of measure), which is calculated as

$$\mathrm{MSE}(\hat{y}) = \mathrm{bias}^2 + \mathrm{var}(\hat{y}) . \tag{3.1}$$

The word accurate is often used as a synonym for unbiased. In common use, however, the word accurate is often used to convey more than simply being unbiased but is used to mean correct. Therefore, we discourage the use of the term accurate in discussions of the statistical properties of sampling programs.

3.1.3 Precision and Confidence Intervals

In addition to obtaining a point estimate of some characteristic of a population, the degree of confidence we have in that estimate is also important to determine. Unless we sample the entire population, our point estimate is unlikely to match the true population value exactly. Thus, a critical concept is that the precision of an estimate is a measure of how likely it is that our estimate is close to the unknown true value. The precision of estimates is often expressed as the standard error (SE) of the estimate or as a confidence interval (CI) around the estimate.

When estimates are viewed as one possible outcome of many possibilities, such that repeating the same procedure would likely result in a different sample being taken with a different point estimate, the estimates can be treated as coming from a statistical distribution, and the SE is simply the square root of the variance of that distribution. For a normal distribution, 68% of the estimates would fall between one SE on either side of the true mean. Similarly, a CI can be thought of as a range within which most, commonly 95%, of the estimates would be expected to lie.

Assuming an unbiased estimator, the precision of estimates is affected by the inherent variability of the attribute being measured and the number of observations of the attribute that are obtained during sampling, in addition to the sampling design (covered later). In general, to obtain estimates of a given level of precision, more samples are required for population attributes that are highly variable than for population attributes that are relatively invariable. Before sampling, the investigator should decide what level of precision is acceptable and then determine how many samples are needed based on prior knowledge of the level of variability that is expected for the attribute of interest, such as would be obtained from a preliminary survey. The level of precision that is acceptable is often determined from the practical needs of the investigator or agency. For example, a fishery scientist may wish to know the mean length of walleyes in a particular lake following a new regulation and would like to have 95% CIs around the mean that were less than ± 50 mm. Here, the acceptable level of precision (±50 mm) could be set to exceed the level of interannual variation that would occur in the absence of a regulation change. Alternatively, the acceptable level of precision could be set at some arbitrary level that is deemed by the agency or investigator to be acceptable.

3.1.4 Random versus Nonrandom Sampling

Generally, samples must be drawn randomly from the population of interest to ensure the sample is representative of the entire population. When samples are drawn nonrandomly or using subjective criteria, measured attributes will usually be biased, though the degree of bias cannot be determined from the samples. For example, sampling is sometimes focused in areas where fish are known to aggregate to avoid spending sampling effort in areas where fish are known to be scarce. Unfortunately, attributes (e.g., length) of fish in areas where they aggregate may differ from attributes of fish in areas where fish are scarce. Thus, such a sample may not be representative of the entire population. The only way to avoid such potential bias is to sample units randomly from the population.

3.2 SAMPLING DESIGNS

3.2.1 Overview of Sampling Designs

In this section, we discuss four commonly used sampling designs, for which both a naturally defined sampling unit (individual fish) and an artificially defined sampling unit (sampling grid locations) can be used (Figure 3.2). Although other

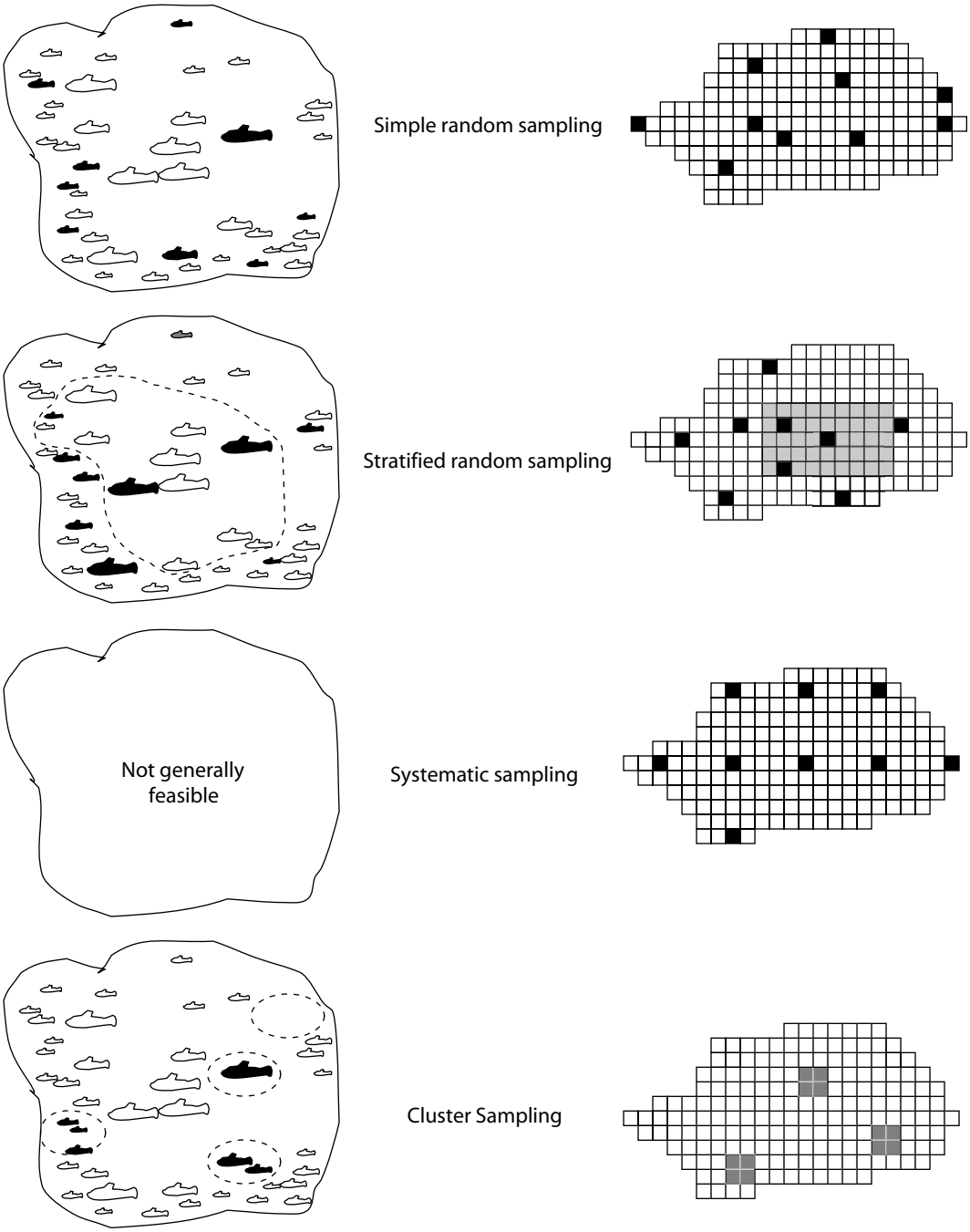


Figure 3.2 Sample frames and sampling unit selection for basic sampling designs. The column on the left illustrates the case in which the sampling unit is individual fish or groups of fish for cluster sampling. The column on the right illustrates the case in which the sampling unit is a grid location in a lake or a group of grids for cluster sampling. Units shaded in black are selected for sampling. The dotted line or shaded region in stratified random sampling indicates stratum boundaries.

sampling designs are available (e.g., Cochran 1977; Thompson 1992; Lohr 1999), we will cover the basic designs that are applicable in most situations. The critical consideration distinguishing each sampling design is how sampling units in the population are defined and how they are selected. In simple random sampling, each sampling unit in the population has an equal probability of being included in the sample, and each sampling unit is selected independently of other units (section 3.2.2). In stratified random sampling, each sampling unit in the population is first assigned to a stratum, and then a simple random sample is independently drawn from each stratum (section 3.2.3). In cluster sampling, the population is divided into primary sampling units, termed clusters, and secondary sampling units, termed elements (section 3.2.4). In systematic sampling, all the sampling units in the population are arranged in a sequence, and then from a random starting point, every kth sampling unit is included in the sample (section 3.2.5). In systematic sampling, selection of the first sampling unit determines all other units in the sample, so sampling units are not independently selected. In large populations, lack of independence does not generally lead to biased estimates of the mean but typically results in a biased estimate of the variance if dependency among sample units is not taken into account. Each of these sampling designs is described in more detail below. To present computational methods concisely and facilitate comparisons among sampling designs, we present the basic formulae for estimating the mean in Table 3.1, which are summarized from Cochran (1977), Thompson (1992), and Lohr (1999).

3.2.2 Simple Random Sampling

In simple random sampling, a sample of size n is randomly selected from a population with N sampling units. Implementing simple random sampling is easiest when all of the sampling units can be enumerated before sampling begins, as in the example of seining sites within a lake (Figure 3.1) or in the example of sampling grids within a lake (Figure 3.2). Sample units are often selected for sampling without replacement (selecting each sample unit no more than once) by using a random number table or generator (Wilde and Fisher 1996). When the sampling frame is unknown before sampling begins, such as in the example of fish within a lake (Figure 3.2), implementing true simple random sampling may be impossible, thereby leading to the use of an alternative design (e.g., cluster sampling, section 3.2.4). Simple random sampling is often less efficient and less precise than are other designs but illustrates concepts and estimators that are inherent in other designs and also provides a basis for understanding the efficiency of other designs, as we will illustrate in the ensuing parts of section 3.2.

3.2.2.1 Estimation of Mean Values

An important property of simple random sampling is that the sample mean and variance provide unbiased estimates of the population mean and variance regardless of the shape of the distribution in the population being sampled (Cochran 1977; Lohr 1999). The sample mean is simply calculated as the sum of

the observations (y_i) divided by the sample size (n; Table 3.1; Box 3.1). Several equivalent formulae are available for calculating the sample variance, but one approach is to sum the squared deviations (differences) between observations of the sampling units (y_i) and the sample mean, and divide by $n - 1$ (Table 3.1; Box 3.1). One can view the sample mean as a simple prediction for each observation, and the sample variance is the average squared deviation between observations and their predicted value. In most fisheries applications, the proportion of the population that is sampled (also known as the sampling fraction, n/N) is small, and the SE of the sample mean is estimated by taking the square root of the sample variance, divided by the sample size (Table 3.1). When the sampling fraction is large (e.g., $n/N > 0.5$), the confidence in estimates of the sample mean is increased, and consequently the SE is reduced:

$$\mathrm{SE}(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}\,. \tag{3.2}$$

The term $(1 - n/N)$ is called the finite population correction. As the sampling fraction approaches 1 (the entire population is sampled), the finite population correction approaches 0, and the SE also approaches 0. In a census of the entire population, the SE of the sample mean would be 0 because all possible sampling units in the population would be included in the sample. Many books on sampling theory include formulae and derivations that include the finite population correction, but it is typically negligible in practice. Therefore, our summary of formulae (Table 3.1) excludes the finite population correction factor.

The SE is often used as a measure of precision of estimates and describes the variability that would be expected if the sampling process could be repeated a large number of times. For a normal distribution, approximately 68% of the distribution is found between 1 SE above and below the mean. An alternative method of conveying the precision is to estimate confidence limits (CLs) on point estimates. Estimating CLs on the mean requires knowledge of the distribution of the mean or assumptions about the shape of the distribution. For large samples (e.g., $n \geq 50$; Zar 1999), the distribution of the mean approaches a Student's t-distribution with $n - 1$ df, so CLs can be estimated as

$$\begin{aligned} &\text{Lower CL} = \bar{y} - t_{\alpha,\, n-1}\mathrm{SE}(\bar{y}), \text{ and} \\ &\text{Upper CL} = \bar{y} + t_{\alpha,\, n-1}\mathrm{SE}(\bar{y}), \end{aligned} \tag{3.3}$$

where $t_{\alpha,\, n-1}$ is the value of the t-distribution (commonly available from a table in a statistics book or equation in a spreadsheet) for an α equal to the probability of making a type I error (often 0.05) and an n of a given sample size. The $\mathrm{SE}(\bar{y})$ is the SE of the mean (see Box 3.1 for an example). The t-distribution is often used to estimate approximate confidence limits for small samples, but the bootstrap method can also be used to estimate CIs and is often recommended for small sample sizes (see Efron and Tibshirami [1998] for details on this approach).

Table 3.1 Summary of formulae for computing the mean and associated measures of precision for common sampling designs. Note that these formulae do not include the finite population correction factor ($1 - n/N$; where n/N is the proportion of the population sampled). Variable definitions are as follows. For simple random sampling, y_i = measurement of item i and n = sample size. For stratified random sampling, y_{ih} = measurement of item i in stratum h; h = stratum index; N_h = number of elements in stratum h; n_h = units sampled from within each stratum; L = total number of strata; and N = number of elements in population. For cluster sampling, y_{ij} = measurement of element j in cluster i; n = number of primary units or clusters sampled; m_i = number of secondary elements i sampled; and M_i = number of secondary elements in cluster i. For regression, or double sampling, $\bar{x}$ = the mean of estimates for the subsample, $\bar{y}$ = the mean of measured values for the subsample, and $\bar{X}$ = the mean of estimates for the entire sample.

	Mean	Variance	SE (mean)
Simple random sampling	$\bar{y} = \frac{\sum y_i}{n}$	$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$	$SE(\bar{y}) = \sqrt{\frac{s^2}{n}}$
Stratified random sampling	$\bar{y} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$	$s_h^2 = \frac{\sum (y_{ih} - \bar{y}_h)^2}{n_h - 1}$	$SE(\bar{y}) = \sqrt{\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h}}$
Single-stage cluster sampling	cluster total $= y_i = \sum y_{ij}$ mean cluster total $= \bar{y} = \frac{\sum y_i}{n}$ mean per secondary unit $= \frac{\sum y_i}{\sum M_i} = \hat{R}$	Can treat (clusters) as simple random sample See ratio estimators	Treat clusters as simple random sample of size n $SE(\hat{R}) = \frac{1}{\sqrt{n}\,\bar{M}} \sqrt{\frac{\sum (y_i - \hat{R}M_i)^2}{n-1}}$

Two-stage cluster sampling	Estimated cluster total $\hat{y}_i = \frac{M_i}{m_i}\sum_{j=1}^{m} y_{ij}$	$s_1^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$	$SE(\bar{y}) = \sqrt{\frac{s_1^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n(n-1)}}$
	Mean cluster total $\bar{y} = \sum_{i=1}^{n}\frac{\hat{y}_i}{n}$		
	Mean per secondary unit within cluster $\bar{y}_i = \sum_{i=1}^{n}\frac{y_{ji}}{m_i}$	$s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2}{n(m-1)}$	
	Estimated population mean per secondary unit $\bar{\bar{y}} = \frac{\sum \hat{y}_i}{\sum M_i}$		$SE(\bar{\bar{y}}) = \sqrt{\frac{\sum_{i=1}^{n} M_i^2(\bar{y}_i - \bar{\bar{y}})^2}{n\bar{M}^2(n-1)}}$
Systematic sampling (one start point)	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	Cannot be directly estimated	Cannot be directly estimated
Systematic sampling (two start points)	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	Can treat data associated with each start point as a cluster	See single-stage cluster sampling
Regression or double sampling	$\bar{y}_{reg} = \bar{y} + b(\bar{X} - \bar{x})$		$SE(\bar{y}_{reg}) = \sqrt{\frac{1}{n(n-2)}\left\{\sum(y_i - \bar{y})^2 - \frac{[\sum(y_i - \bar{y})(x_i - \bar{x})]^2}{\sum(x_i - \bar{x})^2}\right\}}$

Box 3.1 Example of Estimating the Mean Based on Simple Random Sampling

Fifteen sites were randomly selected from an *X* – *Y* grid superimposed on a shallow lake. At each site, the catch of central mudminnow in a throw trap (assumed to be equally efficient at all sites in the lake) was recorded. The goal of the sampling was to determine the mean density of central mudminnows in the lake.

Table Catch of central mudminnow from 15 lake sites randomly selected on an *X* – *Y* grid.

Coordinate and total			
X	*Y*	Catch	(Catch – mean)^2
5	18	3	1.96
15	16	0	2.56
4	9	4	5.76
14	3	1	0.36
11	8	4	5.76
12	5	1	0.36
2	4	1	0.36
3	20	0	2.56
11	7	1	0.36
1	8	0	2.56
2	15	2	0.16
11	2	2	0.16
3	17	3	1.96
3	12	2	0.16
1	10	0	2.56
Total		24	27.60

3.2.2.2 *Estimation of Proportions*

Many characteristics of fishes or their habitats cannot be expressed quantitatively as a continuous variable but can be expressed qualitatively as a categorical variable. For example, the sex of a fish is a qualitative (categorical) trait. Sampling is often undertaken to estimate the proportion of the population (p) that possess some quality or attribute. An attribute that takes on one of only two values (e.g., male or female or mature or immature) is a single or binary classification system, whereas an attribute that falls into one of several categories or classes (e.g., species of fish or length intervals) is a multiple classification system.

For single classification variables, the observation (y_i) is coded as 1 if the individual possesses one attribute or trait and 0 if it possesses the other attribute or trait. The proportion of individuals that possess the trait in the population (p) can

The mean catch and associated measures of precision are calculated as follows, using formulae from Table 3.1.

$$\bar{y} = \frac{\sum y_i}{n} = \frac{24}{15} = 1.6.$$

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{27.6}{15-1} = 1.97.$$

$$\mathrm{SE}(\bar{y}) = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{1.97}{15}} = 0.36.$$

Although the sample size is not large (<30), a normal approximation can be used to estimate approximate 95% confidence limits (CLs).

$$\text{Lower CL} = \bar{y} - (t_{\alpha, n-1})(\mathrm{SE}) = 1.60 - 2.145 \cdot 0.36 = 0.83.$$
$$\text{Upper CL} = \bar{y} + (t_{\alpha, n-1})(\mathrm{SE}) = 1.60 + 2.145 \cdot 0.36 = 2.37.$$

This example can also be used to illustrate how to compute estimates of target sample sizes (equation [3.9]). For example, if we wanted to compute the mean catch with a SE of 0.10, we would start by guessing a sample size of 60 might be adequate. Using this preliminary guess (which is needed to get an initial estimate of the *t*-statistic used in the formula for sample size), we would estimate that the necessary sample size, to the nearest integer, would be

$$n = \frac{(1.97)(2.0)^2}{0.10} \approx 79.$$

Even though this is different than our initial guess of 60, the actual *t*-statistic for 79 is 1.990, which would change our integer estimate of the necessary sample size to only 78.

be estimated by summing the y_i and dividing by n (Table 3.2; Box 3.2; Cochran 1977). The proportion of individuals lacking the trait (q) is termed the complement of p, and is computed as $q = 1 - p$. As with estimates of the mean for quantitative measurements, the estimate of p is unbiased in simple random sampling, as long as the attribute is identified correctly for each individual examined. Although the estimator for the SE of p (Table 3.2; Box 3.2; Cochran 1977) is unbiased, when p is close to 0 or 1 the distribution around p is skewed because p cannot be less than 0 or greater than 1. Therefore, we recommend that CLs for p be estimated from the F-distribution. The lower CL (L_1) for p is (Zar 1999)

$$L_1 = \frac{a}{a + (\mathrm{n} - a + 1)F_{\alpha, \upsilon_1, \upsilon_2}}, \tag{3.4}$$

Table 3.2 Summary of formulae for proportions, and their associated measures of precision, for common sampling designs. The proportion of individuals that possess the trait in the population is given by p and the proportion of individuals lacking the trait, q, is computed as $q = 1 - p$. Note that these formulae do not include the finite population correction factor. Variable definitions for simple random sampling are y_i = observation on unit coded 0 if not in class C and 1 if in class C; a = number of units in class C in sample; n = sample size; and m_i = number of units in cluster i. See Table 3.1 for additional variable definitions.

Sampling design	Proportion in class C	SE(p)
Simple random sampling	$\hat{p} = \frac{\sum y_i}{n} = \frac{a}{n}$	$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}$
Stratified random sampling	$\hat{p}_h = \sum_{h=1}^{H} \frac{y_{ih}}{n_h}$ $\hat{p} = \sum_{h=1}^{H} \frac{N_h \hat{p}_h}{N}$ $= W_h \hat{p}_h$	$\mathrm{SE}(p_{st}) = \sqrt{\sum_{h=1}^{L} \frac{W_h^2 \hat{p}_h \hat{q}_h}{(n_h - 1)}}$
Single-stage cluster sampling	$\hat{p}_i = \frac{\sum y_i}{n} = \frac{a}{n} = p$ $\hat{p}_i = \frac{\sum a_i}{\sum m_i}$	$\mathrm{SE}(\hat{p}) = \sqrt{\frac{1}{n\bar{m}^2} \frac{\sum a_i^2 - 2\hat{p}\sum a_i m_i - \hat{p}^2 \sum m_i}{n-1}}$

Box 3.2 Example of Estimating a Proportion in Simple Random Sampling

One hundred sixteen brown trout were collected at random from a population in a stream with the goal of estimating the proportion in each age-group. The age of each fish was estimated from scales to produce the following data.

Table Age distribution of a random sample of 116 brown trout from a stream.

Age	n
0	55
1	22
2	10
3	18
4	6
5	3
6	1
7	1

The proportion in each age-class was estimated as follows.

$$\hat{p}_0 = \frac{\sum y_i}{n} = \frac{55}{116};\ \hat{p}_1 = \frac{22}{116};\ \hat{p}_2 = \frac{10}{116};\ \hat{p}_3 = \frac{18}{116};\ \hat{p}_4 = \frac{6}{116};\ \hat{p}_5 = \frac{3}{116};\ \hat{p}_6 = \frac{1}{116};\text{and } \hat{p}_7 = \frac{1}{116}.$$

For example, for age 0 the SE (Table 3.2) and CLs (equations [3.4] and [3.5]) were calculated as follows.

$$\text{SE}(\hat{p}_0) = \sqrt{\frac{\hat{p}_0\hat{q}_0}{n-1}} = \sqrt{\frac{0.47 \cdot (1-0.47)}{116-1}} = 0.047.$$

$$\text{Lower CL} = \frac{a}{a + (n-a+1)F_{\alpha, v_1, v_2}} = \frac{55}{55 + (116-55+1)F_{0.05,\, 2\cdot(116-55+1),\, 2\cdot 55}} = 0.3948.$$

$$\text{Upper CL} = \frac{(a+1)F_{\alpha, v_1, v_2}}{n-a+(a+1)F_{\alpha, v_1, v_2}} = \frac{(55+1)F_{0.05,\, (2\cdot 55)+2,\, 2\cdot(116-55+1)-2}}{116-55+(55+1)F_{0.05,\, (2\cdot 55)+2,\, 2\cdot(116-55+1)-2}} = 0.5545.$$

Estimates of the proportion in each age-class, and appropriate measures of precision, are given in the table below.

Table Estimates of the proportion of brown trout in each age-class (p_{age}) and measures of precision.

Age	p_{age}	SE	Lower CL	Upper CL
0	0.47	0.047	0.3948	0.5545
1	0.19	0.037	0.1320	0.2596
2	0.09	0.026	0.0475	0.1418
3	0.16	0.034	0.1028	0.2214
4	0.05	0.021	0.0228	0.0995
5	0.03	0.015	0.0071	0.0655
6	0.01	0.009	0.0004	0.0402
7	0.01	0.009	0.0004	0.0402

where $a = \Sigma y_i$; n = sample size; and the F-statistic is evaluated for the two-tailed level of α, the numerator df $v_1 = 2(n - a + 1)$, and the denominator df $v_2 = 2a$. Similarly, the upper CL (L_2) for p is (Zar 1999)

$$L_2 = \frac{(a+1)F_{\alpha,\, v_1',\, v_2'}}{n - a + (a+1)F_{\alpha,\, v_1',\, v_2'}}, \tag{3.5}$$

where the F-statistic is evaluated for the two-tailed level of α, the numerator df $v_1' = v_2 + 2$, and the denominator df $v_2' = v_1 - 2$ (Box 3.2). The upper and lower CLs for q are obtained by subtracting the upper and lower CLs for p from 1.

For attributes in a multiple classification system, the problem can be simplified by focusing on one class at a time and treating the attribute as a single classification variable where the individual either has the attribute or not. The proportion within any single class and the associated SE is then estimated exactly as for the single classification situation. Box 3.2 illustrates how to calculate proportions, SEs, and CLs for a multiple classification system, where any one class of the multiple classification system can be used to illustrate a single classification variable.

3.2.2.3 *Estimation of Ratios*

Attributes of fishes or habitats are often expressed as ratios of variables that both vary among units, which contrasts with proportions that describe the fraction of a sample that possess a certain attribute, as in section 3.2.2.2. A familiar example is angler catch per effort where both catch and effort vary among individual anglers. Unfortunately, situations in which ratios are estimated are often confused with situations in which a proportion is being estimated. For example, in diet studies, the amount of food consumed, by weight, among various prey taxa is commonly referred to as a proportion but is more appropriate to view as a ratio of the weight consumed of each prey taxon to the total weight consumed. These types of data are best treated as a ratio because the weight consumed of each prey taxon varies among sampling units and the total weight consumed varies among sampling units.

In simple random sampling, the population ratio (R) is estimated from the ratio of the sums of the sampled quantities (Box 3.3; Cochran 1977; Lohr 1999):

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}. \tag{3.6}$$

For catch per effort data, the numerator of equation (3.6) is the sum of catch and the denominator is the sum of effort. For diet data, the numerator is the total weight of one prey taxon and the denominator is the total weight of all prey taxa (as in Box 3.3). This estimator is biased, but the bias tends toward 0 as the sample size increases (Cochran 1977). Cochran (1977) showed that the degree of bias relative to the SE of the estimated ratio can be approximated as

Box 3.3 Example of Estimating a Ratio in Simple Random Sampling

Twenty yellow perch were randomly sampled from a lake, and weights of zooplankton, benthos, and fish in each yellow perch stomach were measured. The goal was to determine the ratio of each prey category to total weight of prey in the diet of the yellow perch population.

Table Ratio of three prey categories to total weight of prey in the diets of 20 yellow perch. Squared deviations of the observed (y) minus the predicted ($\hat{R}x$, where x is the total weight for each fish and $\hat{R}$ is estimated as shown below the table) allows for estimation of population ratios for each prey category (sums of sampled quantities $[y - Rx]^2$).

	Weight in stomach				$(y - Rx)^2$		
Fish and total	Zoo-plankton	Benthos	Fish	Total Weight	Zoo-plankton	Benthos	Fish
1	0.000	0.000	16.217	16.217	4.257	36.537	65.735
2	0.200	2.501	0.000	2.701	0.021	2.233	1.824
3	0.593	0.054	0.000	0.647	0.261	0.035	0.105
4	0.356	0.741	0.000	1.097	0.047	0.110	0.301
5	0.070	1.112	0.000	1.182	0.006	0.450	0.349
6	0.191	1.734	0.000	1.925	0.003	1.033	0.926
7	0.012	0.022	0.000	0.034	0.000	0.000	0.000
8	0.017	2.822	0.000	2.839	0.119	3.111	2.015
9	0.400	2.796	0.000	3.196	0.000	2.575	2.554
10	0.202	2.154	0.000	2.356	0.010	1.627	1.388
11	0.591	0.559	0.000	1.150	0.198	0.017	0.331
12	0.737	0.902	0.000	1.639	0.280	0.085	0.672
13	0.095	0.098	0.000	0.193	0.005	0.001	0.009
14	0.000	0.000	12.090	12.090	2.366	20.306	36.534
15	0.747	1.913	0.000	2.660	0.167	0.849	1.769
16	0.663	0.600	0.000	1.263	0.253	0.017	0.399
17	0.937	0.354	0.000	1.291	0.598	0.016	0.417
18	0.664	0.213	0.000	0.876	0.305	0.013	0.192
19	0.623	0.448	0.000	1.072	0.237	0.002	0.287
20	0.103	2.077	0.000	2.181	0.030	1.600	1.189
Total	7.202	21.099	28.307	56.608	9.160	70.617	116.997

The ratio (equation [3.6]) and SE (equation [3.8]) of each prey category in the diet is estimated as follows.

$$\hat{R}_{\text{zooplankton}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{7.202}{56.608} = 0.127.$$

$$\text{SE}(\hat{R}_{\text{zooplankton}}) = \frac{1}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}} = \frac{1}{\sqrt{20} \cdot 2.830} \sqrt{\frac{9.160}{20-1}} = 0.055.$$

(Box continues)

Box 3.3 *(continued)*

$$\hat{R}_{\text{benthos}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{21.099}{56.608} = 0.373.$$

$$\text{SE}(\hat{R}_{\text{benthos}}) = \frac{1}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}} = \frac{1}{\sqrt{20} \cdot 2.830} \sqrt{\frac{70.617}{20-1}} = 0.152.$$

$$\hat{R}_{\text{fish}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{28.307}{56.608} = 0.500.$$

$$\text{SE}(\hat{R}_{\text{fish}}) = \frac{1}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}} = \frac{1}{\sqrt{20} \cdot 2.830} \sqrt{\frac{116.997}{20-1}} = 0.196.$$

$$\frac{\text{bias}(\hat{R})}{\text{SE}(\hat{R})} = \frac{\text{SE}(\bar{x})}{\bar{X}}, \tag{3.7}$$

where $\bar{X}$, the population mean, generally can be estimated without bias using the sample mean, $\bar{x}$ (see Cochran 1977). Because the SE of x decreases to 0 as sample size increases, the degree of bias also decreases as sample size increases. This expression can be easily computed from sample data to determine if bias is large enough to be problematic. Importantly, a ratio should not be estimated by averaging the ratios for individuals (often termed a mean of ratios) but rather as a ratio of totals (often termed a ratio of means), because a mean of ratios has a larger degree of bias than does a ratio of means, and this bias does not diminish as sample size increases (Cochran 1977). The SE of a ratio is derived from deviations between the numerator of the ratio (y_is) and the product of the denominator of the ratio (x_is) and the ratio ($\hat{R}$) (Box 3.3; Cochran 1977; Lohr 1999):

$$\text{SE}(\hat{R}) = \frac{1}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}}. \tag{3.8}$$

Confidence limits around the estimated ratio are typically approximated using the t-distribution, which is reasonable for large sample sizes but may not represent skewness in the distribution of the estimate for small sample sizes.

3.2.2.4 *Estimation of Sample Size*

The sample size required to estimate the population mean (for example) can be derived for simple random sampling from knowledge of the variance of the y_i in the population when the desired degree of precision is specified. Various strategies have been developed to simplify the process of estimating sample size requirements (Wilde and Fisher 1996), but ultimately the investigator must specify expectations for the outcome of sampling and acceptable levels of precision in estimates. Precision can be expressed on an absolute scale (e.g., ± 10 mm) or on a relative scale (e.g., ± 8% of the mean). When desired precision is expressed in absolute terms (δ), the sample size needed (n; ignoring the finite population correction factor) can be estimated as (Cochran 1977):

$$n = \frac{s^2 t^2_{n-1}}{\delta^2}, \tag{3.9}$$

where s^2 is estimated from a pilot study or prior experience with similar situations and the t-statistic is defined for a given α level from a statistical table or spreadsheet function. Because t_{n-1} depends on the sample size, the estimate of n must be solved by trial and error. In practice, the first guess does not need to be close to the true value, and only two or three iterations are necessary to obtain the solution because the t-statistic does not vary greatly with n.

When the desired precision is expressed in relative terms (r), an estimate of the mean must also be included. As above, a preliminary estimate of the mean can be obtained by a pilot study or prior experience, and sample size can be estimated as

$$n = \left(\frac{t_{n-1} s}{r\bar{y}}\right)^2. \tag{3.10}$$

The sample size that is necessary for specified precision for proportions is analogous to that for a mean value (Cochran 1977):

$$n = \frac{s^2 t^2_{n-1}}{\delta^2}, \tag{3.11}$$

where the variance is estimated as the product of p and q and the investigator specifies the absolute error (δ).

3.2.3 Stratified Random Sampling

Stratified random sampling performs as well or better than simple random sampling in nearly all cases and results in substantial improvement in precision when variation within the strata is less than variation among the strata. In stratified random sampling, the total sample frame containing N sample units is divided

into L subpopulations or strata, each containing N_h sample units. Within each stratum, a simple random sample of n_h sample units is drawn independently. For example, in Figure 3.2 a lake with N sample grids is divided into two depth strata (L = deep and shallow) with N_h sample grids in each stratum, and n_h sample grids are then sampled from within each stratum.

An estimate for the whole population is obtained by weighting estimates from each stratum $\bar{y}_h$ by the fraction of the whole population contained in each stratum ($W_h = N_h/N$). Stratified random sampling is advantageous over simple random sampling because sampling can be allocated disproportionately among strata to ensure adequate precision can be obtained for subpopulations represented by strata. Stratified random sampling requires that the entire sampling frame be divided into strata before sampling begins, so it should not be applied to situations where the strata are defined a posteriori.

3.2.3.1 *Construction of Strata*

To be most efficient, the strata means should differ widely from one another, so that variability between strata is large and variability within each stratum is small. However, the data necessary to specify strata that best partition the variability in the population would require the investigator to complete the survey. Consequently, other features that are readily obtainable and are correlated to the attribute of interest are often used to construct strata. For example, when sampling fish, we often assume that fish associate themselves with habitat conditions (such as water depth), and we construct strata that coincide with habitat boundaries (as in Figure 3.2). Therefore, prior information about the attribute of interest can be used to construct strata whenever available.

How many strata to develop is also a difficult question to answer. The number of strata that can be sampled is obviously limited by the sample size ($L < n$), but a minimum of at least two sample units must be sampled per stratum to allow calculation of the within-stratum variance. In our experience, the number of strata should depend on the quality and amount of available information, so you should use few strata when prior information is not available and more strata when better prior information is available. The sample size within each stratum should be large enough (e.g., at least 10) to provide reasonable estimates for each subpopulation.

3.2.3.2 *Estimation of Mean Values*

The mean value for a stratified random sample is estimated from the mean values within the strata, weighted by the fraction of the entire population of sample units in each stratum (Table 3.1; Box 3.4; Cochran 1977). Stratum means are estimated as described for a simple random sample (section 3.2.2.1), and each stratum mean is weighted (W_h) by the number of units in the stratum sample (N_h) divided by the total number of units in the population (N). The SE of the mean value for a stratified random sample is a weighted sum of the variances of the mean values for the individual strata (Table 3.1; Box 3.4; Cochran 1977). As in simple random sampling, estimates of the stratified mean and its SE are unbiased. Likewise, estimating CLs on the mean requires an assumption regarding the

Box 3.4 Example of Stratified Random Sampling

A grid was superimposed on the map of a shallow lake, and all grid cells were classified as being in one of three depth strata (0–2 m, 2–4 m, >4 m). Ten grid cells were sampled in each depth stratum, and at each site the catch of age-0 yellow perch in a throw trap (assumed to be equally efficient at all sites in the lake) was recorded. The goal of the sampling program was to estimate the mean density of age-0 yellow perch.

Table Catch of age-0 yellow perch at three depth strata within a shallow lake. Variance in parentheses below mean.

0–2-m stratum		2–4-m stratum		>4-m stratum	
Catch and mean	(Catch – mean)^2	Catch and mean	(Catch – mean)^2	Catch and mean	(Catch – mean)^2
0	2.89	4	1.21	7	1.69
2	0.09	2	0.81	5	0.49
2	0.09	3	0.01	7	1.69
2	0.09	5	4.41	7	1.69
3	1.69	2	0.81	5	0.49
1	0.49	4	1.21	5	0.49
3	1.69	1	3.61	7	1.69
2	0.09	3	0.01	6	0.09
2	0.09	2	0.81	3	7.29
0	2.89	3	0.01	5	0.49
1.7 (1.122)		2.9 (1.433)		5.7 (1.789)	

Within each stratum, the mean catch and variance were computed using formulae for a simple random sample (Table 1; Box 3.1 example). The lake contained 320 grid cells, which included 172 in the 0–2-m stratum, 80 in the 2–4-m stratum, and 68 in the > 4-m stratum, so the weight for each stratum (W_h) was

$$W_h = \frac{N_h}{N}.$$

$$W_{0\text{–}2} = \frac{N_{0\text{–}2}}{N} = \frac{172}{320} = 0.5375.$$

$$W_{2\text{–}4} = \frac{N_{2\text{–}4}}{N} = \frac{80}{320} = 0.250.$$

$$W_{>4} = \frac{N_{>4}}{N} = \frac{68}{320} = 0.2125.$$

The stratified mean catch was

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h = (0.5375 \cdot 1.7) + (0.2500 \cdot 2.9) + (0.2125 \cdot 5.7) = 2.85.$$

The SE of the stratified mean catch per effort was

$$\mathrm{SE}(\bar{y}) = \sqrt{\sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h}} = \sqrt{\left(\frac{0.5375^2 \cdot 1.122}{10}\right) + \left(\frac{0.2500^2 \cdot 1.433}{10}\right) + \left(\frac{0.2125^2 \cdot 1.789}{10}\right)} = 0.222.$$

Approximate 95% confidence intervals can be computed (assuming normality) using the same approach as for simple random sampling (Box 3.1).

sampling distribution of the stratified mean. If the stratum means are normally distributed, the *t*-distribution can be used to estimate CLs. In most situations, the degrees of freedom are calculated as the total sample size minus the number of strata. However, Satterthwaite (1946) showed that the effective number of degrees of freedom should be reduced when allocation of sampling effort is not proportional to the weight for each stratum (when n_h/n is not equal to N_h/N). If the finite population correction term is ignored, the effective number of degrees of freedom can be estimated by (derived from Cochran 1977)

$$df = \frac{\left(\sum \frac{s_h^2}{W_h}\right)^2}{\sum \frac{s_h^4}{W_h(n_h - 1)}}. \tag{3.12}$$

3.2.3.3 *Allocation of Samples within Strata*

An important feature of stratified random sampling is that estimates of the mean are unbiased, regardless of the distribution of the target population and regardless of the sampling effort allocated to each stratum (assuming that at least one sample is taken per stratum). Because of this property, estimates of the stratified mean in different periods are directly comparable if the sampling allocation is altered, or even if the strata boundaries are altered (assuming that the sampling frame remains the same). Sampling effort is often allocated to each stratum proportionally to the weight for each stratum ($n_h = W_h \times n$). Although this generally results in higher precision than simple random sampling, the sampling effort can be allocated to minimize the variance of the resulting estimate. Three general rules have been developed to guide the allocation of sampling effort to minimize the SE of the stratified mean. Using these rules, greater sampling effort should be allocated to strata where (1) the stratum is larger, (2) the stratum has a larger variance, or (3) sampling cost per unit is less expensive in the stratum. If the cost per sample c_h varies among strata, then the optimal allocation of sampling effort is (Cochran 1977; Lohr 1999)

$$n_h = n\left(\frac{\frac{N_h s_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}}}\right), \tag{3.13}$$

where n is the total sample size, N_h is the total number of units in stratum h, s_h is the standard deviation (SD) in stratum h, and c_h is the cost per sample in stratum h. If the cost per sample is the same in all strata, the optimal allocation, termed the Neyman allocation, is (Cochran 1977)

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right). \tag{3.14}$$

If variances are specified correctly, the Neyman allocation will always give estimates with smaller SEs than will proportional allocation because larger samples will be drawn from strata with larger variance (s_h in equations [3.13] and [3.14]), thereby reducing the SEs of the stratum means ($\mathrm{SE}\overline{y}_h$ in Table 3.1), which are inversely related to sample size.

3.2.3.4 *Estimation of Proportions*

Estimates of the proportion of sampling units in a population that fall into a defined class are computed much like the stratified mean. Essentially, the proportion is estimated for each stratum using the formula for a simple random sample and the stratum-specific proportions are combined using the stratum weights (Table 3.2; Cochran 1977). Similarly, the SE of the estimate for the proportion in the entire population is a weighted sum of the individual stratum variances (Table 3.2).

3.2.4 Cluster Sampling

In cluster sampling, the population is divided into primary sampling units (clusters) and secondary sampling units (elements). In an example of sampling fish, the secondary sampling units are individual fish and the primary sampling units are groups of fish as might be caught together in a net (Figure 3.2). In an example of sampling grid locations, the secondary sampling units are the individual grid locations and the primary sampling units are blocks of four grid locations (Figure 3.2). In cluster sampling, the primary units are selected independently at random, which in the fish example may be thought of as having randomly selected netting locations. Cluster sampling is single stage if each element (e.g., individual fish) in each cluster (e.g., net) is included in the sample and two stage if only a subsample of each element from each cluster is included in the sample (e.g., individual fish are subsampled from each net). Cluster sampling is distinguished from other designs in that the primary units are sampled independently but the secondary units are potentially correlated. Put another way, fish caught in a net may not be independent because they may be more similar to each other than to randomly selected fish from the entire population.

Cluster sampling is commonly used when the sampling frame is difficult or impossible to construct or the sampling process naturally results in clusters of secondary units. For example, when fish are collected with nets set at random points on a grid (as described in Box 3.1), the net is the primary sampling unit, and the individual fish collected are secondary units. Although catch per net (in numbers or weight) is appropriately treated as coming from a simple random sampling design, the mean weight of fish estimated from this sampling design should be treated as a cluster sample because individual fish within a net may not be sampled independently.

In practice, cluster sampling often results in a situation in which individual elements within each cluster are similar and differences in the means are larger among clusters than within clusters. In the above example, a truly random sample of individual fish in a lake would be very difficult to obtain because trying to

collect one fish at a time would be very inefficient and would likely lead to a much smaller sample size than would using nets that can capture multiple fish. In this case, redundant information is provided by each fish measured (because of their similarity or correlation), and the precision of the overall mean is reduced relative to a simple random sample with the same sample size. Although this seems like a poor sampling strategy relative to simple random sampling, the advantage of cluster sampling design is that sampling is often less expensive and a substantially greater sample size can be obtained.

3.2.4.1 *Single-Stage Cluster Sampling*

In cluster sampling, estimates of several different quantities can be obtained (Box 3.5). For example, when individual fish caught in randomly placed nets are counted and weighed, estimates of the mean number of fish caught per net can be obtained by the usual estimator for simple random sampling. In addition, the mean total weight of fish caught per net can be estimated, but this quantity is an example of a cluster total. Cluster totals are sometimes interpretable statistics (such as presented here) but in other cases are hard to visualize or interpret. If, for example, the lengths of individual fish were measured instead of weight, the cluster total would represent the total length of all the fish caught, a statistic that is of little use. When the statistic of interest focuses on cluster totals, single-stage cluster sampling reduces to a simple random sample for which each cluster total is treated as a single observation.

The mean weight (or length) of individual fish is an example of the mean per secondary unit, which is another statistic that can be computed in cluster sampling. A further complication of single-stage cluster sampling is that simpler formulae may be used when the clusters are of equal size (i.e., the number of secondary units is equal in all clusters) than when the clusters are of unequal size. In most fisheries applications, clusters are of unequal size, so we will emphasize the formulae relevant to such situations.

3.2.4.2 *Estimation of the Mean per Secondary Unit*

As indicated above, the process for estimating mean cluster totals follows simple random sampling (Box 3.5). Estimating the mean per secondary unit, such as the mean weight of individual fish, is conceptually related to stratified random sampling because the mean for each cluster is weighted by the number of secondary units in each cluster. The principal difference, however, is that all strata are sampled in stratified random sampling, whereas only a sample of all clusters is selected in cluster sampling. Estimating the mean per secondary unit is also related to ratio estimation because cluster totals and numbers of elements (secondary units) in each cluster are both random variables, and estimation of the mean per secondary unit naturally uses the number of elements (secondary units) as a divisor (Table 3.1). The mean per secondary unit is simply the sum of cluster totals divided by the total number of secondary units (Table 3.1). Estimating the SE is similar to the procedure for estimating the SE of a ratio, where each cluster total is the y variable and the number of secondary elements in each

Box 3.5 Example of Cluster Sampling

Five throw nets were deployed at random locations along the shoreline of a lake to collect age-0 bluegill. Greater sampling effort would usually be required, but data from these five nets are used to illustrate the procedure. The weight (g) of each of the age-0 bluegill was measured. The goal of sampling was to estimate the mean biomass of age-0 bluegill per net, the mean catch per net, and the mean weight of individual age-0 bluegill.

Table Catch per net and weight per individual of age-0 bluegill caught in five throw nets. Mean fish weight is given by for which computations are shown below table.

Measure and summary statistic	Net 1	Net 2	Net 3	Net 4	Net 5
Catch (M_i)	10	5	7	0	3
Weight (g)	0.495	0.319	0.514		0.610
	0.391	0.419	0.497		0.572
	0.274	0.503	0.374		0.681
	0.470	0.451	0.457		
	0.309	0.491	0.388		
	0.369		0.521		
	0.381		0.539		
	0.308				
	0.420				
	0.326				
Cluster total (y_i)	3.743	2.183	3.290	0	1.863
$(y_i - \hat{R}M_i)^2$	0.47197	0.00102	0.03572	0	0.28516
Mean weight per fish per net	0.374	0.437	0.470		0.621

The mean catch per net is

$$\overline{M} = (10 + 5 + 7 + 0 + 3)/5 = 5.0 \text{ fish.}$$

The mean cluster total (mean biomass per net) is

$$\overline{y}_i = (3.743 + 2.183 + 3.290 + 0 + 1.863)/5 = 2.216 \text{ g.}$$

The mean weight per fish, the SE of the mean, and the CLs are

$$\text{Mean fish weight} = (\hat{R}) = \frac{\sum y_i}{\sum M_i} = \frac{3.743 + 2.183 + 3.290 + 0 + 1.863}{10 + 5 + 7 + 0 + 3} = 0.443;$$

$$\text{SE}(\hat{R}) = \frac{1}{\sqrt{n}\,\overline{M}}\sqrt{\frac{\sum (y_i - \hat{R}M_i)^2}{n-1}} = \frac{1}{\sqrt{5}\cdot 5}\sqrt{\frac{0.47197 + 0.00102 + 0.03572 + 0 + 0.28516}{5-1}} = 0.040;$$

$$\text{Lower CL} = \hat{R} - t_{\alpha, n-1}\text{SE}(\hat{R}) = 0.443 - 2.776 \cdot 0.040 = 0.332; \text{ and}$$

$$\text{Upper CL} = \hat{R} + t_{\alpha, n-1}\text{SE}(\hat{R}) = 0.443 + 2.776 \cdot 0.040 = 0.554.$$

(Box continues)

Box 3.5 *(continued)*

Now, instead of treating the data as a single-stage cluster sample, consider the situation where the same number of fish per net are weighed but 48 fish are caught in Net 1 and 20 fish are caught in Net 3, thereby leading to a two-stage cluster sample with different numbers of fish caught in each net.

Table Catch per net and weight per individual of age-0 bluegill in five throw nets. Computation of the mean weight per secondary unit ($\bar{\bar{y}}$) is given below table.

Measure and summary statistic	Net 1	Net 2	Net 3	Net 4	Net 5
Catch (M_i)	48	5	20	0	3
Weight (g)	0.495	0.319	0.514		0.610
	0.391	0.419	0.497		0.572
	0.274	0.503	0.374		0.681
	0.470	0.451	0.457		
	0.309	0.491	0.388		
	0.369		0.521		
	0.381		0.539		
	0.308				
	0.420				
	0.326				
Estimated cluster total ($\hat{y}_i$)	17.966	2.183	9.400	0	1.863
Mean weight per fish per net ($\bar{y}_i$)	0.374	0.437	0.470		0.621
$M_i^2(\bar{y}_i - \bar{\bar{y}})^2$	3.5044	0.0144	1.2996	0.0000	0.3894

The mean per secondary unit (mean fish weight) is

$$\bar{\bar{y}} = \frac{\sum \hat{y}_i}{\sum M_i} = \frac{17.966 + 2.183 + 9.400 + 0 + 1.863}{48 + 5 + 20 + 0 + 3} = 0.413.$$

The SE of the mean per secondary unit is approximated by

$$\mathrm{SE}(\bar{\bar{y}}) = \sqrt{\frac{\sum_{i=1}^{n} M_i^2(\bar{y}_i - \bar{\bar{y}})^2}{n\bar{M}^2(n-1)}} = \sqrt{\frac{3.5044 + 0.0144 + 1.2996 + 0 + 0.3894}{5 \cdot 15^2 \cdot (5-1)}} = 0.034.$$

cluster is the x variable (Table 3.1). Confidence intervals are obtained as in estimating ratios using the t-distribution.

3.2.4.3 *Two-Stage Cluster Sampling*

In two-stage cluster sampling, a simple random sample of n clusters is selected and then a simple random sample of the elements (secondary units) is subsampled

from within each sampled cluster. In our example of fish in nets (Figure 3.2), the secondary sampling units (fish) are subsampled from the primary sampling units (nets). This differs from single-state cluster sampling, where secondary sampling units are all sampled completely rather than subsampled. Two levels of sampling are employed, so means and variances at two levels are defined. First, the mean per secondary unit within the ith cluster is estimated as in Table 3.1 (Box 3.5; Cochran 1977), where y_{ij} is the measured value for the jth element in the ith cluster and m is the number of elements sampled within each cluster (which has M_i secondary units). From this, the total for each cluster and the mean cluster total are estimated as in Table 3.1 (Box 3.5; Cochran 1977). From the estimated cluster totals, the overall mean per secondary unit is estimated as in Table 3.1 (Box 3.5; Cochran 1977). The variance of the overall mean includes two components that represent the variation between clusters and the variation due to subsampling within clusters. When the number of clusters sampled is small relative to the number of clusters in the population, the SE of the mean per secondary unit can be approximated as in Table 3.1 (Cochran 1977).

3.2.5 Systematic Sampling

In systematic sampling, all sampling units in the population are arranged in a sequence, and then from a random starting point every kth sampling unit is included in the sample. Systematic sampling is often used for ease of execution and convenience. Also, systematic samples are usually spread more evenly over the population, so population attributes may be estimated more precisely than with simple random sampling. However, a major difficulty with systematic sampling based on a single starting point is that the variance and SE of the estimates cannot be directly determined. This occurs because systematic sampling with a single starting point is equivalent to cluster sampling with just one cluster being sampled (the samples are not independent). One way of alleviating this problem is to take a systematic sample with two or more randomly selected starting points. For example, a lake could be divided into grids and several rows of grids could be randomly selected as starting points of evenly spaced grids (Figure 3.2). When systematic sampling is implemented with multiple starting points, the formulae for single-stage cluster sampling apply, with each group of observations associated with each start point treated as a cluster.

Systematic sampling with a single starting point should be avoided whenever sample units are ordered in a linear or nonlinear pattern. Under such circumstances, stratified random sampling produces more precise estimates because strata can be constructed to account for the pattern in the sample units and thereby reduce within-stratum variance that would not be accounted for by systematic sampling (Cochran 1977). Therefore, the choice of systematic sampling versus simple random sampling or stratified random sampling needs to be judged on a case-by-case basis.

The mean value for a single-starting-point systematic sample is estimated in the same way as for a simple random sample (Table 3.1), that is, as the sum of the

observations divided by the number of observations. Assuming that the sampling fraction is relatively small, estimates of the mean using systematic sampling are unbiased, for the same reason that estimates of the mean using simple random sampling are unbiased (section 3.2.2.1). A small amount of bias may occur when the sampling fraction is large, if the number of sampling units in the population (N) is not evenly divisible by the sample size (n) because some units would have a lower probability of being included in a sample than would others. When multiple starting points are used, the mean is estimated as it would be for a single-stage cluster sample (Table 3.1) and is also unbiased if the sampling fraction is relatively small.

Several methods have been developed for approximating the SE for a systematic sample with a single starting point, but we do not recommend their use because they can lead to strongly biased estimates of the SE (Cochran 1977). With two or more starting points, valid estimates of the SE of the mean can be obtained using a single-stage cluster sampling approach (Box 3.6).

3.2.6 Model-Based Estimators

All of the designs we have discussed to this point have focused on sampling a single variable or attribute of interest. Further, all of the designs, when properly implemented, are designed to be unbiased for estimating the mean and proportions. However, in many fisheries investigations, several variables are of interest. Further, relationships among the variables measured provide an opportunity to extract more information than is provided by each variable alone. A familiar example is sampling fish when length and weight are both measured on a subset of fish collected and only length is collected on the remaining fish. Because the two are related, we can infer the weight of fish where only length data are collected. This situation is an example of a model-based design.

Model-based designs are limitless in their variations, given the number of variables that can be measured and the number of relationships among variables that might be considered. This being the case, we will describe a model-based design that illustrates a commonly used approach. The key advantage of model-based designs is that the additional information contained in auxiliary variables can substantially improve precision of estimates. Improvements in precision, however, come at the cost of losing the property of being design unbiased. When choosing models (such as a linear regression) to represent a relationship among variables, the right model is often uncertain. Thus, if the wrong model is chosen, estimates of the mean or proportion can be biased. This is not to say that such an approach is necessarily worse than using a sampling design that is design unbiased. In some situations, the gains in precision may more than offset the bias introduced by having the wrong model. As indicated earlier, precision is appropriately expressed as mean square error (MSE) when bias is present. From equation (3.1), an estimator that reduces the variance component faster than the bias^2 term produces estimates with a smaller MSE. In addition, the concept of MSE applies to situations that are unbiased, but the bias term drops out (being equal to zero).

Box 3.6 Example of Systematic Sampling with Two Starting Points

The width of a stream was measured at sampling locations arranged every 20 m from two random starting points, with 15 points sampled for each random starting point.

Table Stream width measurements based on systematic sampling with two starting points.

Starting point 1		Starting point 2	
Distance upstream (m)	Width (m)	Distance upstream (m)	Width (m)
3	6.1	4	10.3
23	11.4	24	6.5
43	13.7	44	9.0
63	11.3	64	7.6
83	11.7	84	6.3
103	13.3	104	12.0
123	12.1	124	6.2
143	11.5	144	13.1
163	6.4	164	10.2
183	34.8	184	26.5
203	31.7	204	32.1
223	27.2	224	28.6
243	26.7	244	29.2
263	24.1	264	33.5
283	10.1	284	6.0

The estimated mean width and its associated SE are

$$\text{Mean width} = (\hat{R}) = \frac{\sum y_i}{\sum M_i} = \frac{6.1 + 11.4 + 13.7 + \ldots + 33.5 + 6.0}{15 + 15} = \frac{489.2}{30} = 16.31, \text{ and}$$

$$\text{SE}(\hat{R}) = \frac{1}{\sqrt{n}\,\bar{M}} \sqrt{\frac{\sum (y_i - \hat{R}M_i)^2}{n-1}} = \frac{1}{\sqrt{2} \cdot 15} \sqrt{\frac{56.347 + 56.347}{2-1}} = 0.50.$$

Regression or double sampling. A commonly used model-based approach is regression sampling, sometimes referred to as double sampling. In this method, a sample is collected where the auxiliary variable (x_i = independent variable) is measured on all units. A subsample is then selected where the variable of interest (y_i = dependent variable) is also collected. A linear regression between the two variables is formed, and the regression is used to incorporate the information contained in the auxiliary variable into the estimate of the mean for the variable of interest. Typically, this approach is implemented when the auxiliary variable is much less expensive to measure than is the variable of interest, thereby allowing for a greater sample size. An example of this situation is that visual estimates of stream width can be collected much more rapidly (and hence, less expensively) than can actual

measurements of stream width. Thus, the precision of estimates of the mean width of a stream may be improved by taking many visual estimates of stream width while measuring only a subsample of sites to provide a calibration, via linear regression, between measured and visually estimated stream width (Box 3.7). Another example is that the percent water content of fish tissue is easily determined by weighing, drying, and reweighing the tissue and provides a reasonable predictor of the fat content of the tissue (Hartman and Brandt 1995). Fat content is often determined on dry tissues and requires time-consuming extraction by use of solvents in sophisticated equipment. Thus, the mean fat content may be estimated by collecting many measurements on percent water content and only a few concurrent measurements of actual fat content.

An important assumption of this method is that a linear relationship exists between the two variables. If this assumption does not hold, estimates of the mean can be biased, thereby offsetting any gain in precision. Another important consideration is that the value of the y_i for sampling units (e.g., stream width at a particular location or individual fish for fat content), where only the auxiliary (x_i = independent) variable is measured, can be estimated with the regression equation. Therefore, the precision of individual sample units may be relatively poor because the prediction does not match the value that would be obtained by direct measurement. However, the precision of the estimated population mean will usually be increased because of increased sample size. This general principal of sampling reflects the fact that the SE of the mean is inversely related to the sample size; that is, an increased sample size reduces the SE of the mean, which increases precision of the estimated mean.

The formula for estimating the mean using a double-sampling approach is given in Table 3.1, and an example of the application of this method is given in Box 3.7. This example only brushes the surface of the diversity of applications of model-based designs. For a more in-depth treatment see Draper and Smith (1981) for linear regression models and Seber and Wild (1989) and Bates and Watts (1988) for nonlinear regression models.

3.2.7 Advanced Designs

The designs we describe and illustrate above are intended to provide a basis for the appropriate design and analysis of sampling programs. These relatively simple designs provide useful approaches in many situations, and are sufficient for many of the questions posed by fisheries scientists. For more complex situations, these designs can be combined and adapted to suit the needs of the investigator. Many other designs have been developed for specialized situations (e.g., hydroacoustic surveys). Thompson (1992) covers additional designs. One extension to normal sampling designs we would like to highlight are adaptive designs. In adaptive designs, additional sampling is concentrated near sampling points where something interesting happens. For example, in surveys trying to estimate the density of rare species, additional sampling can be concentrated near sampling locations where the rare species is found. Adaptive designs can provide improved precision of density estimates and also have the advantage that more specimens can be collected for length,

Box 3.7 Example of Regression or Double Sampling

The percent of coverage by woody material was visually (Table, column 2) estimated at 25 randomly selected points along a stream (Table, column 1), and the actual amount of woody material coverage was measured at 10 of these points (Table, column 3), with the goal of estimating mean woody material coverage for this reach. The regression between the visually estimated coverage and the measured coverage gave the following equation:

$$\text{Measured coverage} = 7.2937 + 1.0357(\text{estimated coverage}).$$

Table Measurement and visual estimation of percent of woody material coverage along a stream.

Stream location and mean	Woody cover (%) Visually estimated (x_i)	Measured (y_i)	$(y-\bar{y})^2$	$(y-\bar{y})(x-\bar{x})$	$(x-\bar{x})^2$
1	30	34	29.16	5.4	1
2	40	52	158.76	113.4	81
3	60	66	707.56	771.4	841
4	20	26	179.56	147.4	121
5	20	21	338.56	202.4	121
6	20	25	207.36	158.4	121
7	30	38	1.96	1.4	1
8	0	14	645.16	787.4	961
9	40	54	213.16	131.4	81
10	50	64	605.16	467.4	361
11	40				
12	70				
13	80				
14	20				
15	50				
16	80				
17	50				
18	20				
19	80				
20	80				
21	40				
22	90				
23	30				
24	40				
25	40				

The estimated mean coverage using double sampling is

$$\bar{y}_{\text{reg}} = \bar{y} + b(\bar{X} - \bar{x}) = 39.4 + 1.0357 \cdot (44.8 - 31.0) = 53.69,$$

where $\bar{x}$ = the mean of visual estimates for the 10 subsampled stream locations, $\bar{y}$ = the mean of measured values for the 10 subsampled stream locations, and $\bar{X}$ = the mean of visual estimates for all 25 stream locations.

The SE of this estimate is

$$\text{SE}(\bar{y}_{\text{reg}}) = \sqrt{\frac{1}{n(n-2)}\left\{\sum(y_i-\bar{y})^2 - \frac{[\sum(y_i-\bar{y})(x_i-\bar{x})^2]}{\sum(x_i-\bar{x})^2}\right\}} = \sqrt{\frac{1}{10(10-2)}\left\{3086.4 - \frac{[2786]}{2690}\right\}} = 1.58.$$

age, or other biological variables. Because sampling is concentrated near hot spots, the additional samples are not independent of the original sampling locations, and specialized formulae must be used to reduce or remove biases that would occur if the data were treated as coming from a sample of independent observations.

3.3 EXPERIMENTAL DESIGNS

Developing an adequate design to an experiment is perhaps the trickiest and most difficult task that a fisheries scientist faces. Fisheries scientists must balance the need to control the experiment to understand better the results with the need to assure that the design is relevant to natural systems (Yandell 1997). Many experimental designs used by fisheries scientists come from disciplines such as agriculture, where experiments are easier to develop and factors are easier to manipulate. By necessity, fisheries scientists often rely on experimental units, such as lakes or fish, over which they have little control. Lack of control over experimental units is an important reason why developing a sound experimental design and analysis is critical to the success of any fisheries experiment.

The first step in designing an experiment is to develop a clear statement of objectives for the experiment (Cochran and Cox 1957; Yandell 1997). This step should include the questions that are being asked in the experiment or the hypotheses being tested. In the first section of this chapter, we focused on sampling designs for which the goal is generally to describe the attributes of a population. When an experiment is conducted, the goal is often to answer questions focusing on the response to a treatment or to determine the influence of natural or anthropogenic factors. Questions should be clearly focused and reasonably answered. For example, the question, Are more large largemouth bass present after an increase in the minimum length limit? is too vague to be answered through an experiment. A clearer way to phrase the question is, Did the population density of largemouth bass longer than 35 cm increase in lakes where the minimum length limit was increased from 25 cm to 35 cm? The second question is more specific than the first and helps to determine how the experiment should be designed. The second question can now be turned into a testable statistical hypothesis. Hypothesis testing is the formal approach that is used to assess whether evidence supports your question. Hypotheses are set up in two competing claims, the null hypothesis (H_0) and the alternative hypothesis (H_a). The statistical test is set up either to support or not to support the null hypothesis (see Chapter 1). In our example, we could formulate the null hypothesis as the number of largemouth bass longer than 35 cm is the same in lakes with the increased size limit and in lakes with no change in the size limit. The alternative hypothesis is then the number of largemouth bass longer than 35 cm is not the same in lakes with the increased size limit and in lakes with no change in the size limit. Remember the statistical test is set up to support or not support the null hypothesis, so we can either fail to reject or reject that the number of largemouth bass greater than 35 cm differs in lakes with and without the size limit. However, we cannot conclude that the alternative hypothesis is true.

After specifying the research question, the experiment should be described in terms of experimental units, units being sampled (subsamples being taken from each experimental unit), number and type of treatments, number of replicates per treatment, and target population (Cochran and Cox 1957; Brown and Austen 1996). For the largemouth bass question in the previous paragraph, the experimental units are lakes and the treatment is a regulation change. To complete the study design, the fishery scientist would need to specify a population of lakes that would be subjected to the regulation change, to select randomly a sample of lakes for study (lakes = replicates), and to choose a sampling plan with an appropriate gear for capturing largemouth bass longer than 35 cm. To differentiate effects of the regulation change from background variation, sampling would also need to begin before implementing the rule change (temporal controls) and need to include a sample of lakes on which the regulation change was not implemented (spatial controls). Finally, a well-designed experiment should also have an outline of the method of analysis to be applied to the data after the experiment is completed (Cochran and Cox 1957). Methods of analysis for experimental designs are provided in box examples below.

3.3.1 Completely Randomized Design

A completely randomized design is a design in which treatments are applied to the experimental units completely at random, so that each experimental unit has an equal probability of being selected for each treatment (Figure 3.3; Cochran and Cox 1957). For example, we may want to determine how different creel limits affect angler catch rates for walleye in Wisconsin lakes (Box 3.8). In this example, the statistical population being considered includes lakes in Wisconsin that contain walleye, and the experimental unit is an individual lake. Although angler catch rates are determined through a creel survey program (thus constituting a subsample of all anglers), we will treat catch rates as a single observation per lake. In this example, we decided a priori that we were interested only in one factor (i.e., bag limits) and we were interested in three levels of this factor: bag limits of one, two, or five walleye per day. The experiment was designed to have six replicates, seven replicates, and nine replicates for the one, two, and five walleye bag limit treatment, respectively, for a total of 22 lakes in the experiment. Each lake was assigned a treatment level at random, completing the experimental design.

As with simple random sampling designs described earlier, completely randomized designs provide a basic standard against which to compare other designs. Completely randomized designs have the advantage of allowing complete flexibility in the number of treatments and replicates allowed for the experiment (Cochran and Cox 1957). Further, statistical analysis is relatively easy regardless of the number of replicates and treatments (Cochran and Cox 1957). This holds true even when treatments or data are missing in the experiment (Cochran and Cox 1957), a common problem with large-scale field experiments. Finally, completely randomized designs have the advantage of maximizing the degrees of freedom for

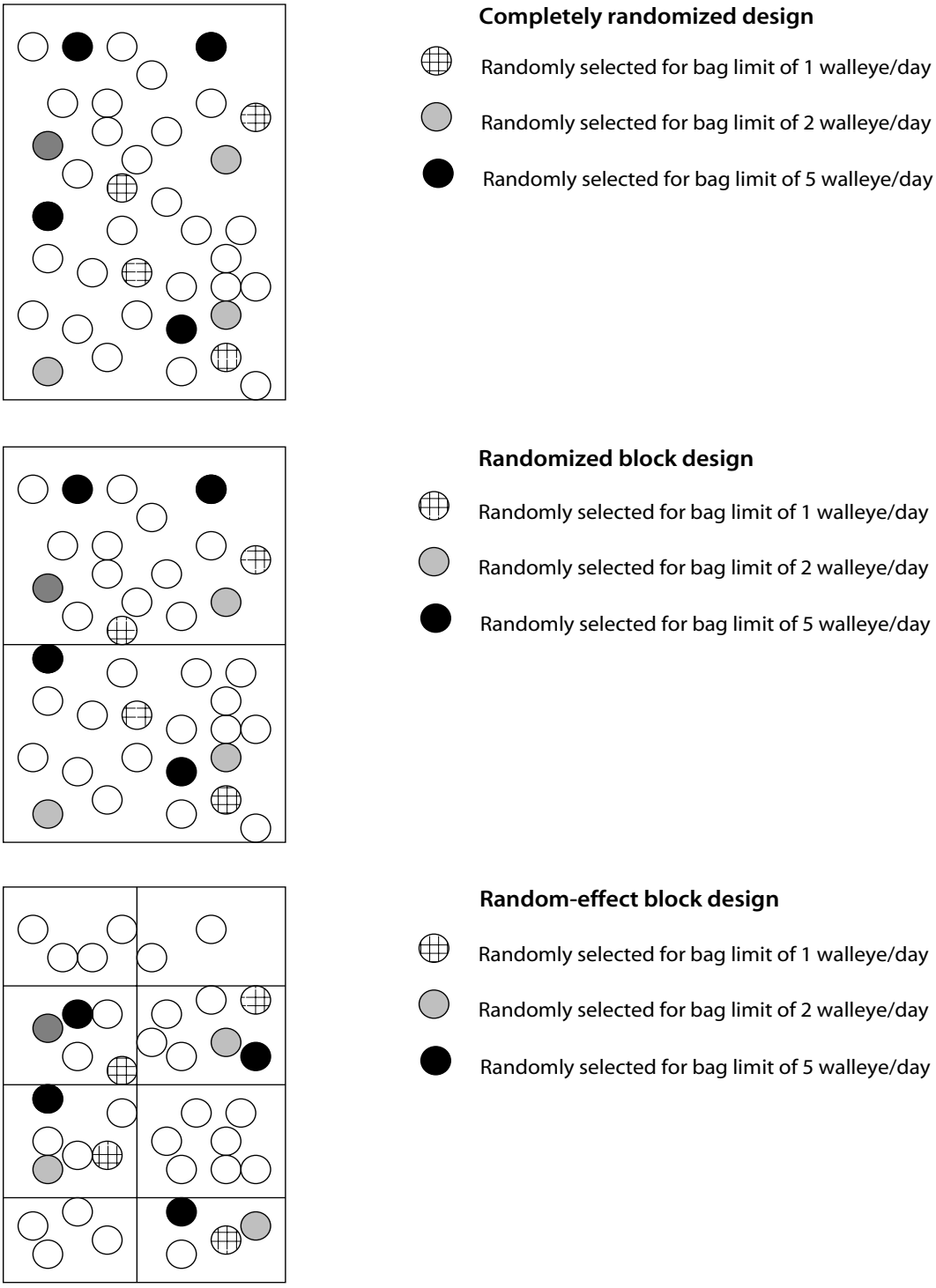


Figure 3.3 Examples of three experimental designs for testing how fish populations in lakes (circles) in a landscape (rectangles) respond to the application of three different daily bag limits in an angling fishery.

Box 3.8 Example of a Completely Randomized Design

The goal of this study was to determine if walleye catch rates differed among Wisconsin lakes with different daily bag limits (Beard et al. 2003). From the thousands of lakes in Wisconsin with walleye populations, six lakes were randomly chosen to have a bag limit of one walleye per day, seven lakes were randomly chosen to have a bag limit of two walleye per day, and nine lakes were randomly chosen to have a bag limit of five walleye per day. For this analysis, the designation of North or South was ignored. A fixed-effects general linear model (GLM, implemented in SAS; SAS 2005) was used for the analysis of these data.

Program

```
   * This data step reads the following information into a data set named
walleye;
   data walleye;
   input lake $ region $ bag_limit catch;
   cards;
   Willow            North          1         2.21
   Mud               North          1         2.32
   Pine              North          1         2.74
   Bass              North          2         2.23
   Perch             North          2         2.25
   Twin              North          2         1.40
   Park              North          2         2.36
   Mendota           North          5         1.78
   Silver            North          5         1.64
   Manistee          North          5         1.97
   Fox               North          5         1.99
   McGee             South          1         2.70
   Deep              South          1         3.63
   Round             South          1         2.82
   Long              South          2         3.09
   Portage           South          2         3.63
   Indian            South          2         2.82
   Wolf              South          5         2.20
   Gull              South          5         1.74
   Black             South          5         2.85
   Goose             South          5         3.01
   Fletcher          South          5         1.72
   ;
   run;
   *These statements call the GLM procedure in SAS, declaring the variable
bag_limit to be a categorical variable, and catch to be the continuous re-
sponse variable. The lsmeans statement requests least-squares means and
standard errors of catch for each level of bag_limit;
   proc glm;
   class bag_limit;
   model catch=bag_limit;
   lsmeans bag_limit/stderr;
   run;
```

(Box continues)

Box 3.8 *(continued)*

Results and Interpretation

Table Results of the GLM procedure in SAS for the dependent variable catch of walleye in 22 lakes with three different bag limits. Abbreviations are given for coefficient of variation (CV), mean square error (MSE), sum of squares (SS), and least-squares means (LSMEAN).

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	2	1.62317576	0.81158788	2.43	0.1153
Error	19	6.35613333	0.33453333		
Corrected total	21	7.97930909			
R^2	0.203423	Root MSE	0.578389		
CV	23.96337	Catch mean	2.413636		

Source	*df*	Type I SS	Mean square	*F*-value	*P* > *F*
Bag limit	2	1.62317576	0.81158788	2.43	0.1153

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Bag limit	2	1.62317576	0.81158788	2.43	0.1153

Least-Squares Means

Bag limit	Catch LSMEAN	SE	*P* > |*t*|
1	2.73666667	0.23612614	<0.0001
2	2.54000000	0.21861033	<0.0001
5	2.10000000	0.19279619	<0.0001

The results of the analysis indicate that the daily bag limit had little effect on walleye catch rate. Although the point estimates of mean catch rate (catch LSMEAN) differ somewhat among bag limits, the differences were not greater than would be expected by random chance. The *F*-value for the entire experiment was 2.43 and the resulting *P*-value was 0.1153, which is greater than the 0.05 alpha value commonly used when testing for significant differences among means. Therefore, we would conclude that bag limits had no significant effect on catch rates.

Sums of squares can be computed in several ways. The simplest to understand and the most widely reported in statistical analysis programs are type III SS, which are computed as the difference in SS between two nested models in which one term is left out. Thus, the SS for each term is simply the difference in SS between the full model with all terms present and a reduced model with the term of interest absent. In contrast, SAS also reports type I SS, which are computed as the difference in SS between hierarchical models in which each term is dropped in sequence, beginning with the right-hand term and proceeding to the left. Thus, the SS for each term depend on the order in which the terms are specified by the user. Type III SS are independent of the order in which terms are specified in the model, so are generally preferred over type I SS.

analysis (Cochran and Cox 1957), thus maintaining statistical power when the number of replicates per treatment is small.

The main disadvantage of experiments using a completely randomized design is that the power of these experiments to detect differences among treatments may be relatively low (analogous to simple random sampling, where the precision of point estimates may be low). Randomized designs are most commonly used in laboratory studies, where experimental units are relatively homogeneous, thereby increasing the power of the experiment. In field studies where experimental units vary greatly from unit to unit (Cochran and Cox 1957), variation among units may obscure systematic differences resulting from the treatment. One way to overcome large variation among units is to increase the number of replicates in the experiment, but this comes at additional monetary cost (Brown and Austen 1996).

Completely randomized designs have been used in fisheries management projects mostly where sites were homogenous or where differences among sampling units were not known. For example, Walsh et al. (2002) compared catches from prepositioned area electrofishing and electric seining at 12 randomly selected stream sites. Similarly, Kocik and Taylor (1994) placed brown trout and steelhead in randomly selected sites within an experimental stream to quantify their survival and growth. In both studies, sampling sites were assumed to be relatively homogenous, thereby minimizing variability not accounted for in the experiment. Although completely randomized designs are uncommon in fisheries, they can be useful in small pilot studies that will provide some information about the experimental unit for better design of a full-scale study. Before continuing on with more sophisticated designs, we consider the analysis of this relatively simple design, and discuss some of the critical considerations for data analysis.

3.3.1.1 *Analysis of Completely Randomized Design*

After the experiment has been conducted and data collected, how do we determine if the treatment(s) led to a response? One tool available is the general linear model (GLM), which contains the familiar analysis of variance (ANOVA) model. General linear model is a term used to refer to an entire class of models that are linear in their parameters (Yandell 1997; Montgomery 2001), which means that no parameter in the model is an exponent or is multiplied or divided by another parameter (Neter et al. 1996). The term general is used because both continuous and categorical variables can be used as predictor variables (Quinn and Keough 2002). In most of these models, we measure a response variable and then determine how this response variable is influenced by one or more predictor variables.

In our creel limit example, the treatments (or predictor variables) are fixed because the bag limits were determined prior to the start of the experiment and then applied according to the completely randomized design. Moreover, we treat the creel limits as categorical variables. This is in contrast to continuous variables (such as lake area), which we will discuss later. When analyzing data from this situation, we use what is called a fixed-effects GLM (Quinn and Keough 2002). The objective of our analysis is to determine whether variation in the means for different treatment levels differs more than would be expected by chance or if

"real" differences in catch rate are related to the bag limit imposed. The GLM for this case can be specified for $i = 1$ to p treatment levels and $j = 1$ to n replicates:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{3.15}$$

where y_{ij} are the observations, μ is the overall population mean of the response variable, α_i is the treatment effect for each level, and ε_{ij} is the unexplained variation among lakes (i.e., statistical error; Quinn and Keough 2002). In our example (Box 3.8), $i = 3$ treatments (bag limits of 1, 2, and 5) and $j = 6$, 7, and 9 replicates. The fixed-effects model can then be used to test the H_0 that all treatment level means (specified as μ_i) are the same:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu.$$

This can also be specified in terms of the test of treatment effects:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_i = 0.$$

An F-test is used to compare the variability among groups to the residual variability (F-ratio = mean squared error for main effects divided by the mean squared residual error), to determine if the observed differences in group means are greater than would be expected by chance. The observed F-ratio is compared to an F-distribution with the degrees of freedom in the numerator and denominator being those used for the two mean square errors. If H_0 is true, both group and residual mean square error should estimate the pooled population error term and the F-ratio should be 1 (Quinn and Keough 2002). In our bag limit example (Box 3.8), we would compare the F-ratio to an F-distribution with 2 df (3 treatments; $p - 1$) for the numerator and 19 df (22 observations – 3 treatments) for the denominator.

3.3.1.2 *Assumptions*

The most important assumption when sampling or performing any experiment is that the treatments are randomly applied to the experimental units (Sokal and Rohlf 1995). Failure to select samples at random or to apply treatments at random may result in biased results that are not representative of the true response (Sokal and Rohlf 1995). Applying treatments to lakes or rivers where the investigator suspects they will be most successful is tempting, but the results of the study will not be applicable to any other lakes or rivers. If the fully randomized experimental design is implemented, this insures that the random-selection assumption is satisfied.

Many statistical analyses assume that sample units or the selection of samples are independent (Sokal and Rohlf 1995; Brown and Austen 1996). That is, changes in one sample unit or one sample subject should not affect other sampling units or subjects. Treatments or subjects must therefore be spatially and temporally independent (Sokal and Rohlf 1995; Brown and Austen 1996). In our example

(Box 3.8), this assumption implies that the bag limit imposed on one lake has no effect on the catch rate of walleyes in nearby lakes. This assumption would be violated, for example, if anglers shifted their effort away from lakes where the creel limit was imposed to lakes where the bag limit was not imposed.

Repeated measurements of water bodies over time, which are often used to detect changes in fish populations caused by management actions (stocking, habitat manipulations, or regulations) in particular water bodies (often in relation to water bodies where the management action was not implemented) are not temporally independent. Lack of temporal independence is called time series bias and can lead to problems in estimating parameters (Walters 1985; Caputi 1988; Hilborn and Walters 1992; Myers and Barrowman 1996). To account for a lack of temporal independence among sample units, a repeated-measures design is often used (section 3.3.8). Lack of spatial independence also occurs in fisheries studies (as noted in the previous paragraph; also see examples in telemetry studies, Chapter 14, and watershed analyses, Chapter 18), and a variety of methods have been developed to account for spatial dependency.

Another important assumption of a GLM analysis is that the residual variance must be constant or homoscedastic among observations and treatments (Sokal and Rohlf 1995; Montgomery 2001). Variability among experimental units commonly increases with an increasing mean, thereby leading to heteroscedastic residual variance (Sokal and Rohlf 1995). Inequality of variance is generally diagnosed by using plots of the residuals against the predictor variable and predicted values and either Bartlett's or the modified Levene test (Montgomery 2001). When unequal variance occurs, the data are often transformed to equalize the variance, or the unequal variance is accounted for using a mixed model (section 3.3.4). For example, prior to transformation, catch data are often highly skewed in their distribution, so variance often differs among treatment levels (heteroscedastic residual variance; Figure 3.4). In contrast, after log transformation, catch data may be normally distributed and have equal variance among treatment levels (homoscedastic residual variance; Figure 3.4).

The final assumption necessary when performing many common statistical analyses is that the residual errors are normally distributed (Sokal and Rohlf 1995; Montgomery 2001). Although large departures from normality can significantly affect inferences from a GLM analysis (Montgomery 2001), this is perhaps the least important assumption because the central limit theorem states that with large sample sizes (e.g., greater than 30), estimates of model parameters often approximate a normal distribution regardless of the distribution of the data (Yandell 1997). Departures from normality are tested using normal probability plots and statistical tests such as the Shapiro–Wilk test or Kolmogorov–Smirnov test (Box 3.9; Sokal and Rohlf 1995). Each of these tests examines different aspects of departures from normality, so they sometimes provide conflicting insights. Our preference is for the Shapiro–Wilk test, which is more sensitive to departures in the tails of the distribution, though the Kolmogorov–Smirnov test is also useful because the test statistic, D, is a readily interpretable measure of the maximum difference between the observed and expected cumulative distributions. When

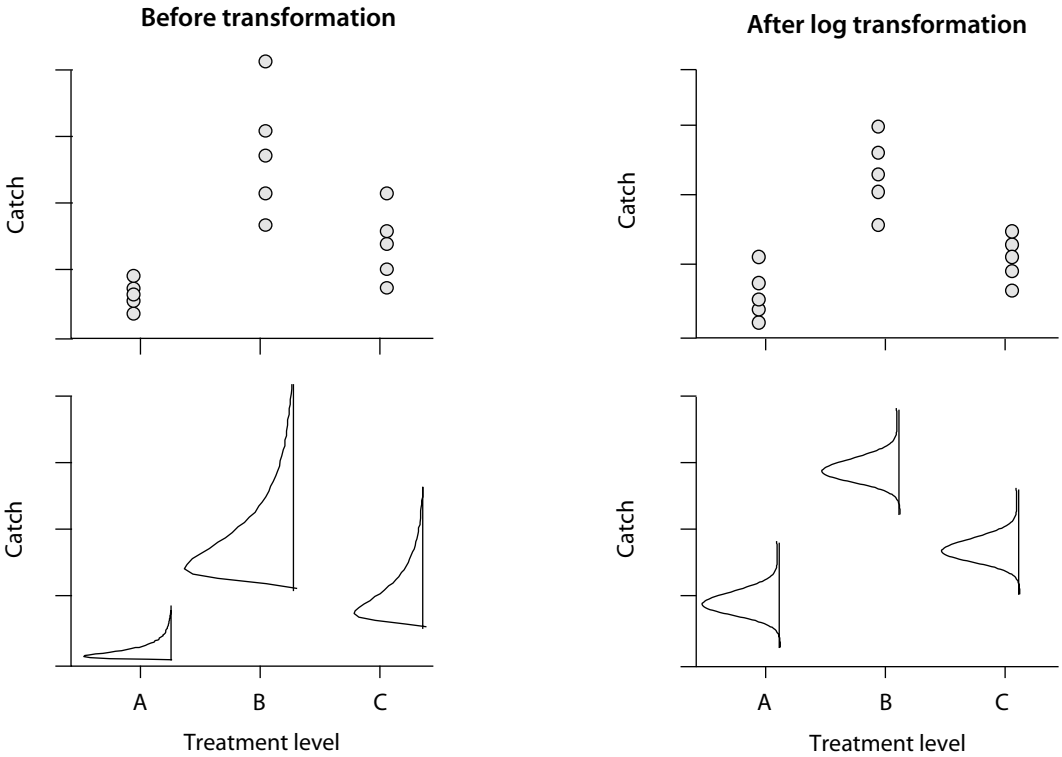


Figure 3.4 Changes to the distribution of catch data following log transformation. The upper panel in each pair of graphs illustrates the distribution of individual data points and the lower panel depicts the hypothetical statistical distribution from which data points were drawn.

examining distributions of residuals for potential violations of the assumption of normal distribution, more concern should be given to distributions with thinner or thicker tails than expected than to distributions that are skewed (Montgomery 2001). Analyses that proceed with nonnormal data will generally lead to fewer significant test results because of reduced power of the test (Montgomery 2001).

One important departure from normality is the presence of outliers, extreme values that lie well outside the distribution of the rest of the data (e.g., more than 3 SDs from the mean) and that are often caused by sampling problems or because the outliers belong to a population that differs from the target population (Montgomery 2001). Outliers can significantly affect the outcome of a statistical analysis and should be examined to determine if they are caused by sampling problems or because they come from another population. If follow-up investigation of an outlier reveals that the outlying datum was caused by a failure of the sampling protocol, the datum can be rejected from the analysis. However, outliers can also be extreme values of the target population so should not be rejected simply because they are outliers.

Box 3.9 Example of How to Test Errors (Residuals) for Normality

In an extension to the example in Box 3.8, the results of the analysis were augmented to examine the normality of residuals.

Program

```
   proc glm;
   class bag_limit;
   model catch=bag_limit;
   lsmeans bag_limit/stderr;
   *The following output request saves a new data set named model_resid,
saving residuals into a variable named resid;
   output out=model_resid r=resid;
   run;
   *These statements call the univariate procedure in SAS, requesting a
normality plot, normality test, and a q-q plot of the variable named resid;
   proc univariate plot normal;
   var resid;
   qqplot resid;
   run;
```

Results

Some output is not shown because it is not critical to this discussion; the pertinent results of this analysis follow.

Table The univariate procedure of SAS was used to evaluate the normality of residuals generated from the analysis in Box 3.8.

Moments

N	22	Sum weights	22
Mean	0	Sum observations	0
SD	0.55015726	Variance	0.30267302
Skewness	0.43350809	Kurtosis	0.01734467

Tests for Normality

Test	Test statistic	Statistic value	*P*-test	*P*-value
Shapiro–Wilk	W	0.933623	$P < W$	0.1460
Kolmogorov–Smirnov	D	0.155156	$P > D$	>0.1500
Cramer–von Mises	W^2	0.120727	$P > W^2$	0.0557
Anderson–Darling	A^2	0.700157	$P > A^2$	0.0602

(Box continues)

Box 3.9 *(continued)*

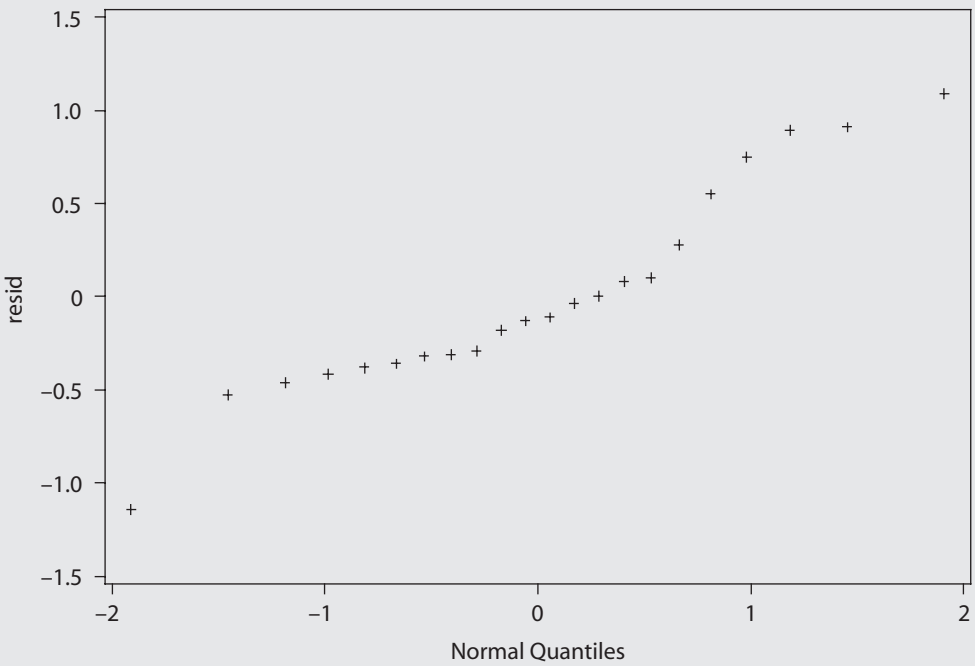


Figure Normal probability plot of residuals versus normal quantiles.

In this example, we used the univariate procedure in SAS to produce a normal probability plot and to provide statistical tests of normality. In a normal probability plot, we are looking for a relatively straight 1–1 line in our plot. In this case, the data show some deviation from a straight line, but none of the normality tests were significant, which indicates that the residuals did not differ significantly from a normal distribution.

When the assumptions of normality or equality of variability are not met, the data can often be transformed into a new scale for which the assumptions are satisfied. However, transformations should be logical and scientifically sound (Yandell 1997). For example, body weight of virtually all organisms will increase as an approximately cubic function of length, so variance in fish weight would likely also increase as an approximately cubic function of length (Brown and Austen 1996). Consequently, the use of a linear model to describe the relationship between weight and length would be incorrect. Transformation of weight and length into their logarithms ($\log_{10}$ or $\log_e$) permits the use of a linear model to estimate parameters of the weight–length relationship and eliminates heteroscedasticity of residual errors. Although transformations may help meet the assumptions of

the GLM, they may also bias parameter estimates (Hayes et al. 1995), so care must be taken in interpreting point estimates.

The selection of an appropriate transformation should rely on an examination of the distribution, mean, and variance of the data. When the mean of a sample is positively correlated with its variance (i.e., variance increases as the mean gets larger), a logarithmic transformation, either base 10 or base *e*, is often appropriate (Sokal and Rohlf 1995). When the mean and variance are similar and do not vary independently, which is often true of count data such as the number of fish caught in a net, the data should be transformed into their square roots to make the variance independent of the mean (Sokal and Rohlf 1995). When the distribution of the data has fewer observations at the mean and at the tails and more observations at intermediate regions than would a normal distribution (platykurtic), which is true of proportion or percentage data such as the percent of lake trout with sea lamprey wounds, the data should be transformed into their arcsines or arcsine square roots (Sokal and Rohlf 1995).

Another approach to transforming the data is to use a nonparametric statistical test. Many nonparametric tests use a rank transformation (i.e., ranking each observation) as a means of reducing the effects of outliers or nonnormality. Coverage of nonparametric methods is beyond the scope of this chapter (see Chapter 1), but they provide a useful suite of methods. As with other transformations, inferences drawn from analyses using a rank transformation do not strictly apply to the arithmetic mean. In the case of rank transformations, inference is generally based on the median, 50th percentile, as a measure of central tendency.

Following transformation, the results of any analysis should be evaluated to determine if the transformation was successful in correcting the observed problem. Also, transforming data to meet statistical assumptions may lead to a model that is not interpretable in the original scale of measurement (Draper and Smith 1981), which may render the model useless for its original purpose. For example, if data for a two-variable model are transformed to meet the assumption of normality, but data for the two variables are each transformed with a different transformation, the resulting statistical model cannot be back-transformed into the original measurement scales for interpretation. Thus, data should be transformed only in the context of an understandable model and its transformation. Sometimes, transformations will still not meet the assumptions necessary for the desired analysis. In such instances, a transformation that achieves approximately equal variances among samples is usually sufficient for analysis, even if the data are slightly nonnormal (Yandell 1997).

3.3.2 Randomized Block Design

One of the disadvantages of a completely randomized design is that natural variation among experimental units obscures the effect of the treatments, thereby reducing the statistical power to detect real differences when they occur. Although more replicates can be taken to overcome low power, a commonly used strategy is to subdivide the population of interest into more homogeneous groups or blocks.

For example, we may already know or suspect that the catch rate of walleye tends to be higher in northern Wisconsin than in southern Wisconsin (Box 3.10). Thus, if we take into account the location of the lake within the state, we can reduce the variability among units within a block. This is directly analogous to the increased precision of stratified random sampling over simple random sampling.

For many fisheries experiments, blocks are often created across time or space (Quinn and Keough 2002). The purpose of blocking is to reduce variability within each group (Cochran and Cox 1957) to estimate means more precisely and to increase power of tests of treatment effects (Quinn and Keough 2002). In a randomized block design, any number of treatments and replicates may be included in the design, and the statistical analysis is straightforward (Cochran and Cox 1957). For a randomized block design to be favored over a completely random design, the precision gained by blocking the treatments must offset the degrees of freedom lost when blocks are used (Yandell 1997). Randomized block designs can be used when blocks are missing, but completely randomized designs are usually better for testing treatment effects if the number of missing blocks is large (Cochran and Cox 1957).

The analysis of randomized block designs is similar to the fully randomized design, except that the effects of the blocking factor are included as an additional effect (Box 3.10). The statistical model for the randomized block design is

$$y_{ijk} = \mu + \alpha_i + \beta_k + \varepsilon_{ijk}, \tag{3.16}$$

here y_{ijk} are the observations, μ is the population mean of the response variable, α_i is the treatment effect for each level, β_k is the effect for each level of the blocking variable, and ε_{ijk} is the unexplained variation among experimental units. This model can then be used to test the null hypothesis that all treatment level means (specified as μ_i) are the same, after taking account for the effect of the blocking variable(s):

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu.$$

Randomized block designs are often used in fisheries. For example, Wilderbuer et al. (1998) compared catch per unit effort of various fish species collected by two different types of trawls (Wilderbuer et al. 1998). In this experiment, the two trawl types were simultaneously hauled, and each paired haul was considered as a block because of variability in catches between trawl runs (Wilderbuer et al. 1998). Similarly, Sammons and Bettoli (1999) examined variation in catch and mean length of largemouth bass caught by electrofishing, blocked by transects sampled. Whalen and LaBar (1994) used stream sections as blocks to compare survival and growth of Atlantic salmon stocked at different densities.

3.3.3 Analysis of Covariance

In the randomized block design, variability due to the effects of the categorical blocking variable(s), such as the designation of north and south in the example shown in Box 3.10, is used to remove the confounding effect of this variability.

Box 3.10 Example of a Randomized Block Design

In an extension to the example in Box 3.8, lakes were first blocked into northern and southern Wisconsin lakes, and then treatments were randomly assigned to lakes in each block. A randomized block design should include the blocking factor during the randomization process. The SAS program for this analysis is similar to a completely randomized design, except that block and an interaction term are included in the model. For brevity, the data are not repeated here.

Program

```
  *The following call to the GLM procedure indicates that bag_limit and
region are categorical predictor variables and catch is a continuous response
variable;
  proc glm;
  class bag_limit region;
  model catch=bag_limit region region*bag_limit;
  lsmeans bag_limit region/stderr;
  run;
```

Results and Interpretation

Results of the above analysis are as follows.

Table The GLM procedure for a randomized block design (blocks being northern versus southern lakes) with the dependent variable catch. This analysis is based on the data presented in Box 3.8.

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	5	4.83082242	0.96616448	4.91	0.0065
Error	16	3.14848667	0.19678042		
Corrected total	21	7.97930909			
R^2	0.605419	Root MSE	0.443599		
CV	18.37888	Catch mean	2.413636		

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Bag_limit	2	1.93517601	0.96758800	4.92	0.0216
Block	1	2.86174438	2.86174438	14.54	0.0015
Bag_limit*block	2	0.43851616	0.21925808	1.11	0.3523

Least-Squares Means

Bag_limit	Catch LSMEAN	SE	*P* > |*t*|
1	2.73666667	0.18109869	<0.0001
2	2.62000000	0.16940231	<0.0001
5	2.07450000	0.14878776	<0.0001

Block	Catch LSMEAN	SE	*P* > |*t*|
North	2.10944444	0.13498300	<0.0001
South	2.84466667	0.13765619	<0.0001

(Box continues)

Box 3.10 *(continued)*

In this analysis, the bag limit ($F = 4.92$; $P = 0.0216$) and block ($F = 14.54$; $P = 0.0015$) both appear to have an effect on angler catch rates. The interaction between these two factors does not appear to be significant ($F = 1.11$; $P = 0.3523$), which suggests that the effect of bag limits was similar in northern and southern Wisconsin lakes (blocks). Therefore, we can re-run the analysis without the interaction term in the model.

```
proc glm;
class bag_limit region;
model catch=bag_limit region ;
lsmeans bag_limit region/stderr;
run;
```

Results of the analysis without the interaction term for region*bag limit are similar to those for the model with the interaction term.

Table The GLM procedure for a randomized block design (blocks being northern versus southern lakes) with the dependent variable catch. This analysis does not include the interaction term of bag limits*region.

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	3	4.39230627	1.46410209	7.35	0.0020
Error	18	3.58700282	0.19927793		
Corrected total	21	7.97930909			
R^2	0.550462	Root MSE	0.446406		
CV	18.49515	Catch mean	2.413636		

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Bag_limit	2	1.95674263	0.97837132	4.91	0.0199
Block	1	2.76913051	2.76913051	13.90	0.0015

Least-Squares Means

Bag_limit	Catch LSMEAN	SE	*P* > |*t*|
1	2.73666667	0.18224431	<0.0001
2	2.59097810	0.16927875	<0.0001
5	2.06035036	0.14918152	<0.0001

Block	Catch LSMEAN	SE	*P* > |*t*|
North	2.10581833	0.13522076	<0.0001
South	2.81951176	0.13664751	<0.0001

Results of this analysis suggest that walleye catch rates differed significantly among daily bag limits ($F = 4.91$; $P = 0.0199$) and between northern and southern Wisconsin lakes ($F = 13.90$; $P = 0.015$). Importantly, differences in walleye catch rates between northern and southern Wisconsin lakes obscured the effect of daily bag limits when the data were analyzed using a fully randomized design (Box 3.8).

Often, however, potential blocking variables are not categorical but are measured on a continuous scale. For example, lakes can be arbitrarily categorized as "small" and "large" based on their surface area but could also be measured in surface area on a continuous scale. The GLM treats these variables as the independent variable in a regression, and inferences based on these models evaluate the effects of a treatment on the response variable, after accounting for the effect of continuous variable(s) on the response variable (Quinn and Keough 2002). This particular application of a GLM is often termed analysis of covariance (ANCOVA).

As in a randomized block design, one of the main advantages of using an ANCOVA design is that unexplained variability in the response variable is reduced, thereby providing greater statistical power to detect and estimate the effects of treatments. A further advantage is that the results provide insight into the effects of covariates and potential interactions with the treatment variable. Thus, a greater understanding of the experimental system can be obtained with ANCOVA. Another advantage of an ANCOVA design is that the value of the independent covariate is not known a priori but is determined at the time of the experiment, thereby reducing the need to have the entire sampling frame sorted into blocks prior to the experiment.

When using ANCOVA, several assumptions must be met. First, a linear relationship must exist between the response variable (e.g., weight) and the independent variable used as a covariate (e.g., length; Montgomery 2001; Quinn and Keough 2002). We must also assume that the covariate values are similar among treatments (Quinn and Keough 2002). The important implication of this assumption is that ANCOVA should not be used to correct for different values of the covariate in each group (Quinn and Keough 2002). For example, if initial lengths of fish in an experiment were different, we should not include initial lengths to correct for this difference (Quinn and Keough 2002). In regression, we must assume that the covariate, x, is fixed and measured without error (Quinn and Keough 2002).

Because continuous factors are included in ANCOVA, the statistical model looks somewhat different than the model for fully randomized or randomized block designs. Multiple expressions of the ANCOVA model are available, but we prefer

$$y_{ij} = \beta_0 + \alpha_i + \beta x_{ij} + \varepsilon_{ij}, \tag{3.17}$$

where y_{ij} are the observations, β_0 is the intercept for the regression between x and y for the population as a whole, α_i is the treatment effect for each level, β is the common slope for the regression between x_{ij} and y_{ij}, and ε_{ij} is the unexplained variation among experimental units. This model can then be use to test the H_0 that all treatment effects (α_i) are the same:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = 0.$$

An equivalent way of thinking about the H_0 is that all of the regressions between the response variable (y) and the covariate (x) have the same intercept. However, before testing the H_0, we must first test whether the slopes of the regression lines for all treatments are the same (i.e., all regression lines are parallel; Neter et al.

1996; Quinn and Keough 2002). To test the hypothesis that slopes are equal for all treatment levels, the interaction between the fixed effect and the continuous variable (covariate) is evaluated using the slope heterogeneity test (Box 3.11; Neter et al. 1996; Quinn and Keough 2002). If the interaction is significant, then the ANCOVA model (equation [3.17]) does not apply and separate regression models should be fit to each treatment level and then compared (Neter et al. 1996). If the interaction is not significant, then slopes are assumed to be equal, and the ANCOVA model (equation [3.17]) is estimated.

The use of ANCOVA is common in the fisheries literature. Many analyses include the effect of a covariate that is important for understanding the effects of one or more fixed treatments. This is especially common when the treatment effect may be influenced by growth in either length or weight. The other common use of ANCOVA in fisheries is to determine how data can be grouped. In many instances, an analyst will be uncertain whether the covariate in question affects the results of an outcome. For instance, Beard et al. (1997) used ANCOVA when building a predictive model of angler catch rate from walleye density to determine if walleye density differed among length-limit regulation categories and years sampled. When length category and years sampled were not significant, walleye densities were grouped together regardless of length regulation and year (Beard et al. 1997).

3.3.4 Random Effects and Mixed Models

In a randomized block design, the entire population of interest is broken into subgroups (blocks) from which units are selected for treatment. In Box 3.10, for example, all lakes in Wisconsin were designated as coming from the northern or southern part of the state. Thus, a randomized block design is analogous to a stratified random sampling design. In many fisheries investigations, blocking is used to reduce variability, but samples are not collected from all blocks within the population. For example, in an experiment to evaluate the effects of a herbicide application on density of age-0 bluegill, we might randomly select five lakes to receive a herbicide treatment and five lakes to receive no treatment. This could be repeated for 4 years, resulting in 20 treatment lakes and 20 control lakes (Box 3.12). We suspect that recruitment of age-0 bluegill may vary annually because of factors such as weather. Thus, we could use year as a blocking factor. However, our interest lies not just in the years selected for study but also in future years. Thus, the blocks (i.e., year) constitute only a sample of all possible years of interest. Such factors are analogous to clusters in cluster sampling (Figure 3.2). When this is the case, the blocking factor is appropriately treated as a random effect in the statistical model. A model that includes both random effects (years, in this example) and fixed treatments (also known as fixed effects; herbicide treatments, in this example) is referred to as a mixed model.

The statistical model for simple mixed models (i.e., with only a single fixed and a single random effect) is similar to that for the randomized block:

$$y_{ijk} = \mu + \alpha_i + \beta_k + \varepsilon_{ijk}, \tag{3.18}$$

Box 3.11 Example of an Analysis of Covariance Design

The goal of this study was to determine how substrate size affected early growth of brook trout eggs. In a lab experiment, a fisheries scientist placed individual brook trout eggs into containers with different substrates. The investigator also believed that egg diameter would affect early growth, so egg size was measured as a continuous covariate. An analysis of covariance model with egg diameter as the continuous variable and substrate as the categorical treatment variable follows.

Program

```
   *The following data step creates a data set named growth containing the
data that follow;
   data growth;
   input id substrate $ egg_diameter growth;
   cards;
    1          Cobble            8.3           20.0
    2          Cobble            8.5           23.5
    3          Cobble           11.2           24.7
    4          Cobble           10.7           29.5
    5          Cobble            9.6           24.3
    6          Cobble           11.8           31.7
    7          Cobble            9.6           22.1
    8          Cobble            8.9           19.0
    9          Cobble           11.2           17.3
   10          Cobble            8.9           23.3
    1          Gravel           10.3           36.4
    2          Gravel            9.5           25.7
    3          Gravel            8.5           13.6
    4          Gravel            9.9           33.9
    5          Gravel            8.6           17.1
    6          Gravel            8.9           22.6
    7          Gravel           10.4           32.0
    8          Gravel           10.8           40.2
    9          Gravel            9.9           26.6
   10          Gravel           10.1           32.9
    1          Sand              9.3           20.4
    2          Sand              8.8           15.3
    3          Sand              9.2           21.6
    4          Sand             10.0           22.9
    5          Sand             10.5           21.2
    6          Sand             10.2           17.4
    7          Sand              9.4           12.4
    9          Sand             10.7           21.8
   10          Sand             11.8           25.0
   ;
   run;
   *These statements call the GLM procedure in SAS, declaring the variable
substrate to be a categorical predictor variable and growth to be the
continuous response variable. By default, the variable egg_diameter is
treated as a continuous predictor variable;
   proc glm;
   class substrate;
   model growth=substrate egg_diameter egg_diameter*substrate;
   run;
```

(Box continues)

Box 3.11 *(continued)*

Results and Interpretation

The results of the analysis follow.

Table The GLM procedure for the dependent variable growth of brook trout. The variable substrate is a categorical predictor variable and the variable egg diameter is treated as a continuous predictor variable.

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	5	1025.104031	205.020806	17.64	<0.0001
Error	23	267.387693	11.625552		
Corrected total	28	1292.491724			
R^2	0.793122	Root MSE	3.409626		
CV	14.23951	Growth mean	23.94483		

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Substrate	2	266.1171709	133.0585855	11.45	0.0004
Egg_diameter	1	557.0652164	557.0652164	47.92	<0.0001
Egg_diameter*substrate	2	311.4589773	155.7294887	13.40	0.0001

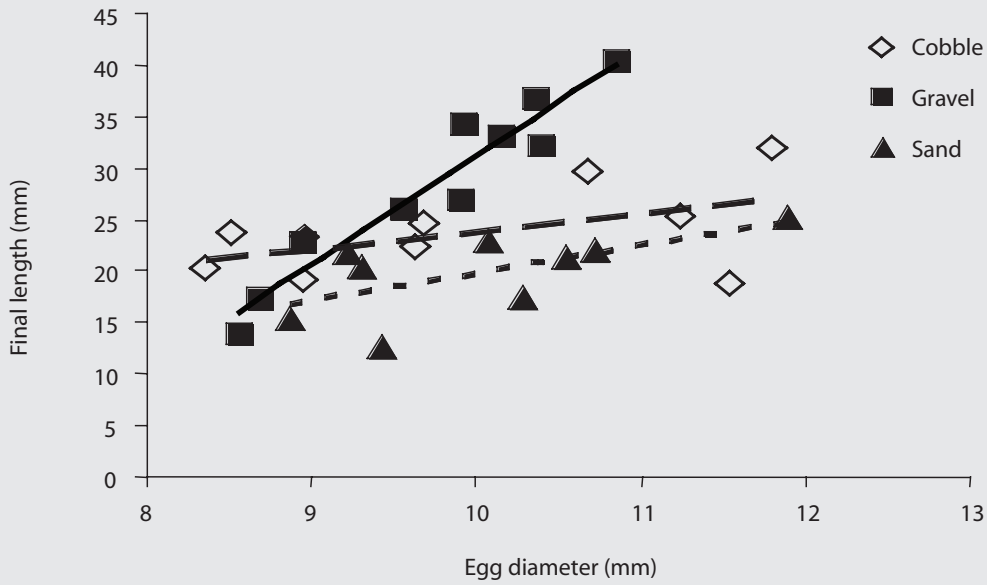
Results of the analysis indicate a significant interaction between egg diameter and substrate (F = 13.40; P = 0.0001), which indicates that egg diameter did not influence final length the same for all substrate classes. Therefore, significance of the two main effects, substrate (F = 11.45; P = 0.0004) and egg diameter (F = 47.92; P < 0.0001), cannot be interpreted because the ANCOVA model (equation [3.17]) does not apply. If the interaction is significant, separate regression models should be fit to each treatment level and then compared.

where y_{ijk} are the observations, μ is the population mean of the response variable, α_i is the treatment effect for each level, β_k is the effect for each level of the random effect variable, and ε_{ijk} is the unexplained variation among experimental units. This model can then be use to test the H_0 that all treatment level means (specified as α_i) are the same, taking into account the effect of the random variable(s):

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu.$$

Mixed models have been used occasionally in fisheries investigations but have also not been used when they would be appropriate. The most common mistake is to treat a random factor as a fixed effect, with the consequence that type I errors are underestimated. Buynak and Mitchell (2002) provide an example where a mixed model was applied in a fishery experiment. The study was designed

With a significant interaction between a continuous and categorical variable, the best way to interpret the results is graphically. For these data, growth generally increased with egg diameter, but the increase was higher in gravel substrate (steeper slope) than it was in cobble or sand substrate (shallower slope). In cobble and sand substrates, the relationship between egg diameter and growth was consistent (similar slopes). In addition, growth was higher in gravel than in sand or cobble for egg diameters greater than about 9.5 mm but lower in gravel than in sand or cobble for smaller egg diameters.



to determine the effects of a slot size limit on smallmouth bass populations. In this study, Buynak and Mitchell (2002) set up a mixed-effects model that tested for differences in density between length limit treatment sites (slot size limit versus no slot size limit) and across years. Year was considered a random variable in this model, because Buynak and Mitchell were interested in determining if the effect of years was the same for all years or differed among years (Buynak and Mitchell 2002).

3.3.5 Factorial Design

The factorial design is used when an investigator wants to investigate the effects of more than one factor on the response variable. In a factorial design, each complete trial of the experiment explores all possible combinations of the levels of

Box 3.12 Example of a Mixed-Model Design

The goal of this study was to determine the effect of herbicide treatment on the abundance of age-0 bluegill in lakes. In theory, treatment with herbicide will create greater access to food resources, so abundance of age-0 bluegill should increase. Funds were available for treating and sampling only four lakes each year, along with sampling an equivalent number of untreated control lakes. To increase the sample size available for the experiment, the fisheries scientists treated lakes over 4 years but were concerned that year-to-year variation in weather could obscure the real effect of treatment.

Program

```
   *This data step creates a data set named herb that contains the following
data;
   data herb;
   input year herbicide $ 13-22  lake_id bluegill_yoy;
   cards;
    2001        Treatment        988          86
    2001        Treatment        116         100
    2001        Treatment        375         163
    2001        Treatment         17         135
    2001        Control          592          62
    2001        Control          677          69
    2001        Control          850          56
    2001        Control          566          50
    2002        Treatment        814         172
    2002        Treatment        397         200
    2002        Treatment        175         204
    2002        Treatment        867         153
    2002        Control          557          51
    2002        Control          106         122
    2002        Control          770          42
    2002        Control          111         127
    2003        Treatment        291         117
    2003        Treatment         76         125
    2003        Treatment         35         153
    2003        Treatment        997         123
    2003        Control          385          89
    2003        Control          712         106
    2003        Control          551          34
    2003        Control          567         197
    2004        Treatment        532          83
    2004        Treatment        424          65
    2004        Treatment        908          59
    2004        Treatment        369          69
    2004        Control          192         137
    2004        Control          371          66
    2004        Control          623          28
    2004        Control          515          23
   ;
   run;
```

```
  *These statements call the MIXED procedure in SAS, declaring herbicide and
year to be categorical predictor variables, and bluegill_yoy (age-0) to be a
continuous response variable. The model statement indicates that the Kenward-
Roger method should be used for computing the degrees of freedom. The random
statement identifies year as a random effect, and the lsmeans statement
requests least-squares means for bluegill density for the different levels of
herbicide treatment;
  proc mixed  covtest;
  class herbicide year;
  model bluegill_yoy = herbicide / ddfm=kenwardroger;
  random year/solution;
  lsmeans herbicide;
  run;
```

Results and Interpretation

Results of this analysis are as follow.

Table The mixed procedure of SAS. Herbicide treatment (fixed effect) and year (random effect) are predictor variables, and age-0 bluegill density is the continuous response variable. The convergence criteria were met. Abbreviations are given for –2 · residual log likelihood (–2Res log like); Akaike's Information Criteria (AIC); small sample corrected AIC (AICc); and Bayesian Information Criteria (BIC). Note a smaller value is better for the information criteria indices.

Iteration History

Iteration	Evaluations	–2Res log like	Criterion
0	1	321.73349015	
1	1	317.52062502	0.00000000

Covariance Parameter Estimates

Covariance parameter	Estimate	SE	*Z*-value	*P* > *Z*
Year	689.45	734.59	0.94	0.1740
Residual	1660.62	451.96	3.67	0.0001

Fit Statistics

–2Res log like	317.5
AIC	321.5
AICc	322.0
BIC	320.3

(Box continues)

Box 3.12 *(continued)*

Solution for Random Effects

Effect	Year	SE estimate	Prediction	*df*	*t*-value	$P > \|t\|$
Year	2001	–9.1751	18.5847	3.77	–0.49	0.6489
Year	2002	24.4509	18.5847	3.77	1.32	0.2626
Year	2003	12.2495	18.5847	3.77	0.66	0.5479
Year	2004	–27.5253	18.5847	3.77	–1.48	0.2169

Type 3 Tests of Fixed Effects

Effect	Numerator *df*	Denominator *df*	*F*-value	$P > F$
Herbicide	1	27	10.53	0.0031

Least-Squares Means

Effect	Herbicide	Estimate	SE	*df*	*t*-value	$P > \|t\|$
Herbicide	Control	78.6875	16.6178	4.52	4.74	0.0067
Herbicide	Treatment	125.44	16.6178	4.52	7.55	0.0010

Results of the analysis suggest that application of herbicide significantly increased the relative abundance of age-0 bluegill ($F = 10.53$; $P = 0.0031$). In control lakes, the mean catch of age-0 bluegill was 78.7 with a SE of 16.6, whereas in treated lakes the mean catch of age-0 bluegill was 125.4 with a SE of 16.6. The effect of the random year effect was not so clear because the covariance estimate for the year effect was 689.45, but the covariance had a SE of 734.59 and a *P*-value of 0.1740. Although this *P*-value is greater than the often-used 0.05, accounting for the potential effects of years is likely an important structural component of the design, and therefore, year should still be included in the model.

factors investigated (Montgomery 2001). For example, an experiment with a levels of factor A and b levels of factor B (where A and B are main effects) includes $a \times b$ treatment combinations. In factorial designs, main effects are generally of primary interest, and if no interactions are present between or among main effects, main effects are simple averages of the effects found for each treatment level (Cochran and Cox 1957; Montgomery 2001). In factorial designs, the factors are considered to be fixed effects (Quinn and Keough 2002).

As an example of a factorial design, the fisheries scientist of an aquaculture facility may be interested in exploring how stocking density and different feeding levels affect the yield of channel catfish in rearing ponds. The fisheries scientist could use only the lowest and highest stocking densities and three feeding levels for the fish, for six possible treatment combinations (Figure 3.5). The fisheries scientist randomly assigns ponds to each treatment combination and runs the experiment. If stocking levels and feeding levels do not interact, the interpretation of

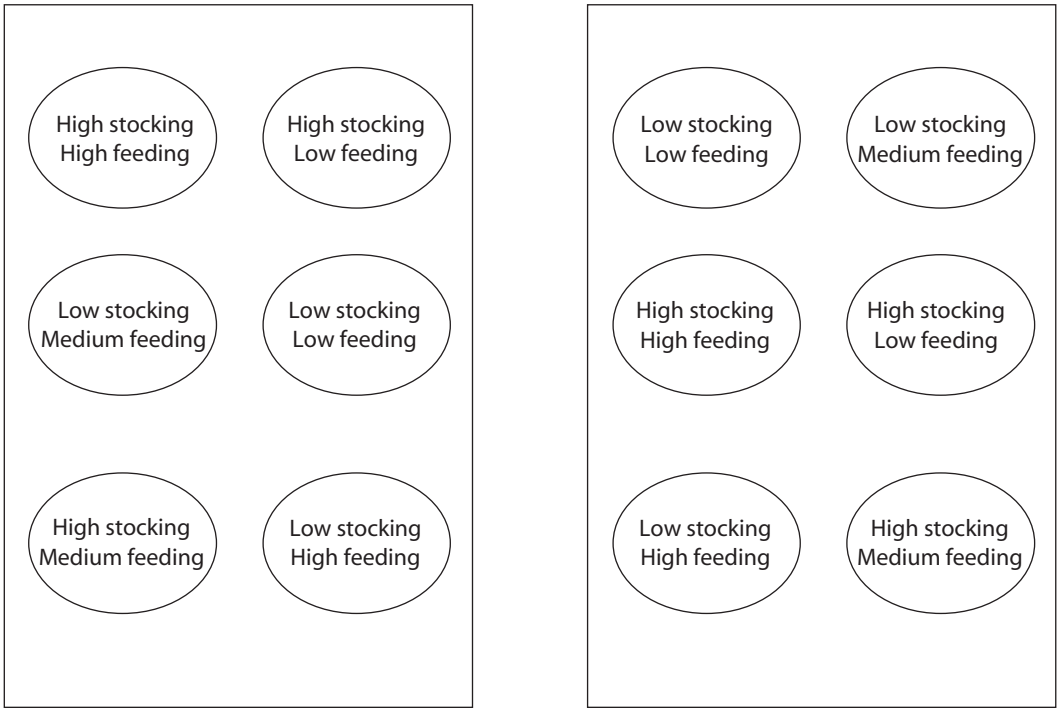


Figure 3.5 Example of a fully randomized factorial design for testing the effect of two stocking densities and three feeding rates. The entire experiment has two replicates for each combination of the factors.

the main effects is straightforward. For instance, the main effect of feeding level, given stocking densities, would be simply the difference between the averages of the results from high and low feeding levels, regardless of the stocking density. That is, we calculate the average for high feeding levels (across all units) and the average for low feeding levels (across all units) and subtract the mean of the low feeding levels from the high feeding levels. The effect is then interpreted as increasing factor *A* from a low to high feeding level causes an average effect equal to the difference between the means (Montgomery 2001). In the absence of interaction, main effects are additive so they are simple to calculate and interpret (Quinn and Keough 2002).

The factorial design helps to understand if and how the effects of each factor interact (Montgomery 2001). Failure to use a factorial design may lead to misinterpretation of results or failure to ascribe results to proper effects. With the presence of an interaction, the effect of factor *A* depends on the level of factor *B* (Montgomery 2001). For example, in our hatchery experiment, at a low stocking density fish may grow at a similar rate regardless of feeding level, but at a high stocking rate feeding rate may affect growth. Interactions are very common in fisheries science because main effects may have synergistic or antagonistic effects (Quinn and Keough 2002). Interactions can make interpretation of main effects difficult and often are easier to interpret when main effects are plotted.

Factorial designs have many advantages over other designs. Factorial experiments are especially useful when the goal of the experiment is to obtain a broad picture of the effects of the factors (Cochran and Cox 1957). If the factors are independent of one another, the factorial experiment can save considerable time and expense (Cochran and Cox 1957; Montgomery 2001). The factorial experiment is most often used in manipulative experiments and in exploratory work, where the factor effects are explored over a range of values (Cochran and Cox 1957; Quinn and Keough 2002).

The statistical model for a simple factorial design with two factors is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \alpha\tau_{ij} + \varepsilon_{ijk}, \tag{3.19}$$

where y_{ijk} are the observations, μ is the population mean of the response variable, α_i is the treatment effect for each level of the first factor, τ_j is the effect for each level of the second factor, $\alpha\tau_{ij}$ is the interaction between main effects, and ε_{ijk} is the unexplained variation among experimental units. This model can then be use to test the H_0 that the means for each level of each factor are the same and that the interaction between the factors is 0.

$$\begin{aligned} H_0 : \alpha_1 &= \alpha_2 = \ldots = \alpha_i; \\ \tau_1 &= \tau_2 = \ldots = \tau_j; \text{ and} \\ \alpha_1\tau_1 &= \alpha_1\tau_2 = \alpha_2\tau_1 = \alpha_2\tau_2 = \ldots = \alpha_i\tau_j. \end{aligned}$$

Factorial designs are commonly used in studies of fisheries management and ecology (Box 3.13). For example, Nowlin and Drenner (2000) used mesocosms to examine the effects of the presence or absence of a planktivore in conjunction with the presence or absence of a fish assemblage on zooplankton densities. Similarly, Dahl (1998) used a factorial design to evaluate the effects of benthivory on benthic assemblages by enclosing standard lengths of stream and then examining the invertebrate assemblage in streams sections with no fish, bullheads, brown trout, and brown trout plus bullheads. In a more complex design, Drenner et al. (1998) examined the effects of nutrient loading, levels of omnivory, and levels of clay on phytoplankton biomass present in mesocosms. In all of these experiments, interactions between factors were suspected, so factorial designs were necessary to understand the effects.

Factorial designs are also beneficial in other types of experiments. For example, Aas et al. (2000) used a factorial design on results of a mail survey to produce hypothetical profiles of fishing opportunities that were based on fishing regulations and expectations of anglers who fished certain waters. Factorial experiments can also be used in computer modeling. For example, Sampson and Yin (1998) used computer simulations of a fractioned factorial design to examine the effects of natural mortality, fishing mortality, and recruitment on the demographic history of a fishery. Factorial designs are common in fisheries, although investigators may not refer to their designs as factorial. If multiple

Box 3.13 Example of a Factorial Design

The goal of this study was to determine how size and stocking location of fingerling Chinook salmon affected survival and subsequent return to the Snake River. Bugert and Mendel (1997) used a 2 × 2 factorial design in which size (subyearling versus yearling) and location of release (on-station versus off-station) were compared to see how these factors affected survival. For this example, we have included only years when all treatment combinations were implemented.

Program

```
data chinook;
input year size$ release$ survival;
cards;
1987 Sub          On    .058
1987 Sub         Off    .155
1987 Yearling     On    .406
1987 Yearling    Off    .319
1988 Sub          On    .058
1988 Sub         Off    .004
1988 Yearling     On    .350
1988 Yearling    Off   1.376
1989 Sub          On    .014
1989 Sub         Off    .008
1989 Yearling     On    .092
1989 Yearling    Off    .320
1990 Sub          On    .047
1990 Sub         Off    .044
1990 Yearling     On    .599
1990 Yearling    Off   3.048
;
run;
```

Because survival was expressed as a percentage, the data were first transformed using the arcsine transformation. The program used to analyze these data follows.

```
data chinook1;
set chinook;
arcsurv=arsin(survival/100);
run;

proc glm;
class size release;
model arcsurv=size release size*release;
lsmeans size release size*release/stderr;
run;
```

Results and Interpretation

Results of this analysis are as follow.

(Box continues)

Box 3.13 *(continued)*

Table The GLM procedure for a 2 × 2 factorial design to assess fingerling Chinook salmon survival with size (subyearling versus yearling) and location of release (on-station versus off-station) as factors (based on Bugert and Mendel 1997). The dependent variable is the arcsin transformation of the percent survival (arcsurv); the number of observations is 16.

Class Level Information

Class	Levels	Values
Size	2	Sub Yearling
Release	2	Off On

General Linear Model

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	3	0.00039779	0.00013260	3.10	0.0672
Error	12	0.00051284	0.00004274		
Corrected total	15	0.00091063			

R^2	0.436827	Root MSE	0.006537
CV	151.6227	Arcsurv mean	0.004312

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Size	1	0.00023428	0.00023428	5.48	0.0373
Release	1	0.00008329	0.00008329	1.95	0.1880
Size*release	1	0.00008021	0.00008021	1.88	0.1958

Least-Squares Means

Size	Arcsurv LSMEAN	SE	*P* > |*t*|
Sub	0.00048500	0.00231130	0.8373
Yearling	0.00813815	0.00231130	0.0042

Release	Arcsurv LSMEAN	SE	*P* > |*t*|
Off	0.00659315	0.00231130	0.0146
On	0.00203001	0.00231130	0.3970

Size	Release	Arcsurv LSMEAN	SE	*P* > |*t*|
Sub	Off	0.00052750	0.00326866	0.8745
Sub	On	0.00044250	0.00326866	0.8946
Yearling	Off	0.01265879	0.00326866	0.0022
Yearling	On	0.00361751	0.00326866	0.2901

Results of the analysis suggest that size at stocking significantly affected survival of juvenile Chinook salmon ($F = 5.48$; $P = 0.0373$), but that release location did not significantly affect survival ($F = 1.95$; $P = 0.1880$). Further, the interaction between release location and size at stocking was not significant ($F = 1.88$; $P = 0.1958$).

factors and interactions are included in the study design, a factorial design is very likely the basis for the experiment.

3.3.6 Nested Design

Nested designs often occur when subsamples are taken from the experimental units included in a study. A common type of nesting in fisheries research is to have individual sample sites nested within lakes or streams. In such studies, individual lakes or streams are experimental units, but we need to account for variation among sites within each lake or stream. Similarly, individual fish sampled from a lake are generally not true replicates but should be treated as a nested subsample. For example, in Figure 3.6, fish are nested subsamples within lakes,

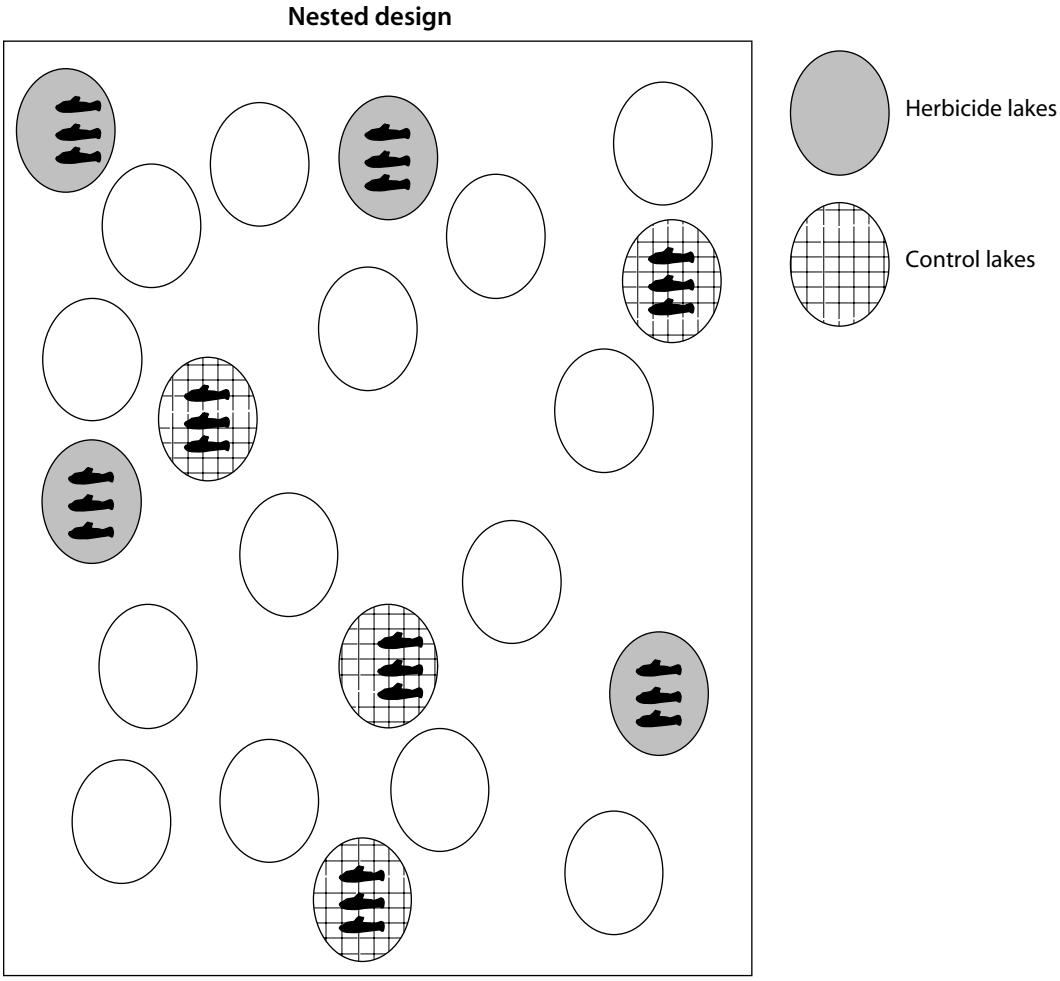


Figure 3.6 Example of a nested design in which lakes are the main experimental unit and individual fish are nested subsamples within each lake. Open circles indicate lakes that were not included in the study as either herbicide lakes or control lakes.

which were randomly selected for application of herbicide or held as control lakes. To account for the experimental design properly, a nested effect needs to be included in the statistical model. The nested effect variable is generally a categorical identifier for each experimental unit. In nested designs, the effects can be either random or fixed, but in the biological sciences the main effect (e.g., the treatment applied) is often fixed and the nested effect (e.g., individual lake identifier) is often random (Quinn and Keough 2002).

The experimental units can sometimes be difficult to identify properly in nested designs (Yandell 1997). For example, in a design in which sampling transects are nested within habitat types in a lake, the primary experimental units are habitat patches, not sampling transects or the lake. Difficulties in identifying primary experimental units in nested designs can also lead to pseudoreplication (Hurlbert 1984), where subsamples (e.g., fish in nets) are confused with truly replicated experimental units (e.g., nets in lakes).

The statistical model for nested designs is similar to that for mixed models (section 3.3.4); an example of a nested design is given in Box 3.14. Nested designs are common in fisheries. The design discussed above, with sampling transects or locations nested in streams or lakes, is appropriate when comparing effects across water bodies. For example, to determine how various benthic taxa varied at different spatial scales, Boyero and Bailey (2001) used a nested design with sampling points nested within riffles nested within streams. Boyero and Bailey (2001) were able to attribute the variation in taxa to these different spatial scales. Using a similar approach, Cole (2001) nested sample cells of different sizes to assess spatial variability in the abundance of clams. Pierce et al. (2001) used a nested design to examine differences in species richness in relation to diel sampling period, sampling gears, and sites, all nested within each lake sampled. Using a random-effects model, Radomski and Goeman (2001) nested developed and undeveloped lakeshore plots within lake development classes to quantify differences in vegetative abundance among lakes and between shoreline types. Toepfer et al. (1999) nested individual leopard darter results within separate trials to separate individual variation in burst speed and numbers when assessing overall swimming performance. Conover et al. (1997) used a nested design to attribute variance in growth rate of young striped bass to individual mothers, nested within the latitude from which they came, to separate genetic and physiological effects of each mother from the effect of latitude.

3.3.7 Split-Plot Design

In a split-plot design, the main experimental units are divided into two or more parts (Cochran and Cox 1957). Different levels of treatments are then applied to a subunit within the main experimental unit. This type of design is similar to a randomized block design, except in the randomized block, the treatment combinations are assigned randomly, not randomly within each main plot (Cochran and Cox 1957). For example, consider a hatchery experiment with two levels of stocking density (high and low) and two feeding levels (high and low). In a split-plot

Box 3.14 Example of a Nested Design

For the example in Box 3.12, where the effect of herbicide treatment on age-0 bluegill density was investigated, we may also be interested in how herbicide treatment affects mean length of age-0 bluegill at the end of the growing season (for this example, assume that length of individual bluegill from each lake in the study was measured). In a nested design, the primary experimental unit is a lake, so each bluegill is not an independent replicate but rather is a subsample from the lake. For brevity, only the lakes sampled in 2001 from Box 3.12 are used in this example.

Table Hypothetical data on lengths of age-0 bluegills from lakes treated with herbicide and control lakes that were not treated with herbicide (an extension of Box 3.12 data).

Summary statistic	Length of age-0 bluegills							
	Treatment lakes				Control lakes			
	988	116	375	17	592	677	850	566
	103	88	97	116	70	83	102	79
	90	95	94	94	79	85	89	72
	98	82	103	112	78	92	82	67
	90	100	94	111	85	85	86	78
	96	84	83	96	65	84	99	83
	88	92	93	111	68	83	88	68
	97	94	90	91	93	79	99	87
	100	79	107	116	80	90	80	75
	89	103	94	109	89	77	81	79
	108	81	86	110	65	85	93	90
Mean	95.9	89.8	94.1	106.6	77.2	84.3	89.9	77.8

Program

The SAS program used to analyze these data follows.

```
  data bluegill;
  input herbicide $ 1-9 lake length;
  cards;
  Treatment        988     103
  Treatment        988      90
  (input data)
  ;
  run;

  * This call to the MIXED procedure is much like in Box 3.12, except that
the random statement is used to indicate that individual bluegills within a
lake are a subsample from a herbicide treatment class;
  proc mixed covtest;
  class lake herbicide;
  model length=herbicide;
  random lake(herbicide);
  lsmeans herbicide;
  run;
```

(Box continues)

Box 3.14 *(continued)*

Results and Interpretation

Results of the analysis follow.

Table The effect of herbicide treatment on age-0 bluegill length. The mixed procedure with the random statement is used to indicate that individual bluegills within a lake are a subsample from an herbicide treatment class. Convergence criteria were met. The estimation method was restricted maximum likelihood (REML).

Model Information	
Data set	WORK.BLUEGILL
Dependent variable	Length
Covariance structure	Variance components
Estimation method	REML
Residual variance method	Profile
Fixed effects SE method	Model-based
Degrees of freedom method	Containment

Class Level Information

Class	Levels	Values
Lake	8	17 116 375 566 592 677 850 988
Herbicide	2	Control treatment

Dimensions

Covariance parameters	2
Columns in X	3
Columns in Z	8
Subjects	1
Maximum observations per subject	80
Observations used	80
Observations not used	0
Total observations	80

experiment, the investigator would randomly select a stocking density for each of four ponds, divide the ponds in half with barriers, and randomly select a feeding rate to apply to each half of each pond (Figure 3.7). The sample size for feeding rate increased from two, using a factorial design with four ponds, to four, using the split-plot design.

When performing a split-plot experiment, the B effect and $A \times B$ interaction (the feeding rate and feeding rate × stocking density effects in the hatchery experiment) are estimated more precisely than are the A effects (stocking density; Cochran and Cox 1957). As described with the hatchery experiment, the degrees of freedom are smaller for the whole unit than for the subunit comparisons. The

Iteration History

Iteration	Evaluations	–2Res log like	Criterion
0	1	579.91706840	
1	1	561.75492286	0.00000000

Covariance Parameter Estimates

Covariance parameter	Estimate	SE	*Z*-value	*P* > *Z*
Lake (herbicide)	37.3500	25.1356	1.49	0.0686
Residual	61.5000	10.2500	6.00	<0.0001

Fit Statistics

–2Res log like	561.8
AIC	565.8
AICc	565.9
BIC	565.9

Type 3 Tests of Fixed Effects

Effect	Numerator *df*	Denominator *df*	*F*-value	*P* > *F*
Herbicide	1	6	9.40	0.0220

Least-Squares Means

Effect	Herbicide	Estimate	SE	*df*	*t*-value	*P* > |*t*|
Herbicide	Control	82.3000	3.2977	6	24.96	<0.0001
Herbicide	Treatment	96.6000	3.2977	6	29.29	<0.0001

Results of the analysis indicate that age-0 bluegill in control lakes were significantly shorter (82.3 mm) than in herbicide-treated lakes (96.6 mm; $F = 9.40$; $P = 0.0220$).

primary advantage of the split-plot design is realized when the *B* and *A* × *B* effects are of greater interest than is the *A* effect, or when the *A* effect cannot be tested on small experimental units because the cost or size of the *A* experimental unit is prohibitive (Cochran and Cox 1957; Montgomery 2001). For example, in the hatchery example the primary interest was in the feeding rate and feeding rate × stocking density effects, whereas the secondary interest was in the stocking density effect. However, the increase in precision of estimating *B* effects can often lead to results where the effect of factor *B* is significant and the effect of factor *A* is not significant (Cochran and Cox 1957). Analysis of data collected using a split-plot design are often complicated and require detailed coding of data to assure analysis

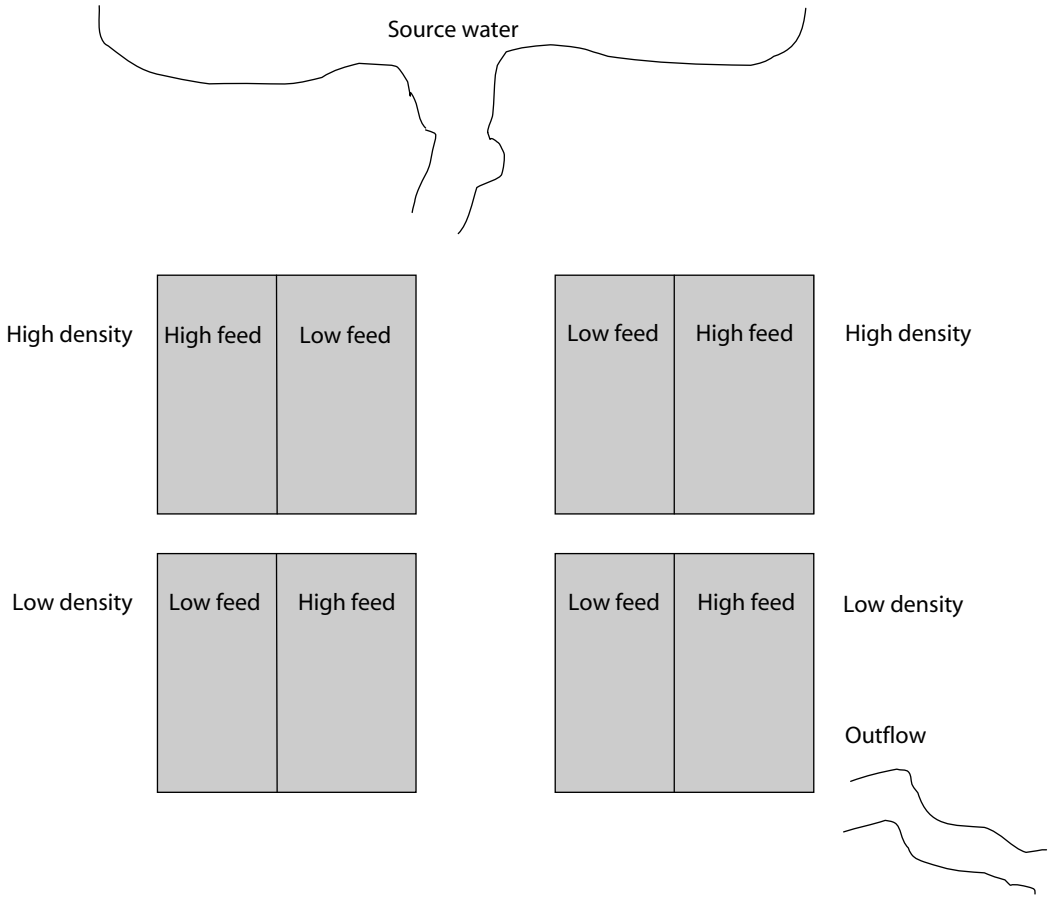


Figure 3.7 Example of a split-plot design to test the effects of feeding and stocking density on growth rates of muskellunge in hatchery ponds. In this experiment, four ponds are each divided in half and two ponds each are randomly selected for high and low stocking density. Within each pond, each side is randomly selected for high or low feeding rates. The overall design includes two replicates for stocking density and four replicates for feeding rate.

programs work correctly. Split-plot designs often contain a mixture of random (e.g., ponds) and fixed effects (e.g., feeding or stocking rates), which may further complicate analysis (Quinn and Keough 2002).

Split-plot designs are uncommon in fisheries, though the repeated-measures split-plot design has been widely applied (Box 3.15; Maceina et al. 1994). More details of that design will be covered in the next section because it combines the aspects of the split-plot and repeated-measures design. An example of a split-plot design by Secor et al. (2000) tested differences in growth performance between anadromous and nonanadromous strains of striped bass; a split-plot design was used to separate growth and salinity effects at three levels of growth and salinity.

3.3.8 Repeated-Measures Design

The repeated-measures design generally refers to experiments in which individual experimental units are observed more than once (Quinn and Keough 2002). In some cases, a single treatment is applied and the experimental unit is observed over time, but in other cases multiple treatments are applied and the experimental unit is observed multiple times. When the same experimental units are observed multiple times, the observations on the response to a treatment are potentially correlated because the same experimental unit is used (Quinn and Keough 2002). Thus, observations are not necessarily independent, and the design and analysis should take this into account. In such experiments, the treatment is typically considered to be a fixed effect and the subject is often a random effect (Montgomery 2001). As with split-plot designs, repeated-measures designs are often complex, and the analysis depends on the details of the situation (e.g., how many times the units are observed and how observations are correlated in time).

Repeated-measures designs are not commonly used in fisheries. A repeated-measures design was used to estimate the retention rate of coded wire tags in paddlefish, which were marked in four locations with coded wire tags. Each individual fish then was examined monthly to determine if tags had been retained or lost in each location, and total tag retention rate was estimated (Fries 2001). In an experiment to train grass carp to respond to different types of sound, Willis et al. (2002) used sound at different frequencies to determine if response varied with frequency. Because individual response types were measured on individual grass carp, the type of sound needed to be corrected for the measurement from individual grass carp to assure that individual grass carp behavior was taken into account in analysis of response to the type of sound (Willis et al. 2002).

In fisheries, repeated-measures designs often refer to a specialized version of the split-plot design in which the repeated measures are taken from the same set of sites (Box 3.15; Maceina et al. 1994). The sites selected are usually thought of as random effects in such designs, so repeated-measure designs are essentially split-plot designs that allow for correlation within each nested random effect (Yandell 1997). The repeated-measures design or the repeated-measures split-plot design often assigns treatments to experimental units, which are then measured over different time intervals (which become the plots). The main difference between a split-plot design and a repeated-measures split-plot design is that the split-plot design allocates within-plot treatments to subunits within each plot, whereas the repeated-measures split-plot design allocates within-subjects treatments sequentially to each subject (Quinn and Keough 2002). In a fisheries experiment that uses a repeated-measures split-plot design, sampling stations are often fixed, so treatments are measured repeatedly at the same site (correlation is present) with interactions between site, treatment, and time (Maceina et al. 1994).

The use of repeated-measures split-plot designs has become common in fisheries because of interest in time period effects of sampling at fixed sites (Maceina et al. 1994). Maceina et al. (1994) were the first to advocate use of repeated-measures split-plot designs in fisheries. In one experiment, Maceina et al. (1994) quantified

Box 3.15 Example of Repeated-Measures Split-Plot Design

The goal of this study was to determine the effects of vegetation removal by grass carp on fish biomass. Maceina et al. (1994) sampled the same six coves twice before and twice after treatment. Main plot *A* included cove, treatment, and cove*treatment interaction effects, and subplot *B* included time and time*treatment interaction effects. Maceina et al. (1994) popularized the use of repeated-measures split-plot designs in fisheries, which is appropriate for analyzing data collected through time at fixed stations. The analysis relies on standard analysis of variance techniques.

Program

```
data cove;
input year treat$ time cove area biomass;
cards;
1980 PRE  1 1 1.51 13854
1980 PRE  1 2  .67  4091
1980 PRE  1 3 2.19 17195
1980 PRE  1 4  .63  5138
1980 PRE  1 5  .64  5148
1980 PRE  1 6  .45  2971
1981 PRE  2 1 1.60  6374
1981 PRE  2 3 1.97 21441
1981 PRE  2 4  .74 17830
1981 PRE  2 5  .66  3577
1981 PRE  2 6  .32  2678
1985 POST 1 1 1.83 3209
1985 POST 1 3 2.39 11556
1985 POST 1 4  .88  8132
1985 POST 1 5  .70  5094
1985 POST 1 6  .49  1973
1986 POST 2 1 1.83 10643
1986 POST 2 2  .43   479
1986 POST 2 3 2.39 11103
1986 POST 2 4  .88  2852
1986 POST 2 5  .70  2489
1986 POST 2 6  .49  8898
 ;
data cove;
set cove;
logbio=log10(biomass);
run;

proc glm;
class cove treat time;
model logbio=cove treat treat*cove time treat*time;
      test h=treat e=treat*cove;
      test h=cove e=treat*cove;
run;
```

Results and Interpretation

The main fixed effects are cove and treatment, and main plot is split into time effects. The interactions were estimated to see if any spatial (treat*cove) or temporal (treat*time) correlations affected the results.

Table The GLM procedure to determine the effects of vegetation removal by grass carp on fish biomass. Six coves were sampled twice before (PRE) and twice after (POST) treatment. Main plot *A* included cove, treatment, and cove*treatment interaction effects, and subplot *B* included time and time*treatment interaction effects (based on Maceina et al. 1994). The dependent variable is $\log_{10}$biomass of fishes (logbio), and the number of observations was 22.

Class Level Information

Class	Levels	Values
Cove	6	1 2 3 4 5 6
Treatment	2	POST PRE
Time	2	1 2

The GLM Procedure

Source	*df*	SS	Mean square	*F*-value	*P* > *F*
Model	13	2.46565668	0.18966590	2.15	0.1406
Error	8	0.70628807	0.08828601		
Corrected total	21	3.17194474			
R^2	0.777333	Root MSE	0.297130		
CV	7.943952	Logbio mean	3.740325		

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Cove	5	1.76323921	0.35264784	3.99	0.0409
Treat	1	0.36593102	0.36593102	4.14	0.0762
Cove*treat	5	0.43968044	0.08793609	1.00	0.4767
Time	1	0.01186994	0.01186994	0.13	0.7234
Treat*time	1	0.00436853	0.00436853	0.05	0.8295

Tests of Hypotheses with Type III MS for Cove*Treat as Error Term

Source	*df*	Type III SS	Mean square	*F*-value	*P* > *F*
Treat	1	0.36593102	0.36593102	4.16	0.0969
Cove	5	1.76323921	0.35264784	4.01	0.0768

The type III SS, which are properly calculated using the cove*treat interaction MSE, indicate that the main fixed effects of coves ($F = 4.01$; $P = 0.0768$) and vegetation removal by grass carp treatments ($F = 4.16$; $P = 0.0969$) were significant at an alpha of 0.10, which suggests that the treatment affected fish biomass. Time ($F = 0.13$; $P = 0.7234$) and time*treatment ($F = 0.05$; $P = 0.8295$) effects were not important in explaining differences in fish biomass.

the abundance between years of age-0 black crappie and white crappie that were collected in trap nets at fixed stations over 2 d. The repeated-measures split-plot analysis treated stations as replicates, year as the main treatment effect, and day as the plot effect, along with interactions between years and stations and between days and years (Maceina et al. 1994). Using a similar approach, Pierce et al. (2001) determined the effect of the number of marked fish, station, year, and month on capture efficiency of beach seines for various species of fish. In this experiment, the replicate was the station, the fixed effect was the year, and the plot was the month (Pierce et al. 2001). Year and month were significant in explaining capture efficiency (Pierce et al. 2001). Jackson and Hightower (2001) used individual movement data from striped bass (the plot) to separate variance of individual fish from sex and season. They determined how sex and season affected site fidelity of striped bass (Jackson and Hightower 2001). Finally, to quantify spawning substrate preferences of yellow perch in Lake Michigan, Robillard and Marsden (2001) used a repeated-measures split-plot design that treated stations as replicates, year as the main effect, and substrate types as plots.

3.4 CONCLUSION

One point of potential confusion in experimental design is how to distinguish between fixed factors, fixed blocking effects, and random effects. In the hatchery example above, stocking density and feeding rate were factors assigned randomly to individual ponds because we were interested in understanding the response of the system to these factors. Thus, both of these factors are fixed effects. In contrast, the designation of lakes as being northern or southern in Box 3.10 is not something under the control of the investigator; all lakes are assigned to one of these two groupings before the start of the experiment. As such, this is an example of a fixed blocking factor. However, we may be interested in making predictions for lakes in the northern or southern part of the state, even if they were not present in our sample. We can appropriately make predictions for such a case because unsampled lakes must belong to either the northern or southern block. The trickiest situation is for random effects. Imagine, for example, that the state of Wisconsin had been subdivided into 20 different blocks, but we only selected lakes in three of the 20 blocks. We could still use region as a blocking factor, but the three selected blocks do not constitute the entire population (lakes) of 20 blocks, so we could not make predictions for a lake in one of the 17 blocks that were not included in the experiment. Treating the regional designation as a random effect appropriately allows us to take into account block-to-block variation, thereby enabling predictions about lakes in all 20 blocks.

Throughout this discussion, we have focused on true experiments where levels of treatment can be assigned at random by the investigator. However, in many fisheries studies, we are interested in how naturally varying factors affect fish populations, habitat, or anglers. Strictly speaking, such observational studies do not really fall into the category of an experimental design because we cannot infer cause and effect relationships from such studies. Observational

studies are common in fisheries science and yield insight into the dynamics of fishery systems. Many of the methods we have presented in this section are useful for the analysis of observational studies, but we caution the reader to recognize that the conclusions reached from such analyses are akin to correlation and do not imply causality.

As a final comment, the experimental designs presented here represent only a simple subset of the experimental designs used in practice. Elements of several designs are often used to achieve the goals of an experiment. For example, nested designs are frequently used with a factorial design. This occurs because our unit of measurement (e.g., individual fish) is often part of a larger experimental unit (e.g., lake or pond). Elements of repeated-measure designs are also frequently combined with other experimental design components to allow us to determine how experimental units vary over time in response to treatment. Because of the complexity of many experimental designs and analyses in fisheries, we recommend that you consult with a professional statistician before an experiment is started. This will assure that the proper experimental design is used and that the correct analyses techniques are considered and used. The analysis of data from more complex designs needs to be carefully considered but provides much deeper insights into the biology of fisheries systems.

3.5 REFERENCES

Aas, O. E., W. Haider, and L. Hunt. 2000. Angler responses to potential harvest regulations in a Norwegian sport fishery: a conjoint-based choice modeling approach. North American Journal of Fisheries Management 20:940–950.

Bates, D. M., and D. G. Watts. 1988. Nonlinear regression analysis & its applications. Wiley, New York.

Beard, T. D., Jr., S. P. Cox, and S. R. Carpenter. 2003. Impacts of bag limit reductions on angler effort in Wisconsin walleye lakes. North American Journal of Fisheries Management 23:1283–1293.

Beard, T. D., Jr., S. W. Hewett, Q. Yang, R. M. King, and S. J. Gilbert. 1997. Prediction of angler catch rates based on walleye population density. North American Journal of Fisheries Management 17:621–627.

Boyero, L., and R. C. Bailey. 2001. Organization of macroinvertebrate communities at a hierarchy of spatial scales in a tropical stream. Hydrobiologia 464:219–225.

Brown, M. L., and D. J. Austen. 1996. Data management and statistical techniques. Pages 17–62 *in* B. R. Murphy and D. W. Willis, editors. Fisheries techniques, 2nd edition. American Fisheries Society, Bethesda, Maryland.

Bugert, R. M., and G. W. Mendel. 1997. Adult returns of subyearling and yearling fall chinook salmon released from a Snake River hatchery or transported downstream. North American Journal of Fisheries Management 17:638–651.

Buynak, G. L., and B. Mitchell. 2002. Response of smallmouth bass to regulatory and environmental changes in Elkhorn Creek, Kentucky. North American Journal of Fisheries Management 22:500–508.

Caputi, N. 1988. Factors affecting the time series bias in stock-recruitment relationships and the interaction between time series and measurement error bias. Canadian Journal of Fisheries and Aquatic Sciences 45:178–184.

Cochran, W. G. 1977. Sampling techniques, 3rd edition. Wiley, New York.

Cochran, W. G., and G. M. Cox. 1957. Experimental designs. Wiley, New York.

Cole, R. G. 2001. Staggered nested designs to identify hierarchical scales of variability. New Zealand Journal of Marine and Freshwater Research 35:891–896.

Conover, D. O., J. J. Brown, and A. Ehtisham. 1997. Countergradient variation in growth of young striped bass *(Morone saxatilis)* from different latitudes. Canadian Journal of Fisheries and Aquatic Sciences 54:2401–2409.

Dahl, J. 1998. Effects of a benthivorous and a drift-feeding fish on benthic stream assemblage. Oecologia 116:426–432.

Draper, N. R., and H. Smith. 1981. Applied regression analysis, 2nd edition. Wiley, New York.

Drenner, R. W., K. L. Gallo, R. M. Baca, and J. D. Smith. 1998. Synergistic effects of nutrient loading and omnivorous fish on phytoplankton biomass. Canadian Journal of Fisheries and Aquatic Sciences 55:2087–2096.

Efron, B., and R. J. Tibshirami. 1998. An introduction to the bootstrap. CRC Press LLC, Boca Raton, Florida.

Fries, J. N. 2001. Retention of coded wire tags in four locations in juvenile paddlefish. North American Journal of Fisheries Management 21:962–966.

Hartman, K. J., and S. B. Brandt. 1995. Estimating energy density of fish. Transactions of the American Fisheries Society 124:347–355.

Hayes, D. B., J. K. T. Brodziak, and J. B. O'Gorman. 1995. Efficiency and bias of estimators and sampling designs for determining length–weight relationships of fish. Canadian Journal of Fisheries and Aquatic Sciences 52:84–92.

Hilborn, R., and C. J. Walters. 1992. Quantitative fisheries stock assessment: choice, dynamics and uncertainty. Chapman and Hall, New York.

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

Jackson, J. R., and J. E. Hightower. 2001. Reservoir striped bass movements and site fidelity in relation to seasonal patterns in habitat quality. North American Journal of Fisheries Management 21:34–45.

Kocik, J. F., and W. W. Taylor. 1994. Summer survival and growth of brown trout with and without steelhead under equal total salmonine densities in an artificial stream. Transaction of the American Fisheries Society 123:931–938.

Lohr, S. L. 1999. Sampling: design and analysis. Duxbury Press, Pacific Grove, California.

Maceina, M. J., P. W. Bettoli, and D. R. DeVries. 1994. Use of a split-plot ANOVA design for repeated-measures fishery data. Fisheries (Bethesda) 19(3):14–20.

Montgomery, D. C. 2001. Design and analysis of experiments. Wiley, New York.

Myers, R. A., and N. J. Barrowman. 1996. Is fish recruitment related to spawner abundance? U.S. National Marine Fisheries Service Fishery Bulletin 94:707–724.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill/Irwin, Homewood, Illinois.

Nowlin, W. H., and R. W. Drenner. 2000. Context-dependent effects of bluegills in experimental mesocosm communities. Oecologia 122:421–426.

Pierce, C. L., A. M. Corcoran, A. N. Gronbach, S. Hsia, B. J. Mullarkey, and A. J. Schwartzoff. 2001. Influence of diel period on electrofishing and beach seining assessments of littoral fish assemblages. North American Journal of Fisheries Management 21:918–926.

Quinn, G. P., and M. J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK.

Radomski, P., and T. J. Goeman. 2001. Consequences of human lakeshore development on emergent and floating-leaf vegetation abundance. North American Journal of Fisheries Management 21:46–61.

Robillard, S. R., and J. E. Marsden. 2001. Spawning substrate preferences of yellow perch along a sand–cobble shoreline in southwestern Lake Michigan. North American Journal of Fisheries Management 21:208–215.

SAS (Statistical Analysis System). 2003. SAS/STAT user's guide, version 9.1. SAS Institute, Cary, North Carolina.

Sammons, S. M., and P. W. Bettoli. 1999. Spatial and temporal variation in electrofishing catch rates of three species of black bass (*Micropterus* spp.) from Normandy Reservoir, Tennessee. North American Journal of Fisheries Management 19:454–461.

Sampson, D. B., and Y. Yin. 1998. A Monte Carlo evaluation of the stock synthesis assessment program. Pages 315–338 *in* F. Funk and seven coeditors. Fishery stock assessment models. American Fisheries Society, Bethesda, Maryland.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biometrics 2:110–114.

Seber, G. A. F., and C. J. Wild. 1989. Nonlinear regression. Wiley, New York.

Secor, D. H., T. E. Gunderson, and K. Karlsson. 2000. Effect of temperature and salinity on growth performance in anadromous (Chesapeake Bay) and nonanadromous (Santee–Cooper) strains of striped bass *Morone saxatilis*. Copeia 2000:291–296.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry, 3rd edition. Freeman, New York.

Thompson, S. K. 1992. Sampling. Wiley, New York.

Toepfer, C. S., W. L. Fisher, and J. A. Haubelt. 1999. Swimming performance of the threatened leopard darter in relation to road culverts. Transactions of the American Fisheries Society 128:155–161.

Walsh, M. G., D. B. Fenner, and D. L. Winkelman. 2002. Comparison of an electric seine and prepositioned area electrofishers for sampling stream fish communities. North American Journal of Fisheries Management 22:77–85.

Walters, C. J. 1985. Bias in estimation of functional relationships from time series data. Canadian Journal of Fisheries and Aquatic Sciences 42:147–149.

Whalen, K. G., and G. W. LaBar. 1994. Survival and growth of Atlantic salmon *(Salmo salar)* fry stocked at varying densities in the White River, Vermont. Canadian Journal of Fisheries and Aquatic Sciences 51:2164–2169.

Wilde, G. R., and W. L. Fisher. 1996. Reservoir fisheries sampling and experimental design. Pages 397–407 *in* L. E. Miranda and D. R. DeVries, editors. Multidimensional approaches to reservoir fisheries management. American Fisheries Society, Symposium 16, Bethesda, Maryland.

Wilderbuer, T. K., R. F. Kappenman, and D. R. Gunderson. 1998. Analysis of fishing power correction factor estimates from a trawl comparison experiment. North American Journal of Fisheries Management 18:11–18.

Willis, D. J., M. V. Hoyer, D. E. Canfield, Jr., and W. J. Lindberg. 2002. Training grass carp to respond to sound for potential lake management uses. North American Journal of Fisheries Management 22:208–212.

Yandell, B. S. 1997. Practical data analysis for designed experiments. Chapman and Hall, New York.

Zar, J. H. 1999. Biostatistical analysis, 4th edition. Prentice Hall, Upper Saddle River, New Jersey.