CHAPTER 2
STATISTICAL CONCEPTS

Capture-recapture and removal studies often are not recognized as sampling methods because they are quite unlike those used in much of the usual sampling theory (see standard texts such as *Cochran 1977).* For example, in capture-recapture and removal studies the sampling probabilities are not known and cannot be pre-established. No sampling "frame" is available, and the investigator has relatively little control over the situation.

Progress in science usually is made through experimentation—data are collected and analyzed, and conclusions are drawn. The conclusions drawn from the sample data are meant to go beyond the particular study. Biologists often wish to generalize from a particular experiment to the class of all similar experiments. This type of generalization is termed "inductive inference."

Let us examine two idealized case studies to illustrate some elementary concepts of sampling and inductive inference. First, consider the task of a quality control specialist who must estimate the proportion of newly manufactured light bulbs that are defective. Because 5 million bulbs are manufactured each month, he cannot test each light bulb (make a complete census); therefore, he must sample the bulbs and test the sample for defects. If he randomly samples 1000 bulbs from the population of 5 million, tests them, and finds only 3 defective, he might conclude that about 0.3% of the bulbs manufactured during his experiment are defective. That is, he makes an inference about the population from a sample.

Second, consider a biologist faced with estimating the number of mice on a large tract of land in south-central Wyoming in June of a given year. Because total enumeration (a census) is impossible, he might sample the area by establishing several 20 by 20 trapping grids located randomly throughout the area. (He may, in fact, want to stratify the sample by vegetative type, but he will avoid such considerations for the moment.) If he performs a capture study for six nights at each area, he can estimate the density (the number per unit area) for each of the study areas. The density estimates could be averaged over the areas, and inferences could be made about the density of the population, based on the sample data collected from the grids.

Both of these samples involve sampling a defined population, acquisition of data from the sampling process, and finally estimation and conclusions about the population rather than conclusions about only the sample. A theorem of logic tells us that there is uncertainty in inductive inference and, therefore, that we cannot make perfectly certain generalizations about a population by studying only a sample. However, we can make uncertain inferences, and we can measure the degree of uncertainty if the experiment has been performed in accordance with certain scientific principles *(Mood et al. 1974).* One function of the science of statistics is to provide techniques for making inductive inferences and for measuring their degree of uncertainty *(Ostle 1963:1-16).*

With respect to the subject of statistics, many think of statistics in terms of simple t and chi-square tests, analysis of variance, regression, and other such methods. In fact, the field is far broader than is suggested by the methods for data analysis to which people are exposed in the first two or three courses on statistics. Statistics is not a branch of mathematics, but it is an area of science concerned with the development of a practical theory of information. It involves sampling, design of experiments, analysis of information, estimation of parameters, and testing of hypotheses. It is the basis for inductive inference, and it is an integral part of what is termed the Scientific Method. The following sections introduce basic statistical concepts that are needed for an understanding of the following chapters.

Theory, Reality, and Models

It is essential to understand the difference between theory or theoretical statistical models and reality. The methods presented here and in *Otis et al. (1978)* are approximations or models of reality. No model gives an exact explanation of a real biological or physical phenomenon. A "good" model, however, can be very useful to our understanding of a process.

In this primer we are concerned with the statistical theory of animal trapping experiments designed to enable the estimation of population size or density, or both. We postulate theoretical probability models for sampling animal populations, apply a theory of probability and inference based on rigorous mathematical foundations (see *Otis et al. 1978*), and present several theoretical models for use in the collection and analysis of information in real biological populations. The models are not exact representations of nature. Their utility is measured by the extent to which they assist us in understanding the dynamics of animal populations.

For our purposes we think of a model as a mathematical representation of a postulated set of assumptions concerning a capture-recapture or removal experiment. Such models are stochastic because they allow for the fact that the data arise from a random process. In a stochastic process, the outcome (data) is not completely predictable. Stochastic processes are common in everyday life, and they represent the rule rather than the exception. (This topic will be discussed further later in this chapter.)

Although the biologist need not understand the details or derivation of stochastic models, he should be able to see how these models help to achieve the goal of estimating population parameters. Their role is illustrated in Fig. 2.1. The model is the link between the data and the procedure used to estimate the population parameters contained in the model. Thus, whether or not the fact is stated explicitly, all statistical estimation procedures are based on a model of the sampling experiment or, stated differently,

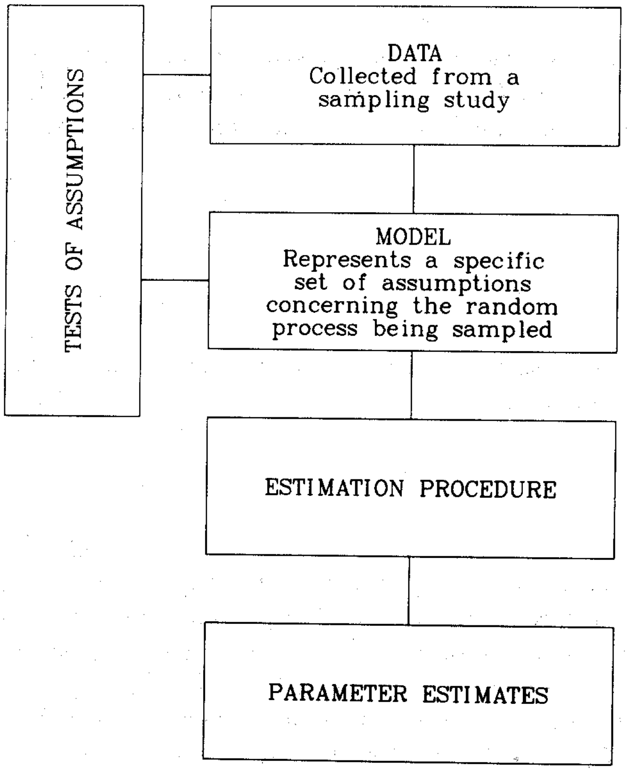

TESTS OF ASSUMPTIONS
DATA
Collected from a sampling study
MODEL
Represents a specific set of assumptions concerning the random process being sampled
ESTIMATION PROCEDURE
PARAMETER ESTIMATES

Fig. 2.1. The roles of data, models, and estimation procedures and their relationships in producing estimates of parameters.

estimation procedures are based on a specific set of assumptions concerning the sampling experiment. In Chapter 3, we deal more directly with models of capture-recapture experiments and the assumptions they represent.

Education in the biological sciences often has not included adequate explanation of either the scientific method or the theory of inference. The books by *Baker and Allen (1968)* and *Goldstein and Goldstein (1978)* provide an introduction to both subjects, and those by *Popper (1962)* and *Medawar (1969)* present more technical discussions. The use of the scientific method represents a broad philosophy concerning rigorous inference. We might ask, "What justifies a conclusion?" The answer to this question always involves "valid methodology."

Valid methodology is a package of essential ingredients: proper hypothesis formulation, design of data collection, conduct of the experiment or sample, rigorous analysis of the data to test the hypothesis, and inference (a conclusion) to reject or support, but never to "accept" the hypothesis. Inference depends critically on study design and data analysis.

Estimation

In the discussion of models, we frequently referred to parameters and estimation procedures (or estimators). These terms are discussed below. [See *Kendall and Buckland (1970)* for related material].

Parameter. A parameter is the true population value of interest, expressed as a number. In capture-recapture studies the parameter of interest is either population size N, the total number of animals in the population, or population density D, the number of animals per unit of area. Examples of other important parameters in biological work are annual survival rate, average clutch size, average number of corpora lutea, and the proportion of males in a population.

Estimator. An estimator is a mathematical expression that indicates how to calculate an estimate of a parameter from the sample data. Estimators are necessary because we almost never know the value of the population parameter. The following formula for calculating a mean is the estimator most commonly used by biologists.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad . \tag{2.1}$$

This formula for the sample mean is an estimator of the population mean μ. The Petersen-Lincoln estimator of population size is simply

$$\hat{N} = \frac{n_1 n_2}{m_2} \quad ,$$

where n_1 and n_2 are the total number of animals captured on the first and second sampling occasions, respectively, and m_2 is the number of marked animals captured on the second occasion *(Seber 1973:59)*. In general, an *estimator* is shown with a "hat" over the parameter to indicate clearly that it is an estimator, rather than the true parameter. For example, $\hat{N}$ and $\hat{D}$ are estimators of the parameters N and D. An *estimate* is the numerical value resulting from substituting the sample data into the estimator. For example, the data set {4, 2, 7, 3, 4}, when substituted into the estimator given in Eq. (2.1), produces the estimate 20 ÷ 5 = 4.

In most practical situations, a "proper" estimator is not obvious. In other words, intuition is often of little help in deriving a good estimator of a parameter. Without a model, we can only guess at valid

Carl George Johannes Petersen

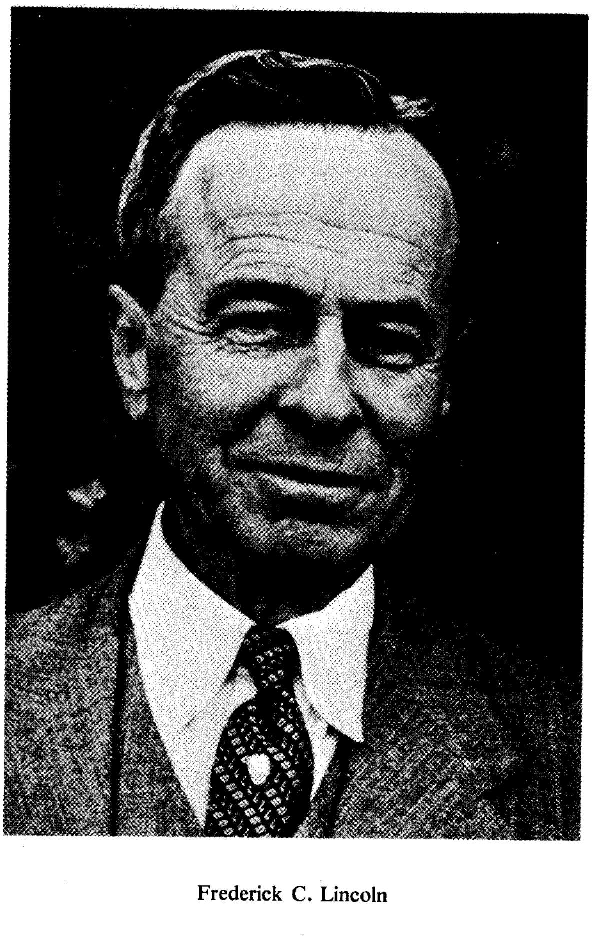
The practice of estimating population size from capturing, marking, and recapturing both marked and unmarked animals had its beginning with the Petersen estimate. Carl George Johannes Petersen was born in Denmark in 1860 and was a famous fishery scientist and the director of the Danish Biological Station from its founding in 1889 until his retirement in 1926.

The estimation method now bearing his name was published in 1896 and stemmed from his work on plaice. He invented a brass tag that he attached to the fish, to study their migrations. When one-third of the marked fish in his study were recaptured by fisherman, Petersen recognized that this information constituted a basis for estimating population size.

Petersen was awarded the LL.D. degree *honoris causa* in 1912 from the University of St. Andrews, Scotland. Much additional information about his work can be found in *J. Du Conseil* 1928, 3(2):135-138 and in the *Report of the Danish Biological Station,* 1940, Copenhagen.

Nearly every terrestrial ecologist is aware of the Lincoln Index, a simple estimator developed by Frederick Lincoln to estimate the size of the waterfowl population in North America from banding and recovery data. The same method was derived before the turn of the century by C. G. J. Petersen for fishery problems and as far back as the 17th century by P.S. LaPlace for human population problems. In the past half century the Petersen-Lincoln method has found many uses in a variety of disciplines.

Frederick Lincoln, born in Denver, Colorado, in 1892, spent his life in the study of birds. He joined the Biological Survey (now the U.S. Fish and Wildlife Service) in 1920 and was responsible for the bird banding program, a cooperative program among Canada, the United States, and Mexico. He developed the flyway concept for management of migratory waterfowl and was the leading authority on the distribution and migration of birds. He devoted much of his energy to developing better methods of trapping and banding birds and to developing procedures for recording, reporting, and analyzing banding data. He was awarded an honorary Doctor of Science degree by the University of Colorado in 1956. Additional information concerning his life appears in Auk 79:494-499, written after his death. (Photograph with permission of the American Ornithologists Union.)



Frederick C. Lincoln

estimators. Such guesses are typically poor (if not incorrect), and no estimates of precision can be made without a model. However, with a proper model relating the data, assumptions, and parameters of interest, we can derive valid estimates of parameters in the model routinely, by very general, available methods. The principal method used in statistical estimation over the past half century has been the method of maximum likelihood (ML), which is discussed later in this section.

Our goal is to use good estimators to produce estimates of the parameters of interest. To evaluate estimators, we need criteria by which to judge them. In statistical theory, two essential criteria arise from the concepts of bias and precision.

Accuracy. Accuracy is defined as "exact conformity to truth" or "freedom from error or defect." This ideal is unattainable in sampling studies and inductive inference; therefore, we rely on the concepts of bias and precision (defined below) as aids in making good inductive inference.

Bias. Ideally, an estimator should be free of bias. That is, if we were to repeat a sampling experiment under the same conditions on a very large number of occasions, each time computing an estimate from the sample data, the average of the estimates should equal the parameter being estimated. Frequently, we denote the "average" value of an estimator $\hat{N}$ by $E(\hat{N})$, read as the "expected" value of $\hat{N}$ or the "average" value of $\hat{N}$ over a very large number or repetitions. Thus if $E(\hat{N}) = N$, we say that the estimator $\hat{N}$ is unbiased. Note that bias is a conceptual quantity because usually we have only one set of data and can compute only one value of $\hat{N}$ from the data. Bias relates strictly to the *average* performance of an estimator.

It is often convenient to discuss the subject of bias or biased estimators in two classes—small-sample bias and model bias. These terms are not well established in the literature, but the distinction between them is important for biologists. *Small-sample bias* is often of negligible importance to a biologist in the analysis of one or only a few data sets. This type of bias decreases as sample size increases. Biologists frequently encounter a "biased" estimator for the first time when estimating the variance from a random sample. One learns that the ML estimator

$$\text{variance} = s_1^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

is biased. However, we find that the expression

$$\text{variance} = s_2^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

is unbiased. When n is 30 to 40 or more, the difference between s_1^2 and s_2^2 becomes negligible. This is an example of small-sample bias. Another example relates to the Petersen-Lincoln estimator.

$$\hat{N}_1 = \frac{n_1 n_2}{m_2}$$

is biased, but

$$\hat{N}_2 = \frac{(n_1 + 1)(n_2 + 1)}{(m_2 + 1)} - 1$$

is unbiased in some instances and virtually unbiased in others (see *Seber 1973:60* for details). However, the difference between the two estimators is negligible if the sample size is reasonably large. For example, compare using $n_1 = 60$, $n_2 = 50$, and $m_2 = 30$.

A much more serious problem deals with ***model bias***. This problem arises when important assumptions such as equal catchability, made in creating the model, are incorrect for a particular situation. An incorrect assumption can cause large bias even when sample sizes are quite large because it is theoretically independent of sample size. An example will illustrate this important concept. Consider a theoretical population of animals in which individuals become trap shy after they have been captured the first time. Given that the population is composed of 400 animals ($N = 400$), the probability of first capture is 0.20 ($p = 0.20$), and the probability of recapture drops to 0.05 ($c = 0.05$), we can ask what bias could be expected if we (incorrectly) assume that the population is equally catchable (that $p = c$) and estimate the population size under this assumption. Using simulation procedures, we find that the expected value of the estimator $E(\hat{N})$ is about 1071, which illustrates that model bias can be very substantial. Formally, if $E(\hat{N}) = 1071$, the bias is $E(\hat{N}) - N = 1071 - 400 = 671$. Interested readers can find this example and others in *Otis et al. (1978:127)*.

Expressing bias as a percentage is often useful; called "percent relative bias," the expression is defined as

$$PRB = \frac{E(\hat{N}) - N}{N} \times 100 \ .$$

In the example above, PRB = 168; that is, $[(1071 - 400)/400] \times 100 = 168$.

Precision. Precision relates to the repeatability of a result. If, for example, a sample is drawn and the total population size is estimated to be 10 700, will the next sample yield an estimate of 400, or 31 900, or will it be near 10 700? Repeatability is an integral concept in science. The precision of an estimator is measured by the sampling variance and its square root, called the standard error of the estimate (Fig. 2.2). People not familiar with the concept of repeatability, with numerical quantities plotted as histograms (bar graphs), or with theoretical probability functions overlying the histograms should examine Fig. 2.3.

In this context, our concern is an estimate of the sampling variance of the estimator $\hat{N}$, denoted $\widehat{\text{var}}(\hat{N})$, as a measure of precision or repeatability. The sampling variance and standard error $[\text{se}(\hat{N}) = \sqrt{\text{var}(\hat{N})}]$ are measures of the variability of the individual estimates around their expected or average value over different samples. Of course, we would prefer to have an estimation procedure that would give very similar estimates from different samples.

The concepts of bias and precision are illustrated in Figs. 2.4 and 2.5. The information in Fig. 2.4, adapted from *Overton and Davis (1969)*, shows a series of targets and shot patterns. If we make an analogy by considering each shot as an estimate (made from sample data by using a given estimator), we can illustrate the concepts in terms of standard frequency diagrams as shown in Fig. 2.5.

Research scientists and managers always prefer the unbiased and precise estimator illustrated in Figs. 2.4a and 2.5a to the very precise, incorrect estimate depicted in Figs. 2.4c and 2.5c, which is considered especially undesirable. Unfortunately, situations like those depicted in Figs. 2.4b-d and 2.5b-d are probably very common in attempts to estimate population size from the data for capture-recapture studies. In this primer we attempt to improve the accuracy of such estimates by emphasizing study design, increased sample size, and improved methods of analysis.

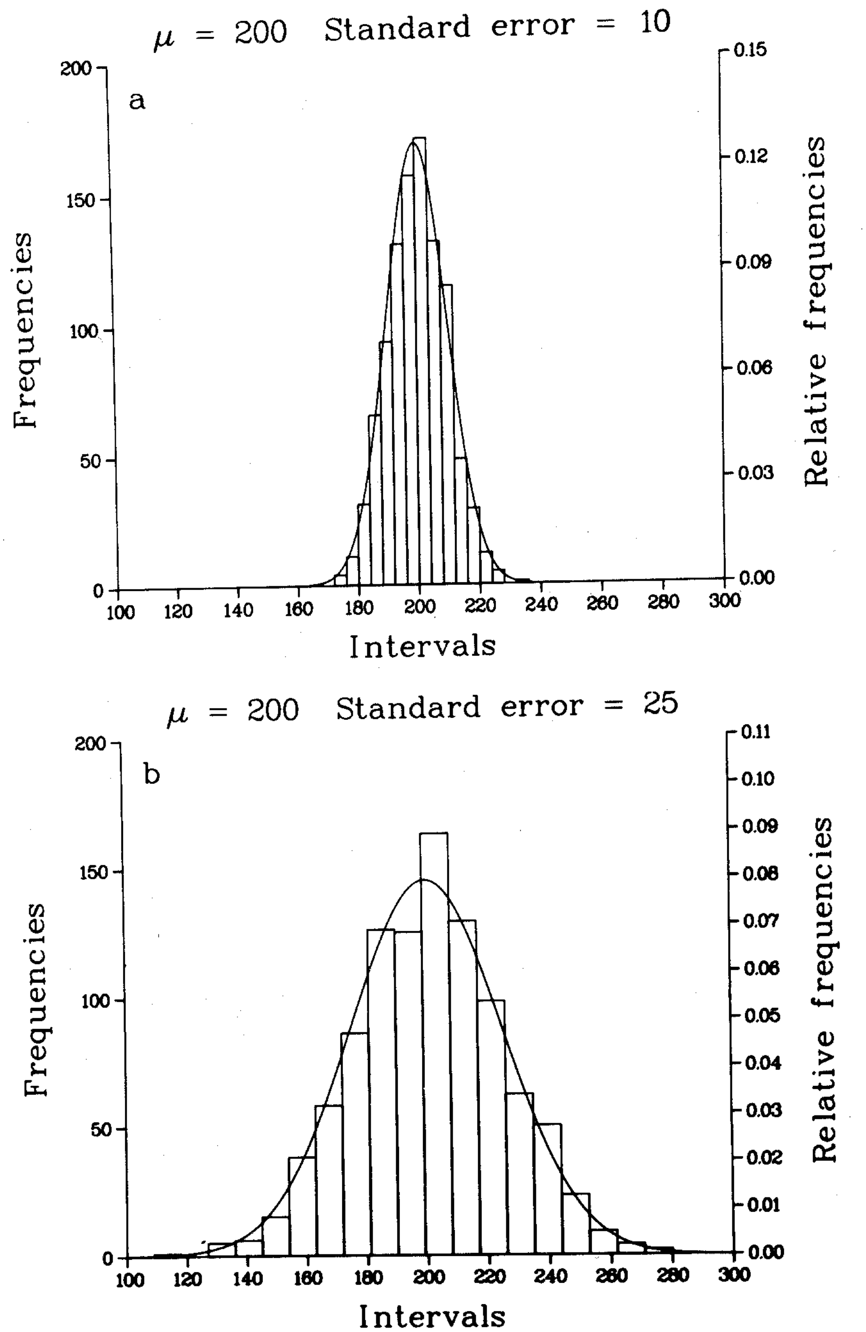


Fig. 2.2. Histograms of two data sets with differing amounts of variability. In a, the data are clustered fairly closely around the population parameter of 200. The spread of the data around the mean is measured by the standard error, which is 10 in this example (variance $= 10^2 = 100$). For example, the range of the data is about 170 to 230 for 1000 data points. In contrast, the data shown in b are much more variable, as reflected by the larger standard error of 25 (variance $= 25^2 = 625$). Here the range is from 130 to 280, also based on 1000 data points. In each instance, a normal curve has been fitted.

Stochastic Processes and Models. Processes that are not completely predictable (deterministic) are termed stochastic. Examples include coin-flipping, card games, all forms of gambling, weather patterns, stock market fluctuations, and, most important in the context here, all sampling data.

It follows that stochastic models are appropriate for data that arise from a stochastic process. Such is the case in capture-recapture and removal-sampling studies. The biologist need not understand the details

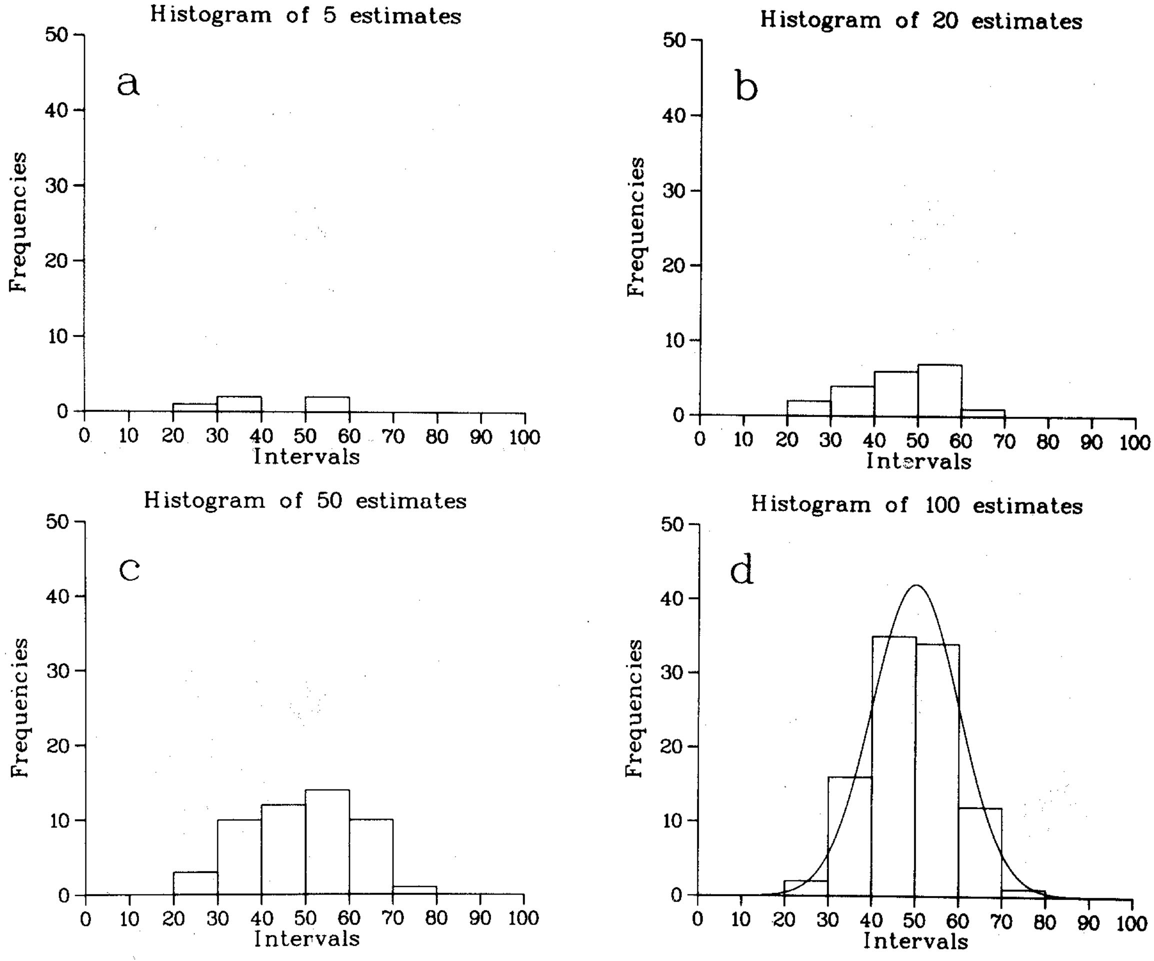


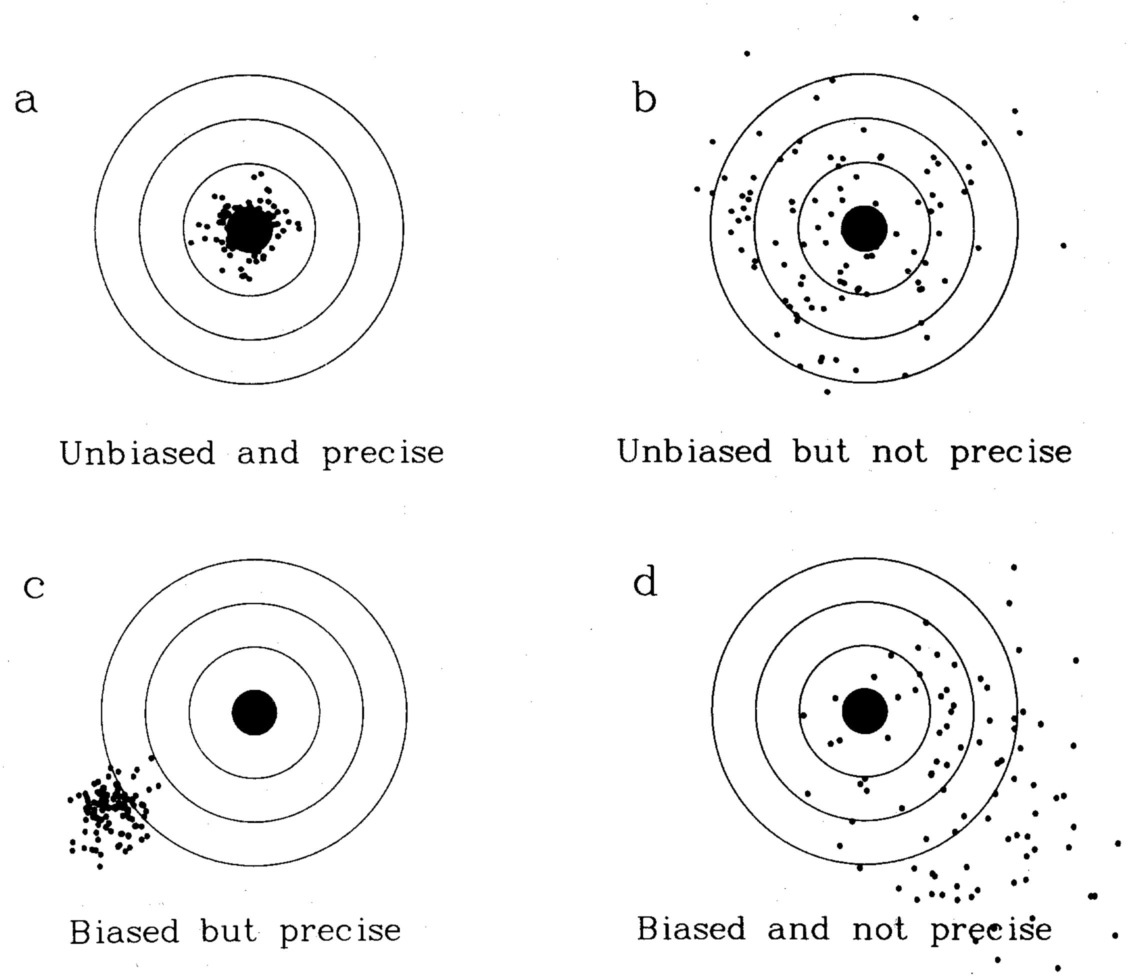
Fig. 2.3. Histograms of estimates based on samples of sizes 5, 20, 50, and 100. Computer simulation studies are useful in assessing whether an estimator has a normal sampling distribution for a given sample size. In the case shown, the estimator is distributed approximately normally as indicated in d.

of stochastic models, model building, or estimation theory to use the methodology presented in this primer; our purpose is to concentrate on concepts rather than on mathematical or statistical derivations.

Variation

Important variation is found everywhere in the biological sciences. It is crucial to understand clearly the two distinctly different types of variation in capture-recapture and removal studies.

Spatial and Temporal Variation. First, there is the obvious variation in space and time in the real world. Neither animal density nor plant cover is uniform over the State of Utah; both are clear examples of spatial variation. Animal numbers fluctuate over time; these changes constitute temporal variation. As another example consider a 20-km portion of stream divided equally into 20 numbered segments. Assume that we know the exact number of fish in each segment. In other words, we know the



Unbiased and precise

Unbiased but not precise

Biased but precise

Biased and not precise

Fig. 2.4. Targets and shot patterns illustrate the concepts of bias and precision. The goal of the marksman is shown in a. Although c might not be too bad for a marksman, who merely needs to adjust his sights, it is the worst case for the biologist attempting to estimate population abundance. It is a highly precise, incorrect estimate. Furthermore, after completing a shooting session, the marksman can approach the target and compare his shot pattern with the bull's eye (true parameter). In contrast, the biologist usually will never know the true parameters; therefore, his inductive inferences from the sample data about the true parameter must be made carefully.

true parameters N_1, N_2, N_3, . . ., N_{20}, and these parameters vary. As a measure of this variation, the population variance σ^2 could be computed by the usual definition,

$$\sigma^2 = \frac{\sum_{i=1}^{20} (N_i - \bar{N})^2}{20} \; . \qquad (2.2)$$

This quantity measures the variation in population size over space (20 segments of the stream). This type of variation is most frequently studied in basic statistics courses, where "sampling error" or "measurement error" is ignored.

Stochastic Variation. Second, there is stochastic variation of basically unpredictable events such as coin flipping, success of a nest, or time of death of an animal. This kind of variation is somewhat more difficult to understand. We will use the same example of the 20-km stream and consider segment 3 of the

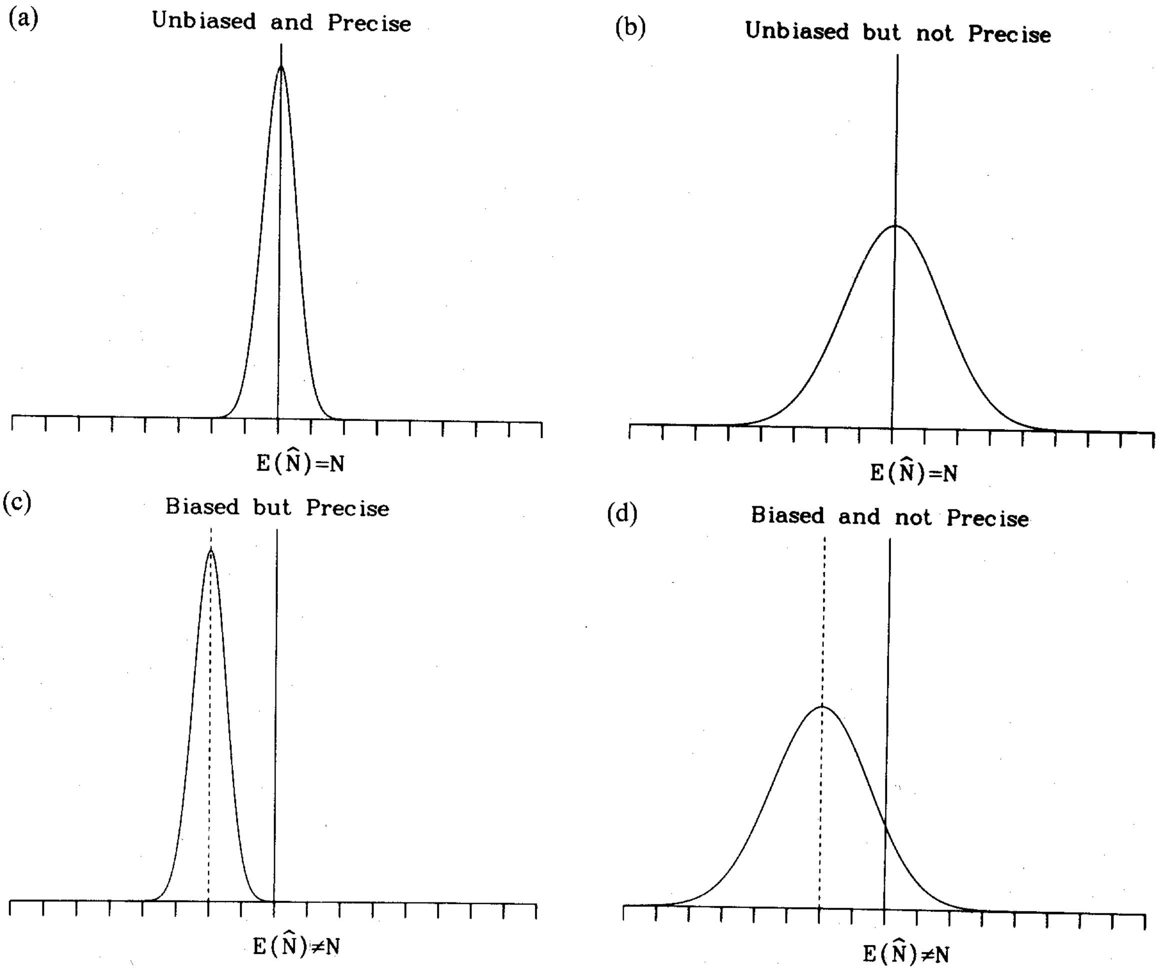


Fig. 2.5. An illustration of the concepts of bias and precision. Here, the information from a large number of samples is used to compute individual values of $\hat{N}$, which then are graphed as frequency distribution. (Alternatively, they could have been shown as simple histograms.) We desire an estimator with properties shown in a, in which the average value of $\hat{N}$ is equal to the parameter N and the variation of the estimates around N is small. Precision is measured by the sampling variance of the estimator $|\text{var}(\hat{N})|$ and relates directly to the spread of the frequency distribution. (Compare these spreads with those shown in Figs. 2.4.a and b.)

stream. In this instance, we have to estimate the number of fish in the 1-km segment. (Of course, a complete census would give us the parameter N_3 directly, but we will assume that such a census is not feasible.) A removal experiment using electrofishing methods on 3-5 occasions and analysis of the resulting sample data would provide an estimate $\hat{N}_3$ of the parameter N_3. If the survey were repeated, we would get a second estimate $\hat{N}_3$ of N_3, and so on. It is this variation among the estimates that we consider in this primer, but it is important to recognize the existence of spatial and temporal variation.

Catching fish, however, is not a perfectly predictable event—it is, instead, a stochastic process. Therefore, we obtain $\hat{N}_3 = N_3 \pm \varepsilon_i$, where ε is sampling error or measurement error. (Note, if $\hat{N}_3$ is an unbiased estimator, then the average or expected value of ε_i must be zero.) Sampling variation occurs whenever we sample from a defined population (N_3 in this instance) and attempt to estimate a parameter associated with the population. We denote sample variation associated with the estimators $\hat{N}$ and $\hat{p}$ as var($\hat{N}$) and var($\hat{p}$), respectively. Usually sampling variances themselves are only estimated, and we employ the notation $\hat{\text{var}}(\hat{N})$ and $\hat{\text{var}}(\hat{p})$ to signify that these are estimators.

Standard Errors and Sample Size. The specific meaning and the concept of the terms "sampling variation" and "standard errors" are sometimes difficult to grasp. Expressions such as var($\hat{N}$) or se($\hat{N}$) are recognized as measures of precision, but the underlying concepts often are not understood clearly. The following example illustrates these concepts.

Consider a small island off the coast of Alaska with a population of colonial sea birds nesting on the island's rocky edges. There are exactly 11 000 nests on the island in a certain year, and in this species the female lays only 1 egg. An investigator must find the proportion of the nests that are successful. In other words, he wants to know the parameter p. We know that p in this example is 0.70, that exactly 70% of the nests were successful; however, this fact is unknown to the investigator. He could contemplate a complete census of all nests and, by classifying the nests as successful (p) or unsuccessful (1 − p), arrive at the exact value of p. This approach is impossible, however, because the time and expense would be prohibitive. Therefore, he must sample a fraction of the total nests and estimate the parameter using the estimator $\hat{p}$.

Let us assume that the investigator decides to randomly select a sample n of size 25 nests. If s is the number of successful nests, then an estimator of p is computed as $\hat{p} = s/n$. He conducts the survey, finds the random 25 nests, and observes that 15 were successful and 10 failed. Thus the estimate of the proportion of successful nests is 15/25 = 0.60. We know that p = 0.70 in this example, and that this estimate of p is not too bad. The investigator, however, has no idea at this point how good his estimate is. In other words, he does not know how close his estimate is to the true, unknown parameter.

The investigator then decides to run a second survey of 25 different nests to check his first estimate. This survey yields s = 20 and an estimate of $\hat{p} = 20/25 = 0.80$ or 80% successful. Now his confidence is shaken a bit and he decides to conduct eight more surveys. The results, including the first two estimates, are as follows: 0.60, 0.80, 0.69, 0.52, 0.88, 0.69, 0.76, 0.64, 0.56, and 0.76. All 10 values are estimates of the same parameter p. The fact that the estimates vary represents sampling variation.

Sampling variation occurs when we sample a population (in this example, 11 000 nests) and *estimate* the parameter p, rather than making a complete census and *computing* the parameter p exactly. It is clear, once we consider the matter, that the variation in the 10 estimates would have been much smaller if a larger sample had been taken (if n = 500 instead of 25) in each of the 10 surveys. Now we can ask, How do I measure how much variation to expect in the estimates for a sample of a certain size? The answer to this question is in the realm of mathematical statistics. Theory exists to enable calculation of the variance of the estimate, var($\hat{p}$) in this example, which is a measure of the sampling variation we can expect. The formula for estimating the variance of a proportion is given by $\hat{\text{var}}(\hat{p}) = [\hat{p}(1 - \hat{p})]/n$. In the first sample of 25 nests, $\hat{p} = 0.60$; therefore, $\hat{\text{var}}(\hat{p}) = (0.60)(0.40)/25 = 0.0096$.

The square root of the variance, a more useful quantity, is called the "standard error." That is, se($\hat{p}$) = $\sqrt{\text{var}(\hat{p})}$ or, in the above example, $\hat{\text{se}}(\hat{p}) = \sqrt{0.0096} = 0.098$. The standard error is used in calculating confidence intervals and coefficients of variation. The estimated coefficient of variation (cv) of an estimate is defined as

$$\text{cv}(\hat{\theta}) = \frac{\hat{\text{se}}(\hat{\theta})}{\hat{\theta}} ,$$

where $\hat{\theta}$ is, in general, an estimate of some parameter θ. In our example from above,

$$\begin{aligned}\text{cv}(\hat{p}) &= \frac{\hat{\text{se}}(\hat{p})}{\hat{p}} \\ &= \frac{\sqrt{[\hat{p}(1 - \hat{p})]/n}}{\hat{p}} \\ &= \frac{0.098}{0.60} \\ &= 0.16 \ .\end{aligned}$$

In biological studies, a coefficient of variation (of an estimator) of 0.10 or less is considered good, so we see that the estimate from the first survey of 25 nests has only fair precision. The most effective means of increasing the precision of our estimate is to increase the sample size.

A 95% confidence interval in our example is computed as $\hat{p} \pm 1.96\hat{se}(\hat{p})$ or $0.60 \pm 1.96 \times 0.098 = 0.60 \pm 0.192$; hence the interval is (0.41 to 0.79). If the investigator had run a large number of independent surveys, each of a random sample of 25 nests, 95% of the confidence intervals would be expected to include the true parameter (which, in our example, is 0.70). Clearly a confidence interval as wide as (0.41 to 0.79) is not of much use. The important concept here is that expressions like $\text{var}(\hat{p})$, $\text{se}(\hat{p})$, and $\text{cv}(\hat{p})$ are measures of precision (repeatability) or sampling variation.

A final point will illustrate the advantage and importance of sample size. Assume that the 10 surveys of nests were pooled and the effort is considered as 1 survey, with a sample size of 250 (10 surveys × 25 nests per survey = 250). Then $\hat{p} = 0.69$, which is very close to the true parameter of 0.70. The sampling variance is then $\hat{\text{var}}(\hat{p}) = [0.69 \times (1 - 0.69)]/250 = 0.000855$, and $\hat{\text{se}}(\hat{p}) = 0.029$. The coefficient of variation of the estimate is only 0.04. The 95% confidence interval is much narrower (0.63 to 0.75), a good indication that the variation in the estimates of p would not vary much from survey to survey. In other words, the repeatability is good, which allows much stronger inference about the parameter of interest. Repeatability is an important part of inductive inference.

A Further Example of Variation. Now let us assume we have five islands, each similar to the others, and each supporting a colony of birds of the same species. The situation is summarized as follows.

Island Number	Unknown Parameter[a]	Estimate	Sampling Variance
1	p_1	$\hat{p}_1 = p_1 \pm \epsilon_1$	$\hat{\text{var}}(\hat{p}_1)$
2	p_2	$\hat{p}_2 = p_2 \pm \epsilon_2$	$\hat{\text{var}}(\hat{p}_2)$
3	p_3	$\hat{p}_3 = p_3 \pm \epsilon_3$	$\hat{\text{var}}(\hat{p}_3)$
4	p_4	$\hat{p}_4 = p_4 \pm \epsilon_4$	$\hat{\text{var}}(\hat{p}_4)$
5	p_5	$\hat{p}_5 = p_5 \pm \epsilon_5$	$\hat{\text{var}}(\hat{p}_5)$

[a] The proportion of successful nests.

If we average the five estimates, we see that both spatial and sampling variation are involved. For example, define $\bar{p}$ as the average of the five estimates.

$$\bar{p} = \frac{\sum_{i=1}^{5} \hat{p}_i}{5} .$$

The variance of $\bar{p}$ is

$$\hat{\text{var}}(\bar{p}) = \frac{\sum_{i=1}^{5} (\hat{p}_i - \bar{p})^2}{4} . \qquad (2.3)$$

It should be clear that the $\text{var}(\bar{p})$ has two components of variation: sampling variation (given in Column 4 of the table above) and spatial variation among the islands [similar to the expression in Eq. (2.1)]. Spatial variation enters the computation of $\bar{p}$ because physical and biological factors may cause differences among the islands. The separation of the sources of variation in expressions like Eq. (2.3) is a difficult subject known as "variance components" in statistics. We will not explore the subject here because it would take us too far afield. In this example, both spatial and sampling variation are likely to be quite important, and biologists should keep the two sources in mind.

As a second example, consider the sunfish in all the small ponds (potholes) in a particular county in Minnesota. There are 89 ponds, varying in size from 0.1 to 1.8 ha. Each pond is assigned an identification number, 1, 2, . . ., 89.

Spatial variation arises because the actual population size N of sunfish will differ among the 89 small ponds. Some may not have any sunfish (N = 0), whereas others may have large populations. We also might suspect that population size could vary with pond size. Temporal variation relates to the actual population size N in a particular pond as it changes over time, because of births and deaths. Temporal variation is often seasonal. Note that both spatial and temporal variation relate to changes in N.

The stochastic component is encountered in many studies of animal populations because we usually cannot count each member of a population to determine N. Instead, a sampling procedure, such as capture-recapture, must be used to *estimate* N. The sample data (the X matrix or some summary of it) are used with an estimator, such as $\hat{N} = (n_1 n_2)/m_2$, to compute an estimate of the population size for a given point in time and space. The estimates vary with each sample drawn (see Fig. 2.3). It is this variation that we call sampling variance; it is a measurement error, caused by the stochastic nature of the sampling and capturing process and denoted as $\text{var}(\hat{N})$. Sampling variation can be illustrated by looking at pond 32. Because the pond is small, electrofishing was used to sample the population of sunfish each day for 4 days. The fish were returned to the pond after each sample day. An appropriate estimator was used to estimate N from the first 4-day sampling study, and the estimate was 243 sunfish. A second 4-day sampling study was conducted, and the estimate was 202 sunfish. A total of five 4-day sampling studies yielded estimates of the parameter N as 243, 202, 157, 231, and 192. This variation is called sampling variation. Fortunately, it can be estimated without having to conduct replicate samples. Sampling variation is a measure of the precision among the estimates.

Many practical problems necessarily involve both types of variation. However, this primer treats the estimation problems for a given point in time and space, such as a single large trapping grid within a larger area, and thus spatial and temporal variation are not relevant here.

Properties of a Good Estimator

Because estimators are functions of random variables (the sample), they possess probability (sampling) distributions. Estimators, therefore, must be derived from probabilistic (for stochastic) models. A good estimator

(1) is robust to crucial assumptions: it is not very sensitive to the failure of some important assumptions. It is robust to model bias.

(2) exhibits minimum variance: it is the most precise estimator possible. It makes full use of all the information in the sample.

(3) is distributed normally: for the sample sizes usually encountered, the distribution of the estimator is normal. If not normal, the distribution should be known, at least approximately.

(4) is unbiased, given the assumptions: at least the small-sample bias is zero when sample size is large.

All of these properties are more complex than we have indicated; however, those listed should provide a working basis for an understanding of the material that follows. [The interested reader should consult a text on mathematical statistics, such as *Lindgren (1968:266-278).*]

A word of caution is appropriate here because most capture-recapture and removal analysis methods fall somewhat short of our expectations for a good estimator. For example, most estimators have a slight small-sample bias, most are nonnormal (skewed to the right) for the sample sizes typically encountered, and most are not robust to the failure of certain assumptions. Poor coverage of confidence intervals sometimes is due to nonnormality. However, estimators are derived by using the ML method, which guarantees that they will be minimum variance estimators, at least asymptotically. Many methods described in the literature are *ad hoc* methods, and their properties are generally unknown. Such deficiencies call for careful design, field work, and analysis. These needs are the central focus of the material to follow.

Estimation Methods

The data from capture-recapture or removal studies are collected from samples, thus requiring a probabilistic treatment of the data to derive good estimation and inference procedures. As we have shown, model formulation in this context begins with a set of explicit assumptions. A probability model for the sampling distribution of the X matrix (the basic data) is derived to express the assumptions quantitatively. A probability model is a form of mathematical representation of the observed data under a specific set of assumptions and, as such, it provides a basis for quantitatively and explicitly incorporating the specific assumptions about closure and capture probabilities and for developing the point and interval estimators by rigorous statistical estimation techniques. Most parameter estimators used here were derived by using the maximum likelihood (ML) method.

Estimators derived by this method are optimal, at least for large samples. This general method of estimating parameters was derived in the early 1920s by the famous statistician and geneticist, Sir Ronald A. Fisher, and it has been the backbone of statistical estimation theory for more than 50 years. Alternative estimation procedures, such as method of moments and minimum chi-square, have been developed, but the ML method is generally accepted as the best. The interested reader is referred to any book on mathematical statistics for additional material on the ML method; one example is *Kempthorne and Folks (1971:242).*

Random Sampling. Many authors state that random sampling is required in capture studies. This assumption stems from ball and urn experiments (Fig. 1.1), in which marked and unmarked balls are shaken completely, and a random sample is taken at the end of each sampling occasion.

Traditional sampling methods include procedures for drawing random samples. Use of the procedures requires knowledge of the sampling probabilities (for finite populations). Deliberate control over the elements to be sampled is required. However, such control is clearly absent in capture studies of animal populations. The sampled animals are not selected by the investigator; the capture probabilities are not preset, nor are they even fixed during the course of the study. It is unrealistic to think that an animal captured in one corner of the trapping grid may be captured subsequently in the opposite corner. There is simply no basis for thinking that samples are drawn randomly in capture studies of animal populations. *Mendenhall et al. (1971:187)* and *Johnson and Kotz (1977:248:250)* present a different view, although they recognize some practical problems. *Feller (1950:45)* gives an example of the capture-recapture method for a hypothetical fish population (essentially the Petersen-Lincoln estimate) and mentions in a footnote that the method is used widely in practice.

The concept of random sampling does not apply to situations assumed in Models M_b, M_h, M_{bh}, M_{tb}, M_{th}, and M_{tbh}. The goal of these models is to provide an analysis of the sampled data in the face of behavioral response and heterogeneity, both of which are contrary to the traditional role of random sampling.

Robustness of an Estimator. In addition to the important properties of bias and precision, an estimator also may be judged by its robustness to the failure of certain assumptions. In the previous section, we described how every estimation procedure is based on a model that represents a specific set of assumptions concerning the population or process being sampled. The concept of robustness relates to the question, How well does the estimator perform if one (or more) of the assumptions on which it is based is false? If the performance of an estimator is little affected by the failure of an assumption, it is said to be robust to the particular assumption (also see *Otis et al. 1978:15*). As an example, recall the discussion of model bias in the previous section, where we considered a capture-recapture experiment in which the probability of first capture and recapture were 0.20 and 0.05, respectively. We found that if we incorrectly assumed that probability of first capture and probability of recapture were equal, the estimation procedure based on these assumptions was very biased and therefore performed poorly. Thus the estimator is not robust to the assumption that the probability of recapture and first capture are equal. Unfortunately, most methods for the analysis of capture-recapture data are not very robust. In particular, the assumption of equal catchability is important, because most traditional methods (the Petersen-Lincoln and Schnabel estimators and the Zippin removal method) are not robust to failure of this assumption *(Burnham and Overton 1969; Otis et al. 1978:123-133).*

Closed-Form Solutions to the ML Estimator. In general, ML estimators of unknown parameters, such as N or D, are found by using calculus techniques on a function closely related to the stochastic model and called the "likelihood function." In many cases, the ML estimator can be written in

Zoe Emily Schnabel

The "Schnabel estimate" has been the backbone of population size estimation, assuming closure, for the past 40 years. It provides an easy-to-compute method for estimating population size in the case where animals are captured, marked, and recaptured over t occasions. Before Schnabel's work, only t = 2 occasions could be handled by the Petersen-Lincoln method.

Zoe Emily Schnabel (Mrs. George S. Albert since the late 1930s) completed an A.B. degree at Oberlin College and an M. A. degree in mathematics at the University of Wisconsin (1937). She taught mathematics and biometry at Ohio State University and mathematics and statistics at the University of Tennessee before retiring in 1969.

Schnabel's work on capture-recapture studies, however, was done between 1936 and 1938, when she was a graduate assistant in the Computing Laboratory of the Mathematics Department in Madison, Wisconsin. The Laboratory had been established in the early 1930s to assist university researchers with the statistical analysis of data. Members of the mathematics faculty served as consultants, and Schnabel and others assisted with the computations using calculating machines of the era.

The Schnabel estimate was an outgrowth of the work done by Schnabel, E. Hull, and M. Ingraham in the Mathematics Department and D. Juday, a limnologist in the Biology Department. Schnabel concluded her paper with an observation that, unfortunately, many have disregarded: "It should be emphasized, however, that none of the solutions can be expected to provide more than an estimate of the general order of magnitude of the total population." (Photograph taken about 1937-1938.)

Calvin Zippin became interested in removal sampling while at Johns Hopkins University in the early 1950s, when he began consulting with a group in vertebrate ecology. The group was involved in trapping rats in the Baltimore area to estimate population size. The removal method was known, but its statistical properties, including the standard error formulas for population size and capture probabilities, had not been explored thoroughly. This consulting work developed into his doctoral dissertation at Johns Hopkins.

Zippin joined the faculty of the University of California, San Francisco, in 1953, and he has worked there since then concentrating on biometry and the epidemiology of cancer. He remains interested in capture-recapture and removal sampling.

Zippin, shown at the left, is standing with Sir Ronald A. Fisher. Fisher, an outstanding pioneer in statistical theory and practice, worked with several scientists involved in capture-recapture studies. Some of his theoretical developments in mathematical statistics form the basis for much of what we now call the field of statistics. (The photograph was taken in the early 1950s in North Carolina.)

Calvin Zippin

a simple form that is easy to use. For example, the Petersen-Lincoln estimator is the ML estimator of N for a special case of one of the models discussed in Chapter 3. The formula is written as

$$\hat{N} = \frac{n_1 n_2}{m_2} ,$$

where n_1, n_2 are the total number of animals captured on the first and second sampling occasions, respectively, and m_2 is the number of marked animals captured on the second occasion. We say the estimator "exists in closed form."

Many closed-form estimators found in the published literature are only approximations to the exact ML estimator; examples are the *Schnabel (1938)* and *Zippin (1956)* estimators.

Numerical Solutions to the ML Estimator. In capture-recapture models we rarely find that the exact ML estimators exist as simple formulas (as shown above). To illustrate this, consider the model developed by *Darroch (1958)* in which there are four sampling occasions and the capture probabilities

John N. Darroch

In many ways, John Darroch's work represents a cornerstone in capture-recapture theory. He studied optimal estimation in the model underlying the Schnabel method for closed populations, laid the foundations for the fully stochastic open model developed later and independently by George Jolly and George Seber, and developed the theory for stratified populations—a subject that later captured the attention of Neil Arnason. Also, he supervised the Ph.D. program for Seber at Manchester University.

Dr. Darroch received his undergraduate and early graduate training in mathematics and statistics at Cambridge University in England. He took a lectureship in mathematical statistics at the University of Cape Town, South Africa, in 1955. There, he became interested in the problem of estimating the number of species in a marine environment. This interest led to three papers on capture-recapture, which were published in *Biometrika* and were accepted as a Ph.D. thesis at the University of Cape Town. He returned to England for 3 years before going to Australia in the early 1960s. He is now at the Flinders University in South Australia. (Photograph taken in late 1950s.)

are assumed to vary only by time. The approximate ML estimator for N for this model is the unique value of N that satisfies

$$\left(1-\frac{M_5}{N}\right)=\left(1-\frac{n_1}{N}\right)\left(1-\frac{n_2}{N}\right)\left(1-\frac{n_3}{N}\right)\left(1-\frac{n_4}{N}\right) ,$$

where $M_5 = M_{t+1}$ = number of individuals caught during the study and t equals the number of sampling occasions. For example, the total caught on each of four occasions might be $n_1 = 30$, $n_2 = 15$, $n_3 = 22$, and $n_4 = 45$, and the total individual animals caught at least once, M_5, might be 79. Then, the ML estimate of N is the solution of the equation

$$\left(1-\frac{79}{N}\right)=\left(1-\frac{30}{N}\right)\left(1-\frac{15}{N}\right)\left(1-\frac{22}{N}\right)\left(1-\frac{45}{N}\right) .$$

There are efficient numerical methods to solve such equations. However, simple trial and error and a little patience will solve an equation as simple as this one. In general, the ML estimator for Darroch's model is derived by solving the equation

$$\left(1-\frac{M_{t+1}}{N}\right)=\prod_{j=1}^{t}\left(1-\frac{n_j}{N}\right) .$$

For t greater than 2, this equation cannot be solved algebraically for N. In other words, it is not possible to arrange the symbols algebraically in such a way that only N appears on one side of the equation and all other terms appear on the other side. The equation can be solved, but only on an iterative basis, by using a sophisticated trial and error numerical procedure. We say the equation does not have a simple, closed-form solution. Complex probability models often do not have simple estimators; nonetheless, complex models appear necessary to describe many capture-recapture studies adequately.

Although we cannot show simple closed-form estimators for most of the models to be discussed, it is the ML concept that is important and we leave it to the computer to do the arithmetic. The numerical methods employed are given in detail in *Otis et al. (1978:103-114).* The concept that is so important here involves the notion of a likelihood function.

Likelihood Function. Formally, the likelihood function is the joint probability density function of the sample data. In the context here, it is a function of the integer-valued parameter N and the real-valued parameter $\underline{p}$ (the vector containing all the probability parameters necessary to the model), given the discrete sample data contained in the $\underline{X}$ matrix (a matrix of zeros and ones).

The notation is not as complex as it may seem. For example, $\mathcal{L}(N,p|\underline{X})$ denotes the likelihood function of the unknown parameters N and p, given a specific set of sample data contained in the $\underline{X}$ matrix. As we will see in the next chapter, this is notation for the likelihood function for Model M_0 (Fig. 2.6). Two more examples will be given. For Model M_t, we have capture probabilities that may vary among sampling occasions; that is, $p_1, p_2, p_3, \ldots, p_t$. If we let these values be denoted as the vector $\underline{p}$, then the likelihood function for Model M_t can be denoted as $\mathcal{L}(N,\underline{p}|\underline{X})$ (*Otis et al. 1978:106*). In Model M_b, the parameter p is the probability of first capture, the parameter c is the probability of recapture, and the likelihood function for Model M_b is denoted as $\mathcal{L}(N,p,c|\underline{X})$.

The likelihood function is a formal way to express quantitatively the relative "likeliness" of several values that may be considered as candidates for $\hat{N}$. The ML method selects as the value for $\hat{N}$ the most likely one, on the basis

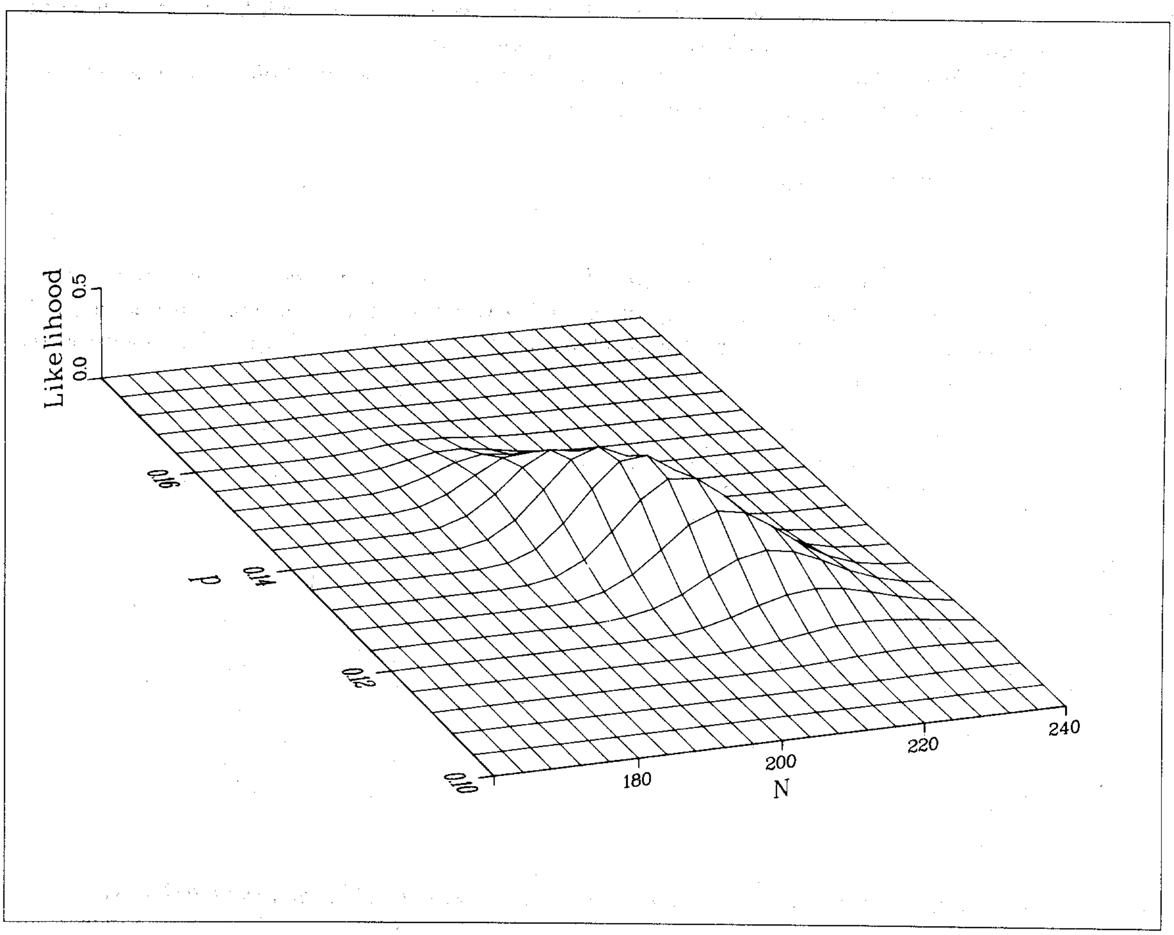


Fig. 2.6. A three-dimensional graph, showing the likelihood function for a given data set under Model M_0. This model involves only two unknown parameters: N = population size and p = constant capture probability. Given the data, the ML estimates of N and p are found as those values that maximize the likelihood function. The idea is that we are trying to find those values of N and p that make our data seem "most likely." Graphically, we see that values of N = 208 and p = 0.136 are approximately the ML estimates in this example. Of course, with different data, the likelihood function would be of a somewhat different shape and the values of N and p that maximize the function also would be different.

of the available data (hence the name, maximum likelihood). The use of likelihood theory and the ML method extends our intuition and ability to make inductive inference. To understand these concepts, consider a 2-sample capture experiment in which 40 animals are captured, marked, and released ($n_1 = 40$) in the first sample. In the second sample, 50 animals are captured; of these 25 (half) have marks from the first occasion ($n_2 = 50$, $m_2 = 25$). Before computing the ML estimate from the Petersen-Lincoln estimator, let us use our intuition and see what can be inferred from the data about population size. First, there must be at least 50 animals, because we caught that number on the second sample. In fact, at least 65 animals must be present (40 from the first occasion plus the 25 unmarked from the second occasion). However it seems unlikely that N = 1000 because half the animals caught on the second occasion were marked. Moreover, it seems fairly unlikely that the population could be as large as 500 or even 400. For example, if N = 400, only 40/400 = 10% of the population would have been marked before the second sample; if $m_2 = 50$ we would have expected only 5 marked animals in the second sample. Intuitively, we have reason to believe that N is at least as large as 65 and probably well below 400. The ML estimate is the value selected as the most likely, given the data we observed; $\hat{N} = (n_1 n_2)/m_2 = 80$.

The likelihood function is difficult to deal with directly because it involves products of often complicated terms. The likelihood function for Model M_o (see Chapter 3) is

$$\mathcal{L}(N,p|\underline{X}) = \frac{N!}{(N - M_{t+1})!}\, p^{n.}\,(1 - p)^{tN-n.}\ ,$$

where n. is the total number of captures and recaptures. By taking the natural logarithm of the likelihood function, we can deal with the sums of the terms; dealing with sums is nearly always more desirable than dealing with products of terms (see *Larson 1969:224-226*). This function, denoted as $\ell n\mathcal{L}(N,p|\underline{X})$, is called "the log likelihood function." The log likelihood function for Model M_o is

$$\ell n\mathcal{L}(N,p\,|\,\underline{X}) = \ell n\left[\frac{N!}{(N - M_{t+1})!}\right] + (n.)\ell n(p) + (tN - n.)\,\ell n(1 - p).$$

The term

$$\ell n\left[\frac{N!}{(N - M_{t+1})!}\right]$$

can be written more simply as

$$\sum_{j=N-M_{t+1}+1}^{N} \ell n(j)\ .$$

Details of likelihoods for some capture-recapture models are given in *Otis et al. (1978:102-114)*.

Basis for Rigorous Inference

Often, capture-recapture data are analyzed, and conclusions are drawn from them by *ad hoc* procedures. For example, M_{t+1} is used frequently as an "index to abundance." Another index used frequently is the number of animals captured per 100 trap nights. However, the use of indexes in science is to be discouraged because indexes lack the basic factors (Fig. 2.1) required for making inferences about parameters based on data. Indexes are useful only when they have been calibrated with the parameter of interest by using, for example, the theory of double sampling (*Cochran 1977*).

Initially, we must know what assumptions may be needed and which of them seem realistic. (See the previous section on Theory and Reality.) These assumptions should be built into a stochastic model that deliberately relates the sample data to the unknown parameters of interest. Then, a good estimation procedure is required. This procedure (Fig. 2.1) is essential for making inferences from data. A final, integral step is to test and evaluate the assumptions. This step is especially critical in capture-recapture and removal studies because most estimators are not robust to certain assumptions about capture and recapture probabilities. Tests of model assumptions are computed in program CAPTURE.

Confidence Intervals

An estimate without both a measure of precision (the sampling variance) and an assessment of the relevant assumptions is not trustworthy and must be regarded as scientifically invalid. A single estimate of N is not meaningful without a measure of the sampling variation in the estimator. While the variance, standard error, and coefficient of variation are measures of sampling variation (or precision), the construction of a confidence interval for the parameter of interest represents a much stronger inferential statement. A confidence interval usually is written as

$$P[a \leq N \leq b] = 1 - \alpha \ ,$$

where a and b are the lower and upper bounds calculated from the sample data, N is the parameter of interest, and $1 - \alpha$ is the significance level. The value of α is frequently chosen to be 0.05. The bounds for the interval are constructed from a given formula, depending upon the distributional assumptions made about the estimator. For example, confidence intervals for the mean μ of a normal population are familiar to most biologists (see *Bliss 1967:186-204*). In this example, the bounds for the 95% confidence interval are computed as

$$\bar{y} \ \pm \ 1.96\,\text{se}(\bar{y}) \ ,$$

where se is the standard error. The confidence interval statement implies that if one repeatedly drew a random sample from the population, computed the estimate $\bar{y}$ of μ, and computed the 95% confidence interval, then 95% of the intervals would cover the parameter μ (Fig. 2.7).

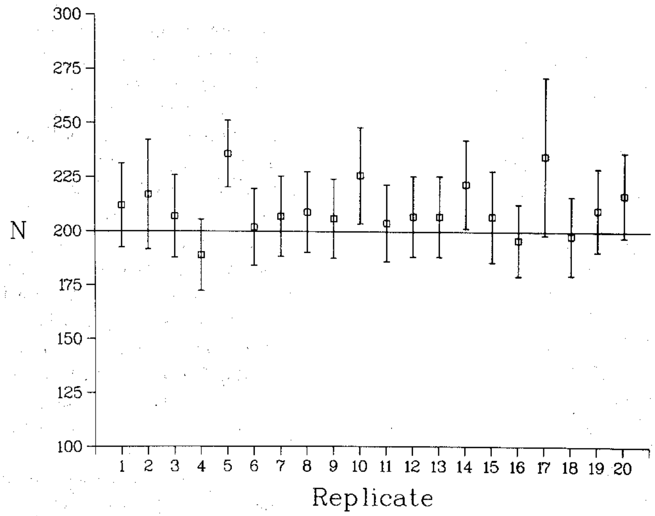


Fig. 2.7. Estimates and 95% confidence intervals plotted for 20 independent surveys. In each simulated case, the true parameter is N = 200. Note that only 17 of the 20 (85%) intervals cover the parameter. The average valueof $\hat{N}$ appears to be biased high. The bias, of course, could explain why the coverage may be a little below the nominal 95% level. In several capture-recapture models, the actual coverage, unfortunately, is substantially less than 95%.

Most methods in the statistical literature, and all methods in this primer, for constructing confidence intervals are based on the assumption that estimators are distributed as normal random variables; it is thus very desirable for an estimator to be distributed normally. Distribution of ML estimators becomes approximately normal as the sample size increases. Unfortunately, significant nonnormality (positive skewness) often occurs with capture-recapture estimators, partly because sample sizes are too small (see *Otis et al. 1978:133-135*).

Thus, in capture-recapture sampling it is often difficult to find a procedure to compute the lower and upper bounds (not always symmetric around the estimate) of the confidence interval so that, in fact, they include the parameter N 95% of the time. In other words, the actual coverage is often less than 95%. Poor coverage also is often due to poor estimates of the sampling variances, to biased estimators, or to nonnormal distribution of the estimators. (Refer to *Otis et al. 1978:126* and *133-135* for examples and discussion of these problems.) In this primer, 95% confidence intervals are constructed as $\hat{N} \pm 1.96 [\hat{se}(\hat{N})]$.

Tests of Hypotheses

A statistical hypothesis is a statement about one or more parameters of the population of interest. A decision concerning the validity of the hypothesis is made based on the value of a test statistic calculated from the sample data. The test statistic frequently has a common distribution, such as chi-square, normal, t, or F (see *Mendenhall and Scheaffer 1973:325-365*). Mathematical statistics is employed to derive the

C. H. N. Jackson

C. H. N. Jackson, D.Sc., made several theoretical contributions to the analysis of capture-recapture data in a series of papers published during the 1930s and early 1940s. His work stemmed from his life-long interest in tsetse flies in the Tanganyika region (now Tanzania). Jackson, an Englishman, consulted with R. A. Fisher on statistical questions during the 1930s.

Jackson proposed several methods based on a variety of assumptions. He gave point estimators and, usually, sampling variance estimators for his methods. In most respects, his work was far advanced for the time.

Jackson was born in 1900 and was awarded Ph.D. and D. Sc. degrees from Cambridge University, the latter for his population studies on tsetse flies. He was awarded the Order of the British Empire not long before he died.

Capture-recapture studies were merely a small part of Jackson's long professional career; his publications cover the period 1927-1955. He was a distinguished entomologist working near Old Shinyanga with the Tsetse Research Center, now the Uganda Trypanosomiasis Research Organization in Tororo, Uganda. Those wishing to gain further insight into this famous entomologist should read the paper by Potts and Jackson (1952); "The Shinyanga game destruction experiment," Bull. Entomol. Res., **43**(2);363-374. (Photograph shows Jackson apparently at middle-age, perhaps in the 1930s or 1940s; courtesy of P. M. Mwambu.)

theoretical distribution of the test statistic if the null hypothesis (H_O) is true. From such distributions we obtain critical values—numbers that are compared to the value of the test statistic to decide whether H_O is rejected. For every significance level (α value) there is a corresponding critical value. Usually α is set at 0.05 or 0.01. However, the experimentor is free to set the significance level of the test at any value, although very rarely does this value exceed 0.10. Thus, in a sense, the user is specifying the chance that a true null hypothesis will be rejected (Type I error). At this point, an interesting tradeoff is made. If the user is willing to increase the chance of making a Type I error from, say, 0.04 to 0.10, then the corresponding chance of accepting a null hypothesis that is not true (Type II error) is decreased. This result of statistical testing theory explains why different significance levels may be used in different testing situations. It is the responsibility of the experimentor to decide which type of error is the more serious in a specific situation.

To test one hypothesis (specifically, the null hypothesis H_O) against another (termed the alternative hypothesis H_A), a study is designed, data are collected and analyzed, and if the results are unlikely under this hypothesis, the null hypothesis is rejected. If the results seem probable under the null hypothesis, there is no reason to reject it. The test statistic measures the degree to which the results conform to the null hypothesis.

For example, we might test the null hypothesis H_O, stating that a penny is fair (that 50% of all tosses will be heads and 50% tails), against the alternative hypothesis H_A, stating that a penny is not fair, by flipping the penny 500 times and observing the outcome. Intuitively, if we observed 50 heads in 500 tosses we would consider that the result was improbable under the null hypothesis and that H_O should be rejected in favor of H_A. We would conclude that the penny is not fair. On the other hand, 248 heads would be considered a likely result, very close to the 250 we would expect, and we would have no reason to reject H_O. In this intuitive example, we have used the "number of heads obtained" as the value of our test statistic.

When results are obvious, the decision to reject or not reject is clear. In actual practice, however, experimental results are not always so clear and intuition may be of little help. Statistical theory then provides objective methods for making inductive decisions and for evaluating the goodness of the inferential procedures. Simply stated, the decision is whether to reject H_O.

A basic philosophy of science is that the truth of a null hypothesis cannot be proven. We can reject a null hypothesis on the basis of data from a proper experiment. If the experiment is replicated several times and each time H_O is clearly rejected, the evidence becomes very convincing that H_O is false. Conversely, if in repeated experiments, properly conducted, we fail to reject H_O, we continue to entertain the possibility that H_O is true. We can never truly "accept" H_O, but repeated failure to reject it adds to its authenticity.

Error Types and Distributions under the Null Hypothesis. In hypothesis testing, two types of errors can be made.

- A Type I error is the rejection of a null hypothesis (H_O) that is true. The probability of a Type I error is denoted as α (the significance level).
- A Type II error is the acceptance of a null hypothesis (H_O) that is false. The probability of a Type II error is denoted as β.

The possible outcomes of hypothesis testing are illustrated in Fig. 2.8. Commonly, 100α (in percent) is referred to as the significance level of the test (for example, 5% and 1% are frequently used).

Nearly all of the relevant tests in *Otis et al. (1978)* and this primer are distributed as chi-square (χ^2) variables if sample size is large. For these tests, H_O is rejected if the test statistic is larger than the critical value. Various chi-square distributions are shown in Fig. 2.9, with rejection (significance) regions. The concept that a test statistic, such as chi-square test, has a distribution is difficult to understand. In many cases, a test statistic follows a chi-square distribution if the null hypothesis H_O is true. If such a test is replicated 5, 25, 50, and 100 times, a strong tendency toward a chi-square distribution is observed (Fig.

Decision	Null Hypothesis H_0	
	True	False
Reject H_0	Type I Error (α)	No Error
Do Not Reject H_0	No Error	Type II Error (β)

Fig. 2.8. These are four possible outcomes of a statistical test of hypotheses and their associated errors.

2.10). Again, the concept of repeated samples is the basis for the theory that indicates distribution of a particular test is chi-square under the null hypothesis.

The test of closure has a test statistic with a standard normal distribution (mean = 0, standard deviation = 1). The test, shown in Fig. 2.11, is one-sided. Rejection of H_0 (closure) is based only on negative values of the test statistic (see *Otis et al. 1978:120*).

The power of the test (in percent), defined as $(1 - \beta)100$, relates to the ability of the test to reject H_0 if it is false. If a test routinely fails to reject a false hypothesis, we say it lacks power. The power of a test

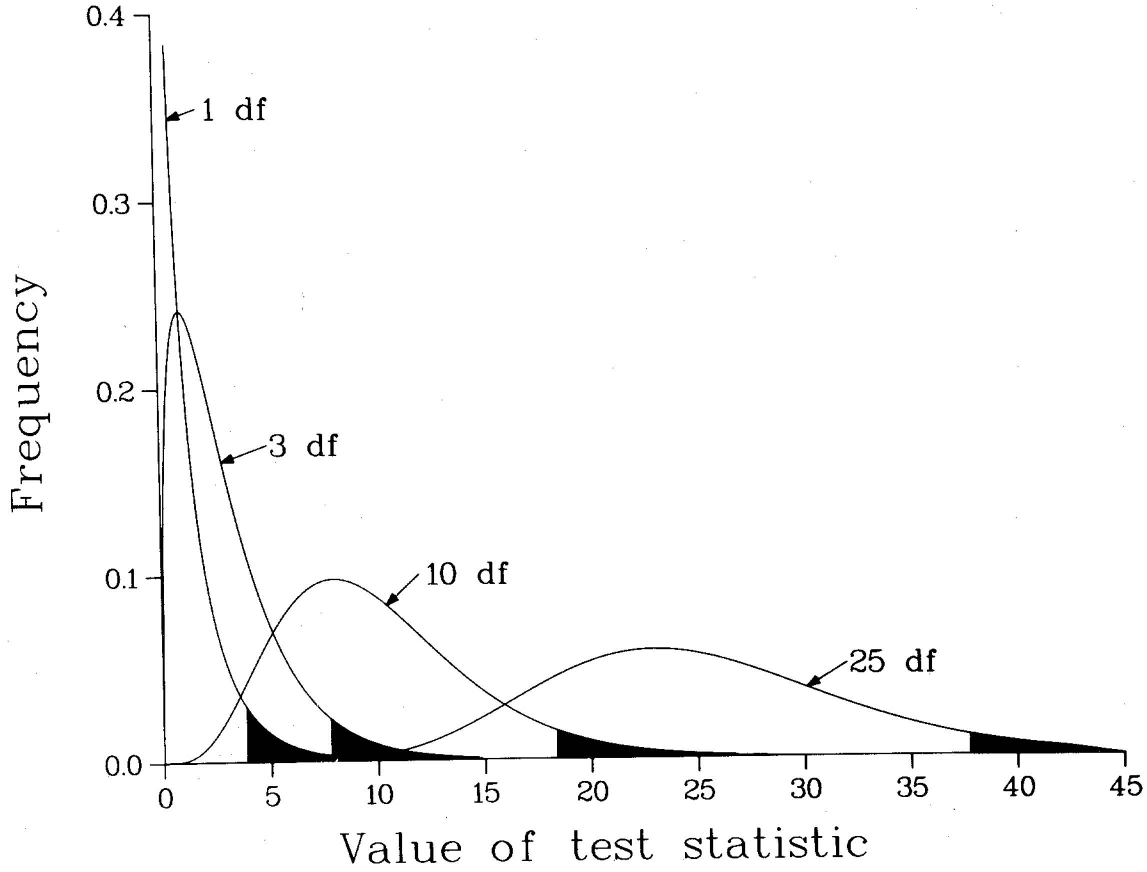


Fig. 2.9. The chi-square distribution for 1, 3, 10, and 25 degrees of freedom (df). In each case, the 0.05 rejection region is shown as a shaded area. All seven test statistics in *Otis et al. (1978:115-119)* are distributed as chi-square under the null hypothesis H_0. The interpretation of a test statistic that is distributed as chi-square is simple. Suppose that a test statistic for a particular H_0 is distributed as chi-square with 25 df [written as $\chi2_{(25)}$]. If the computed value of the test statistic exceeds about 35.5, we will reject H_0 (at the 0.05 significance level). The concept is that a value as large as 35.5 is very unlikely if the test statistic is, in fact, distributed as $\chi2_{(25)}$. We consider it sufficiently unlikely, so we decide to reject H_0. The output from program CAPTURE in *Otis et al. (1978:92)* gives various test statistics that are distributed as chi-square.

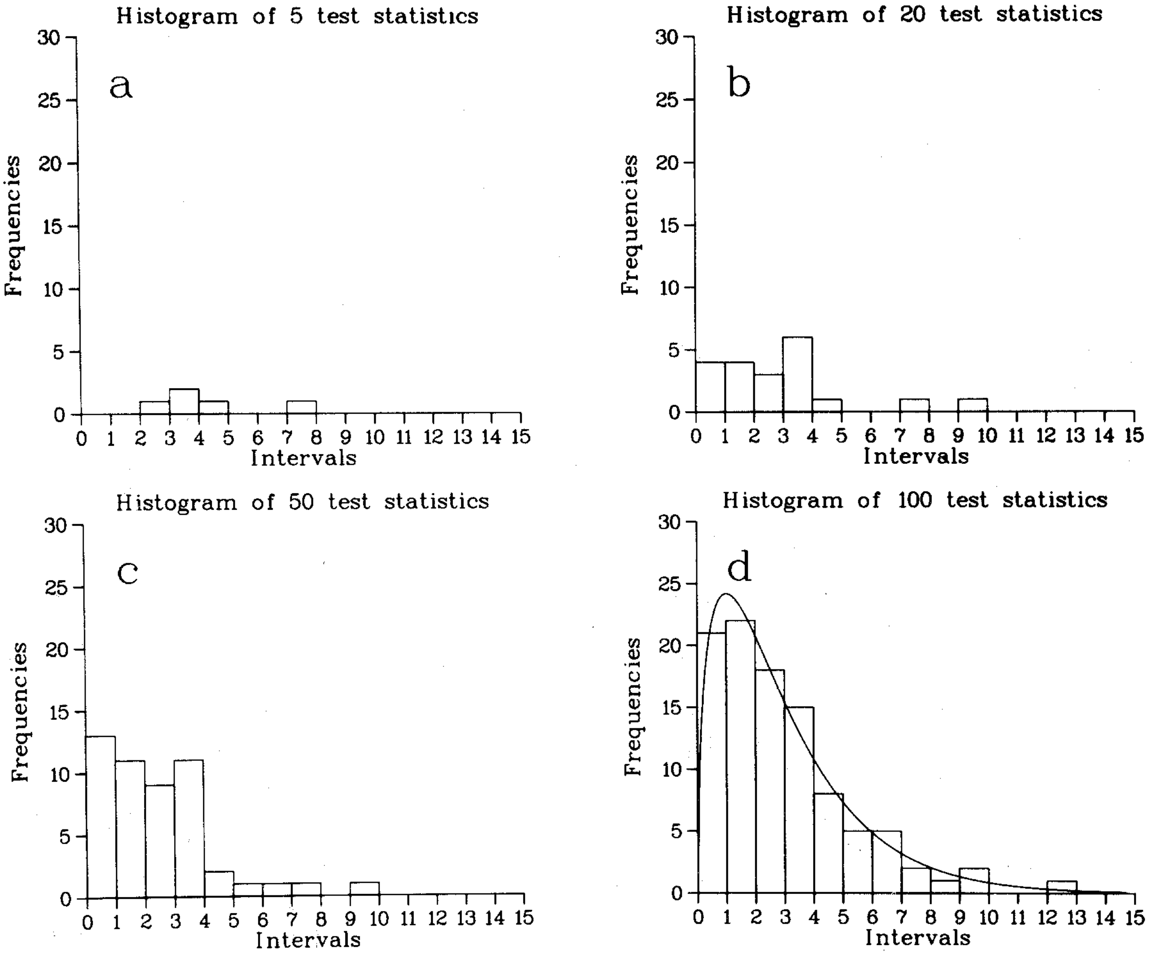


Fig. 2.10. Histograms of test statistics based on 5, 20, 50, and 100 sample data sets. The test statistics are distributed as chi-square under the null hypothesis. In d, the observed value of the test statistic at 12-13 represents a Type I error.

can be computed based on the theory of statistics. For now, only the concept of power is important (Figs. 2.12 and 2.13).

Error Types and Distributions under the Alternative Hypothesis. We have discussed and illustrated the concept of the test statistic distribution under a null hypothesis H_0 and how this distribution actually determines the test statistic values that cause rejection of H_0, for a given significance level (α) of Type I error. A test statistic also has a distribution under the alternative hypothesis H_A. The shape of this distribution determines the power of the test or, equivalently, the size of the Type II error. Figs. 2.12 and 2.13 illustrate this concept.

Hypothesis Testing in Capture-Recapture and Removal Studies. In capture-recapture and removal studies we encounter two basic hypothesis-testing situations. To illustrate them, let us suppose that we are considering two models of a capture-recapture study, Model M_0 and Model M_t. The first test is for "goodness of fit." An example of this test is

H_0: Model M_t fits the data

H_A: Model M_t does not fit the data.

The question asked here is whether Model M_t is an adequate representation of the data.

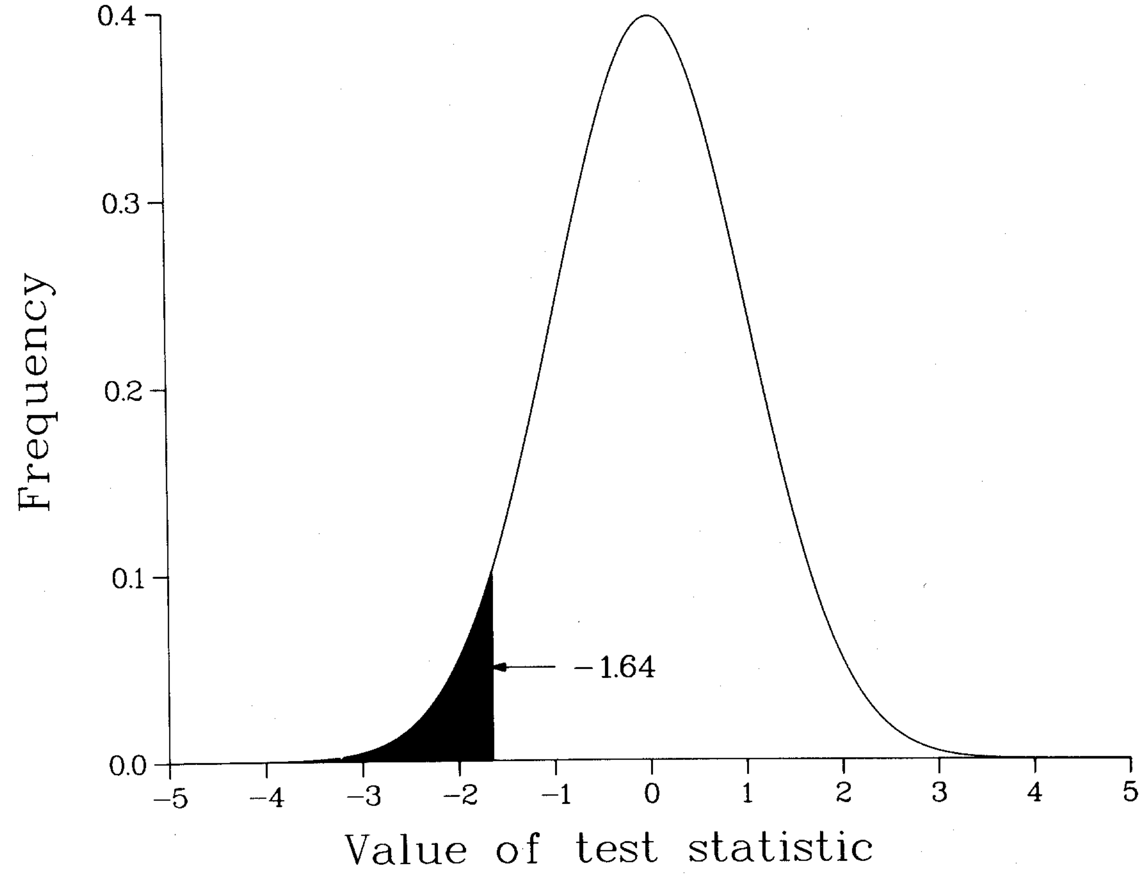


Fig. 2.11. Test statistics frequently are distributed as a standard normal distribution with a mean of 0 and a standard deviation of 1. Such a standard normal distribution is shown, with a one-tailed rejection region for $\alpha = 0.05$ (shaded). The closure test (Otis et al. 1978:120-121) has this distribution and a one-sided (negative) rejection region. If the test statistic under H_O (closure) is less than -1.64, H_0 is rejected at the $\alpha = 0.05$ level.

The second test might be termed a "simple alternative" test. An example of this test is

H_O: M_o fits the data as well as M_t

H_A: M_t fits significantly better than M_o.

In the simple alternative tests discussed in this primer, the model under H_A is more general than the model under H_O. This test is a comparison of the two models. Thus, the question asked here is, Does the more general model (Model M_t) fit the data significantly better than the simpler model (Model M_o), or does the simpler model do as well?

The fundamental difference between these two tests is that the first is concerned with the question of whether one specific model provides a good fit to the data, whereas the second compares a specific model to a more general model to see which provides the better fit to the data. *Begon (1979:55-75)* presents a section on testing assumptions in capture-recapture models that is easy to understand.

Capture-recapture and removal studies represent not only very difficult testing problems, but also difficult modeling and estimation problems. The many technical reasons for these difficulties are beyond the scope of this primer. We will, however, mention four causes of testing problems.

(1) With small samples, the distribution of a test statistic may not follow the theoretical (large-sample) distribution very well (*Fig. 2.14).

(2) The test may have poor power and thus may make rejecting the null hypothesis difficult when, in fact, H_O is false (*Fig. 2.15).

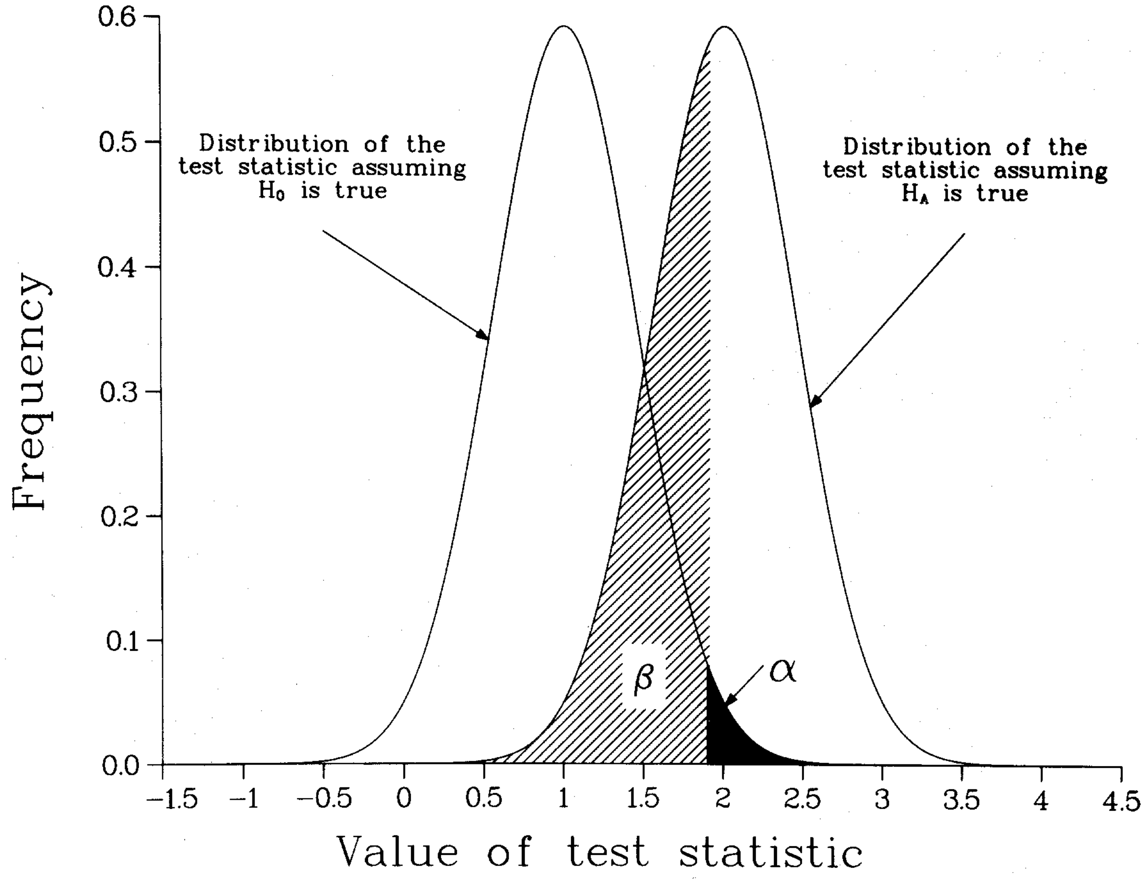


Fig. 2.12. An example of the distribution of a test statistic under the null hypothesis H_0 and under the alternative hypothesis H_A. The probability of a Type I error, denoted as α, is black; the probability of a Type II error, denoted as β, is crosshatched. In this example, $\alpha = 0.05$ and $\beta = 0.45$; the power of the test of the null hypothesis is $1 - \beta$ and is considered to be fairly poor in this example.

(3) The battery of tests to be described (see *Otis et al. 1978:115-120*) are very dependent on each other. Dependence, which is to be expected because all the tests are computed with roughly the same data, makes interpretation difficult. (The difficulty is alleviated, in part, by thc automatic model selection algorithm in CAPTURE.)

(4) Certain tests cannot be computed unless substantial amounts of data are available for analysis; *Leslie's (1958)* test is an example.

In dealing with capture-recapture and removal methods, the user of this primer need not be concerned with the derivation of the test or how the test statistic is distributed. Program CAPTURE computes the value of the required test statistics and the observed significance levels. Biologists, therefore, need only to interpret the results.

Simulation Methods

Since the late 1960s, computer simulation has been used to study the performance of various estimators of population size from capture studies. Simulated populations provide a set of essential features: (1) the primary parameter N is known exactly; (2) the capture probabilities are known and can be manipulated at the will of the investigator; (3) the assumptions can be deliberately met or violated; (4)

Patrick H. Leslie

Patrick Leslie, "George" to many friends and colleagues, is probably best known (in the context of the subject here) for his work in estimating population size and death rates in populations of small mammals. During the late 1940s and early 1950s he collaborated on a series of papers about voles with Dennis and Helen Chitty while working in the Bureau of Animal Population with Charles Elton. Before this work, he had developed several regression-based methods. His most important work dealt with the population projection matrix methods—the "Leslie matrix."

Leslie took his undergraduate training in physiology in 1921 and earned the Doctor of Science degree at Oxford. He joined the Bureau of Animal Population Research at Oxford in 1935 and continued there as Senior Research Officer until his retirement in 1967. Incredibly, in view of his accomplishments, Leslie had no formal training in advanced mathematics; his talent for applying mathematical approaches to ecological problems did not become apparent until after he was 35 years old. Leslie's career is reminiscent of that of Sir Ronald Fisher, in that they both maintained contact with the real problems of colleagues in other fields and loved to explore real data. Further information can be found in *Nature* 239(5373):477-478, in an article written after his death. (Photograph taken in 1949 by D. A. Kempson, Bureau of Animal Population, Oxford University.)

the proper stochastic sampling variation can be emulated; and (5) the simulated study can be repeated exactly, if necessary (*Bishop and Sheppard 1973*).

Simulated populations are very useful in answering a host of questions concerning the small-sample properties of an estimator under data from its model (bias), or under other models (robustness); the confidence interval coverage; the power of hypothesis tests; and other important issues. Simulated population data are inexpensive to generate and useful for many purposes. For example, Tables 17-19 of *Otis et al (1978:60-62)* involve the analyses of 2400 simulated data sets.

We use simulation to generate a data matrix X with known properties. Let us start by considering the first animal on the first trapping occasion. How can we determine if it is to be captured or not? We begin by looking at the capture probability (a parameter) for this animal. Let us say that this is 0.3; that is, $p = 0.3$. Of course, an animal is either captured or not—it cannot be 0.3 caught. In addition, the data must be simulated to preserve the chance (or stochastic) element in sampling studies. Therefore a uniform random number between 0 and 1 is generated by the computer. This number is compared to 0.3. If the random number is less than 0.3, the animal is "captured" and the first element of the X matrix is set to one, indicating the animal was captured. The sample procedure is performed for animal 2, and so on. Input values for the capture probabilities p_{ij} are available within the computer. For example, if animals are assumed to be trap happy, then animal 1 will have $p > 0.3$ on subsequent trapping occasions.

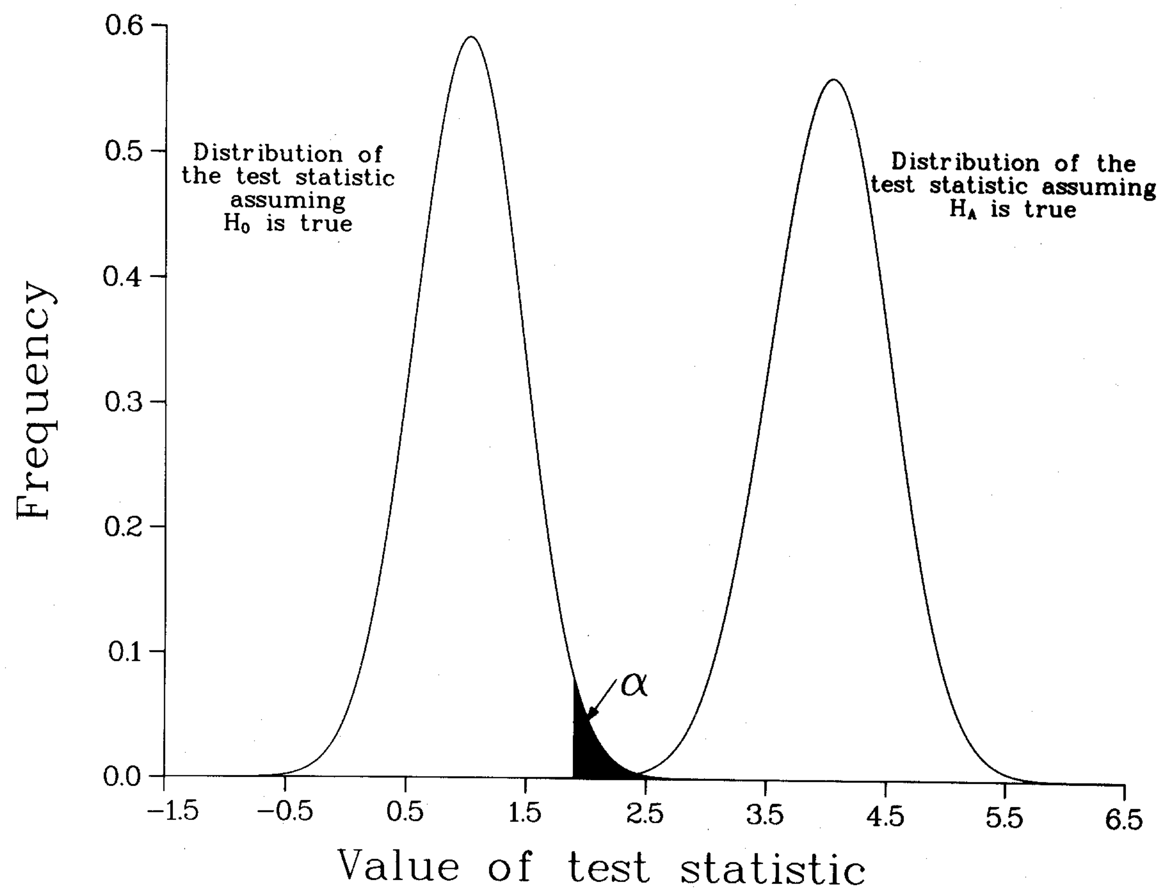


Fig. 2.13. A second example of the distribution of a test statistic under the null hypothesis H_0 and the alternative hypothesis H_A. The probability (α) of a Type I error is shown in black; the probability (β) of a Type II error is essentially zero. The power of the test $1 - \beta$ is nearly 1, indicating a very good and powerful test of H_0. (Compare with Fig. 2.12.)

Summary

1. Sampling a population enables valid inferences to be made about various parameters if proper procedures are used in the field and during the analysis.
2. A mathematical model is required to link the sample data with the necessary assumptions and to provide a basis for parameter estimation. Stochastic models are needed because the sample data result from processes with a strong random component.
3. Estimators for capture-recapture and removal studies should be unbiased and precise. Proper consideration of basic principles enables estimators to have these properties. Estimates of population parameters must have a measure of precision to be of value in making valid inductive inference.
4. Variation is everywhere in capture-recapture and removal experiments. The two types of variation (spatial and temporal, and stochastic) must be recognized in biological work.
5. Adequate sample size and the magnitude of the capture probabilities are critical elements to consider in the design of a study.
6. Random sampling is inappropriate to most capture-type studies, and the methods discussed here make no use of this assumption.
7. Only the estimator for Model M_h can be computed easily by hand. The comprehensive computer program CAPTURE is required for essentially all of the analyses described here.
8. Tests of hypothesis are important to assess the tentative assumptions and to select the best model.

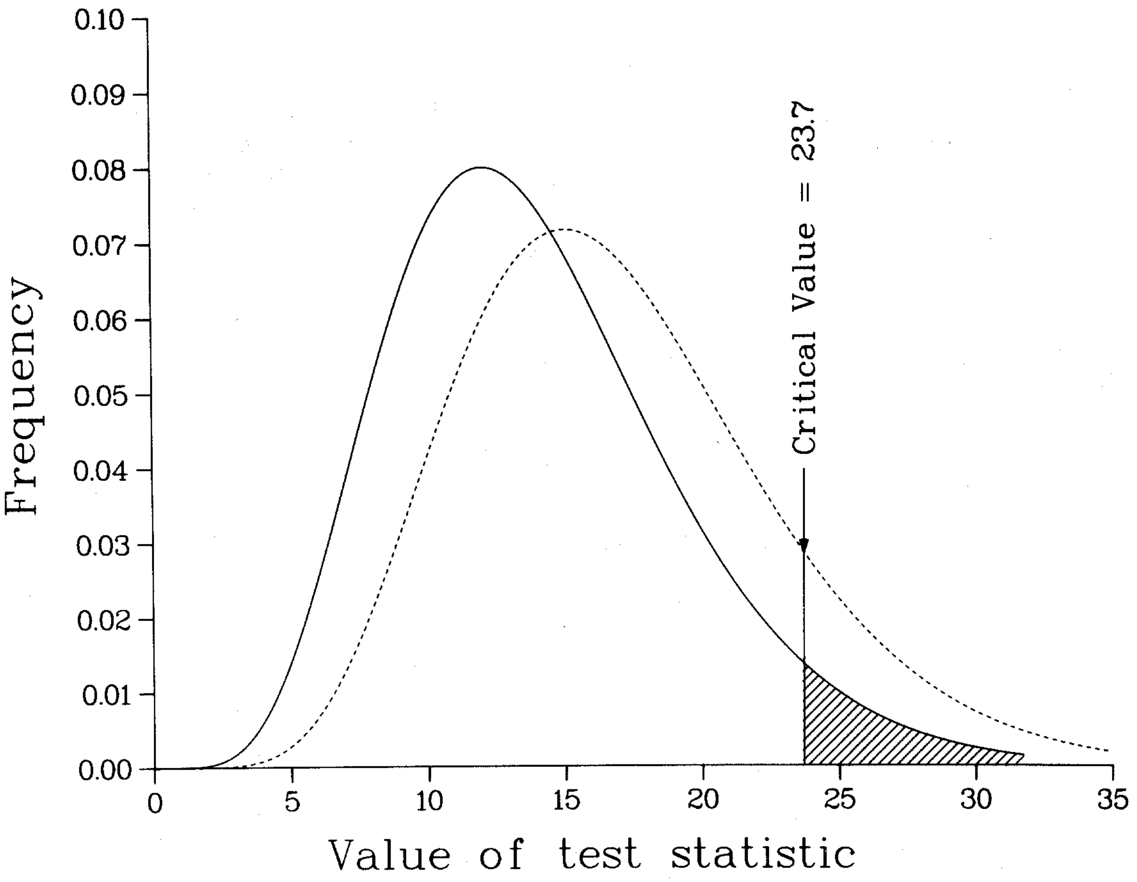


*Fig. 2.14. The fact that a test statistic may not follow the theoretical distribution if sample size is small can present problems. Shown here is the distribution of a test statistic that asymptotically follows a chi-square distribution with 14 df (solid line). However, with the sample sizes encountered in this series of experiments, the distribution was approximated poorly by the chi-square distribution. The rejection region for $\alpha = 0.05$ for the chi-square distribution is shown in the crosshatched area. Note, however, that the critical value 23.7 defines a significance level on the actual sample distribution that is much larger than 0.05. Because the sample distribution is not well approximated by the chi-square distribution, we would reject the null hypothesis in this example when it is, in fact, true more often than 5% of the time (because more than 0.05 of the area of the actual distribution is to the right of 23.7).

Questions and Exercises

1. Are simple formulas available that enable biologists to compute ML estimates of N for most models?
2. Why is a model needed to estimate parameters from data?
3. If an *ad hoc* method provides $\hat{N}$ in close agreement with an estimate from a method with a rigorous underlying theoretical basis, does this agreement provide substantial support for the *ad hoc* method?
4. What might be a reasonable coefficient of variation for $\hat{N}$ for research studies? For management-oriented studies?
5. Compute the average, say $\bar{N}$, its standard error $\hat{se}(\bar{N})$, and its coefficient of variation (cv) for each study below. If N = 20, which estimate is precise, which is biased, and which is both precise and biased or neither? (See Figs. 2.4 and 2.5.)

Study 1	20	21	17	19	22
Study 2	25	26	22	24	27
Study 3	5	8	24	30	33
Study 4	25	38	46	50	57

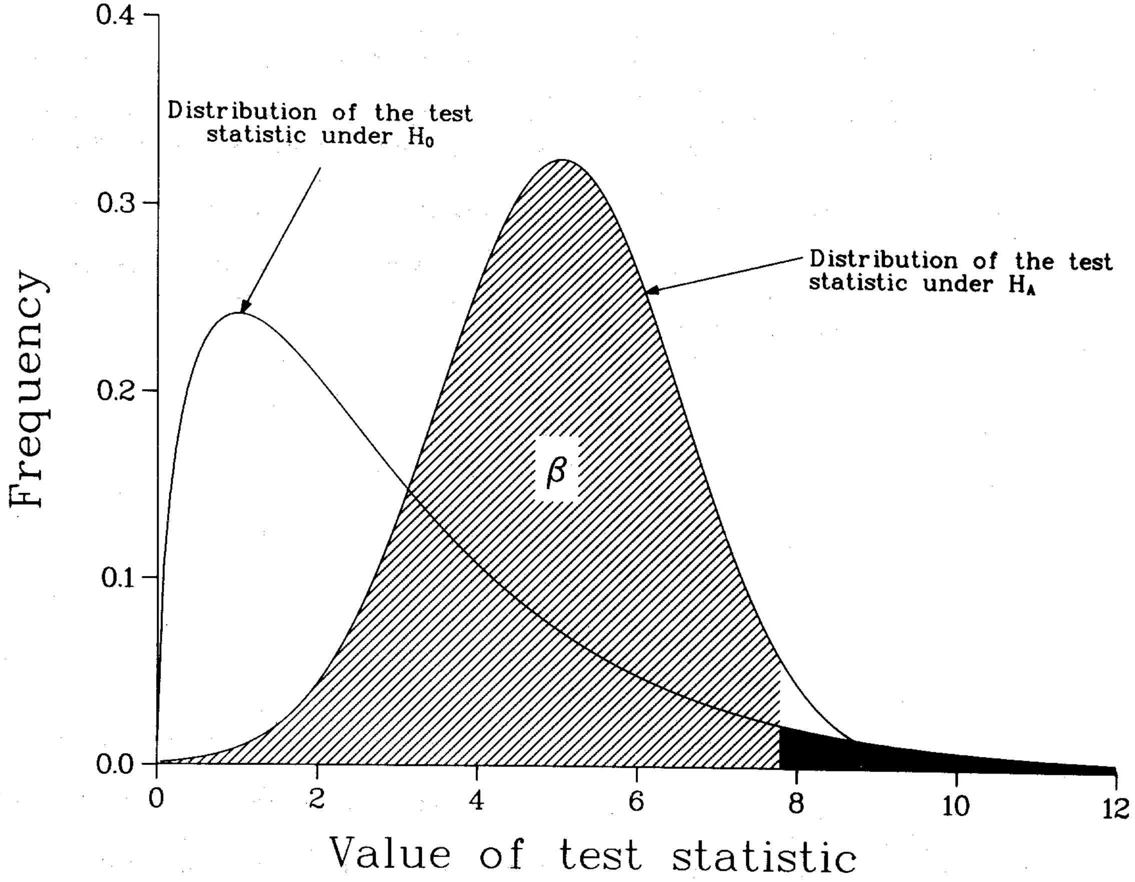


*Fig. 2.15. Distribution of a test statistic with very poor power (power = 1 − β). The very large β region, which is the probability of accepting H_0 given H_A is true, is shown in the crosshatched area.

6. Consider T_1 and T_2, two test statistics of a specific hypothesis; T_1 has power 0.13 and T_2 has power 0.89. Which would you prefer?
7. Name some statistical distributions that test statistics commonly follow, if sample size is large.
8. What are the null and alternative hypotheses for a goodness of fit test?
9. What are the two ways in which a hypothesis test can fail to give the "correct" result?
10. Based on a large sample, you compute an estimate of a parameter θ, as $\hat{\theta} = 141$, with $\hat{se}(\hat{\theta}) = 13.1$.
    a. What is the 95% confidence interval?
    b. Is a true value of $\theta = 95$ a reasonable value?
    c. Similarly, is $\theta = 135$ plausible?
11. The answers to 10b and 10c were somewhat intuitive.
    a. Can formal hypotheses be formed for 10b and 10c?
    b. What is the form of the test?
    c. Compute and interpret the test statistics.
12. A colleague shows you an estimate of population size for snails in geographic areas A and B. In area A, $\hat{N} = 4306$, and in area B, $\hat{N} = 3911$. What can be inferred from these estimates?
13. You work for an agency and your supervisor tells you of the agency's concern for the Wabo tributary of the huge Lake Powell. The concern is over the possible reduction in the number of spawning lake bass in this area caused by oil drilling and exploration in this area of the lake. You are told to "find two technicians and get out there and find out what we need to know." The following questions should occur to you. How would you answer them?
    a. What is the population of interest?
    b. Is a sample or a census called for?
    c. What is the parameter of interest?

d. What sampling methods might be useful if a sample is required?

e. Suppose a good estimate of population size before exploration and drilling is available. Formulate a formal null and alternative hypothesis of interest.

14. You see in the literature that a certain parameter estimator had a 95% confidence interval of (31 to 91) for a given sample. Does the true parameter lie within this specific interval with probability 0.95? Why?

15. If you tested a null hypothesis and made a Type I error, what would you conclude? Is your conclusion correct?

16. Consider the results of a 5-year study of mice in an old field in Wisconsin. Grid trapping was done with live traps for 7 days and the data for each year were analyzed carefully. The estimates of population size ($\hat{N}$) appear below, along with the true parameters (N).

Year	$\hat{N}$ (±se)	N
1	115 (±15)	100
2	170 (±30)	150
3	150 (±33)	200
4	256 (±40)	225
5	42 (±8)	50

a. What is the cv for each estimate?

b. Are the individual estimates fairly good?

c. Can good inferences be made from the five estimates about the actual population changes over the 5 years?

17. Why have exact ML estimators for many of the capture-recapture and removal models not appeared in the literature until recently?

18. Is it necessary to know the details concerning likelihood functions and estimation theory before using some of the analysis methods presented here and in *Otis et al. (1978)*?

19. Give two or three reasons why a stated 95% confidence interval may cover the true parameter less than 95% of the time.

20. You have defined a null hypothesis, collected appropriate data, computed a proper test statistic, and found the observed significance level is 0.007. What can you conclude? Why?

21. If var($\hat{N}$) = 625, what is se($\hat{N}$)?

22. Examine Table N.3.6 in *Otis et al. (1978:127)*. Is the estimator for Model M_t robust to trap-happy and trap-shy populations?