Nov 73

STATISTICAL CONCEPTS AND TECHNIQUES IN POPULATION ESTIMATION USING THE PETERSEN METHOD[1/]

BURWELL GOOCH[2/]

ABSTRACT

Statistical concepts that are meaningful to mark-recapture population estimation studies are defined and discussed. The relevance of probability sampling is explored. It is proposed that establishment of uniform ratios of marked to unmarked fish throughout the population eliminates the requirements of probability sampling. Two unusual characteristics of the estimator of population size involving bias and variance are investigated, and a means of circumventing their consequences suggested. Statistical tests are described for testing two different hypotheses involving association of marked and unmarked fish. A method of estimating loss of marked fish due to migration is presented. Examples illustrating the concepts and techniques are provided.

Many fisheries biologists are inadequately prepared for certain basic tasks required of them in the practical world of resource management. A typical example involves the practice of population estimation by use of the Petersen method of mark and recapture. In Montana this is one of the most important tasks in which biologists are currently engaged. Unfortunately, few of the statistical concepts relevant to this practice, including such elementary and ubiquitous notions as

---

[1/] This work was performed as part of Federal Aid in Fish Restoration Project Montana F-4-R-23 Job I-b.

[2/] Data Processing Bureau, Montana Department of Administration, Helena, Montana 59601.

accuracy, precision, and bias, are adequately understood. Part of the reason for this situation is that, in contrast to the spate of papers devoted to mathematical development of mark-recapture methods, which for the most part are of no practical value to the average biologist, very little is available that bridges the philosophic gap between elementary statistical concepts and some of the fundamental but cryptic characteristics and assumptions underlying these methods.

The notes that follow were developed in an attempt to at least partially fill this void. Thus, the primary concern of this paper is to supplement, rather than paraphrase, the existing literature. Obviously, some of the information that appears here did not originate with the author. To the author's knowledge, however, most of the material, including the synthesis as a whole, has not appeared in print elsewhere.

Preparation on the part of the reader is assumed to consist of one or two courses in statistics and familiarity with the basic notion and techniques of the Petersen method.

DEFINITION OF TERMS

Writing a paper of this nature is much like building a house. It must begin by laying a solid foundation that serves to support the main body of the structure. In this case, the foundation consists of statistical concepts that are basic to the rest of the paper. If the reader does not fully understand the meaning of these concepts, then he can expect to gain only a superficial appreciation of the main results presented later on. Although most of the terms defined here should be familiar to readers, at least by name, explicit and precise definitions of some are difficult to find in elementary statistical texts.

ESTIMATOR: Mathematical formula used for calculating an estimate of some population attribute.

Example: $\hat{N} = (M+1)(C+1)/(R+1) - 1$

ESTIMATE: A number calcluated on the basis of sample data substituted in an estimator.

Example: $510 = (145+1)(195+1)/(55+1) - 1$

SAMPLING VARIABILITY (sampling error): Variability among sample estimates arising from variability among the population units (or individuals), different combinations of which are included in different samples.

EXPECTATION OF ESTIMATE: The average value of all possible sample estimates that can be generated in a given sampling situation. That is,

$$E(\hat{N}) = \Sigma \hat{N} P(\hat{N})$$

where $\hat{N}$ = estimate of population attribute N

$E(\hat{N})$ = expectation (expected value) of $\hat{N}$

$P(\hat{N})$ = probability of occurrence of $\hat{N}$

Σ = summation over all possible values of $\hat{N}$.

BIAS: Difference between the expected value of an estimate and the value of the population attribute being estimated. That is,

$$B = E(\hat{N}) - N$$

where B = bias in $\hat{N}$.

Although B may be positive, negative, or zero, it normally is thought of as being constant in any given sampling situation. It also usually arises from a variety of sources, only some of which may be identifiable, but each of which acts in its own way to either inflate or deflate an estimate. If B is zero, $\hat{N}$ is said to be an unbiased estimate of N.

PRECISION: A measure of the variability of sample estimates in terms of their deviations from expectation.

ACCURACY: A measure of the variability of sample estimates in terms of their deviations from the population attribute being estimated.

VARIANCE OF ESTIMATE: A common measure of precision. It is the expected value of the squared deviations of estimates from their expectation. That is,

$$V(\hat{N}) = E(\hat{N} - E(\hat{N}))^2.$$

MEAN SQUARE ERROR: A common measure of accuracy. It is the expected value of the squared deviations of estimates from the population attribute being estimated. That is,

$$MSE(\hat{N}) = E(\hat{N} - N)^2.$$

PROBABILITY SAMPLING: The foundation of statistical sampling theory. In the context of any given sampling situation, probability sampling implies the following conditions:

(1) All possible unique samples can be defined (at least conceptually) in terms of the population units included in each; that is, each individual in the population is uniquely identifiable in some way (e.g., by name, number, mark, etc), and the composition of each possible sample can be predetermined by listing all included individuals.

(2) The probability of selection of any given sample is known.

(3) The estimation procedure provides a single estimate for each sample.

Conditions (1) and (2) imply the existence of some method of selecting population units such that mathematical deduction of sample composition and probability of selection is permitted.

The significance of the three conditions listed is that they permit us to generate the probability distribution associated with the set of all possible estimates, from which flows the derivation of all characteristics of these estimates.

SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT (random sampling): A method of probability sampling with the following defining features:

(1) Sample size (n) is determined before sample selection begins.

(2) The sample is selected one unit at a time without replacing selected units.

(3) On the i^{th} draw $(i = 1, \ldots, n)$, each unit remaining in the population has the same probability of selection.

In terms of the conditions required for probability sampling, these three properties are equivalent to saying that:

(1) The total number of different samples is

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}$$

(2) Considering all samples together, each population unit appears the same number of times, viz., $\binom{N-1}{n-1}$, and occurs with every other unit in the population the same number of times, viz., $\binom{N-2}{n-2}$.

(3) Each of the $\binom{N}{n}$ different samples has the same probability of selection, viz., $1/\binom{N}{n}$.

We will now investigate some important relationships among the foregoing concepts. One objective in doing this is to dispel two rather widespread myths, viz., (1) biased estimates differ from the population attribute by the amount of the bias; and, (2) unbiasedness is always preferable to biasedness.

To begin, we may note that an individual estimate is the sum of three quantities: the population attribute being estimated, any bias inherent in the sampling and estimation procedure, and a deviation from expectation. That is,

$$\hat{N} = N + B + d \tag{1}$$

where $d = \hat{N} - E(\hat{N})$

Since the expectation of a sum equals the sum of expectations, and the expectation of a constant is the constant, it follows that

$$\begin{aligned} E(d) &= E(\hat{N} - E(\hat{N})) \\ &= E(\hat{N}) - E(E(\hat{N})) \\ &= E(\hat{N}) - E(\hat{N}) \\ &= 0, \end{aligned}$$

and thus that

$$\begin{aligned} E(\hat{N}) &= E(N + B + d) \\ &= E(N) + E(B) + E(d) \\ &= N + B \end{aligned} \tag{2}$$

Hence, the definition of bias. Furthermore, expression (1) demonstrates that bias <u>and</u> sampling variability simultaneously influence the value of an individual estimate. Thus, since B is constant for all $\hat{N}$, whereas d may be positive, negative, or zero in any given sample, the effect of bias may be compounded, negated, or unaffected. It is only the <u>expected value</u> of the estimate, in general, that differs from the population attribute by the amount of bias.

Now,

$$\begin{aligned} MSE(\hat{N}) &= E(\hat{N} - N)^2 \\ &= E((\hat{N} - E(\hat{N})) + (E(\hat{N}) - N))^2 \\ &= E(\hat{N} - E(\hat{N}))^2 + 2E(\hat{N} - E(\hat{N}))(E(\hat{N}) - N) + E(E(\hat{N}) - N)^2 \\ &= V(\hat{N}) + 2(0)(B) + E(B)^2 \\ &= V(\hat{N}) + B^2 \end{aligned} \tag{3}$$

In words, the mean square error of the estimate equals the variance of the estimate plus the square of the bias. If bias is zero, the mean square error and variance are equal. Since precision is a generalization of the term variance (i.e., refers to variability around expectation), whereas accuracy is a generalization of the term mean square error (variability around the population attribute), we may conclude that accuracy is a more inclusive concept because it encompasses both precision and bias. If bias is zero in a particular sampling situation, then accuracy and precision of estimates are synonymous.

Finally, assume that in a given sampling situation, we have available two estimators, $\hat{N}_1$ and $\hat{N}_2$, of N, where $E(\hat{N}_1) = N$, and $E(\hat{N}_2) = N + B$. To many investigators, $\hat{N}_1$ would appear to be the preferred estimator because it is unbiased. The important consideration often overlooked is the relative sizes of $V(\hat{N}_1)$ and $MSE(\hat{N}_2)$. If the latter is smaller than the former, then $\hat{N}_2$ is actually the preferred estimator because it is more accurate (which implies in this situation that it is also more precise).

PROBABILITY SAMPLING AND THE PETERSEN METHOD

In order for a sampling and estimation procedure to be free from bias, three basic requirements must be met. First, sample selection must be based on some kind of probability sampling. Second, the estimator must be unbiased. Third, data must be recorded accurately.

The ability to meet these three requirements in any given sampling situation depends considerably on the nature of the population, the attribute of interest, and the available sampling techniques. As an example, we may consider the problem of estimating the size of a fish population, in particular the "open" type characteristic of streams, using the Petersen mark-recapture method as described by Vincent (1971, 1974).

Under these conditions, by far the most critical and difficult requirement to satisfy is that involving probability sampling, in particular those features relating to the definition and probability of selection of the individual samples. From previous discussion, we know that our ability to determine the composition and probability of selection of individual samples depends on the existence of a suitably defined sampling procedure. Normally, this procedure is such that either (1) the sampler randomizes the order in which sample units are selected (e.g., by shuffling a deck of cards, spinning a basket of numbered balls, or generating a sequence of random numbers); or (2) the population units present themselves for selection in random order, without any interference from the sampler (e.g., red blood cells circulating in the bloodstream, or bacteria in a water supply). Implicit in the latter alternative is that population units behave more or less as gas molecules, randomly distributing themselves among their physical confines independently of each other.

In population estimation situations, these two notions are often expressed as requirements that either (1) the fishing effort be randomly distributed with respect to the population units; or, (2) marked and unmarked members of the population be randomly distributed with respect to each other.

In regard to the former alternative, usually the only practical way of attempting this with electric shocking equipment is on a geographic basis. Unfortunately, such an approach is not really relevant because it results in cluster-type sampling, whereas what is required is a method that is equivalent to simple random sampling. Thus, if this procedure is used to overcome the problem of nonrandom association of marked and unmarked fish, resulting estimates will be biased and have larger than otherwise variance.

In regard to the latter alternative, it is well known that fish do not behave as gas molecules, but rather exhibit a significant amount of territoriality and homing behavior. Furthermore, because of the nature of their habitat, it is more realistic to consider many populations of fish as consisting of somewhat discrete subpopulations. This is particularly true of our Montana stream populations.

Consequently, we may conclude that the conventional assumptions regarding fishing effort and distribution of marked fish are inapplicable to the typical population estimation investigation, at least in Montana.

There is, however, a third alternative that may be more easily met: wherever subpopulations of fish occur, the ratio of the marked number to the unmarked number be the same, viz., M/(N-M), where N is the total population size, and M is the total number of marked members. Although this is precisely the objective of random association of marked and unmarked fish, it is possible in some circumstances to achieve the same result without such an unrealistic constraint.

For example, in flowing waters that do not exceed the capture depth of fish shocking equipment, it should be physically possible, during the mark trip, to cover the entire extent of the study section, marking and releasing at many points throughout. If the stream is too wide or swift for one crew to cover in one pass through the section, then the crew should make several passes (or several crews should make simultaneous passes), until the investigator has satisfied himself that every group of fish in the section has been sampled for marking. The result of this procedure is to maximize the chances of obtaining uniform ratios of marked to unmarked fish throughout the section. In effect, the investigator does not depend on random mixing for the establishment of uniform ratios, but instead ensures this through his own efforts.

One particularly valuable feature of this third alternative is that there is a relatively simple way of evaluating its success in any given situation (as discussed in the section on Tests of Hypotheses). Thus, the necessary elements are available to help ensure the reliability of individual estimation studies, if such is possible, and consequently to provide a useful criterion for determining the applicability of the method in different situations.

A well-known source of error in the Petersen method that is not overcome by this approach is different catch rates of fish related to size. In general, electric shocking equipment is more efficient in catching larger fish than smaller fish. Thus, even under otherwise ideal conditions, if the proportion of small fish varied from one subpopulation to another (as we would expect), the ratios of marked to unmarked fish could not be uniform. For fish in a size range within which efficiency of capture is relatively uniform, however, we would expect these ratios to be similar.

Another source of error that can cause problems is catch rates variable by habitat type. For example, brush-lined, undercut banks are more difficult to fish than open riffle areas. As a result, more effort must be expended marking fish in the former areas than in the latter in order to achieve the same proportions of marked fish throughout.

BIAS IN PETERSEN ESTIMATES

As noted previously, one of the three requirements for unbiased estimates (of any kind) is an unbiased estimator. In this section, we will examine two common sources of estimator bias in the Petersen method of population estimation.

Although it is more realistic to reject the assumption of probability sampling in the context of most population estimation studies, unfortunately this precludes any mathematical analysis of the method. Thus, in order to make the problem at all tractable, mathematicians responsible for the development of Petersen estimation theory have assumed, for the most part, that recapture sampling is equivalent to simple random sampling without replacement. At least three different estimators have thus been derived, viz.,

$$\hat{N} = MC/R \text{ }^{1/}$$

$$\hat{N} = M(C+1)/(R+1) \text{ }^{2/}$$

$$\hat{N} = (M+1)(C+1)/(R+1) - 1 \text{ }^{3/} \qquad (4)$$

where M = number marked at time of mark run

C = number caught at time of recapture run

R = number of recaptures in C

These are similar in most respects, although only expression (4) is potentially unbiased. However, this estimator is characterized by three undesirable features, two of which are rather unique. First, its unbiasedness is a function of sample size. Second, over a partial range of sample sizes, both the variance and mean square error of estimates increase with sample size. Third, the form of the variance can only be approximated. Those features concerning the variance are treated in more detail in the following section.

Under simple random sampling, M and C in the formulas shown are assumed to be known constants, and R is a random variable. In any real situation, however, all three quantities are likely to be random variables. In fact, M is variable in two senses. First, although the number to be selected for marking may (should) be determined in advance, the number actually selected rarely is the same as the predetermined value. Second, and much more important, the number of marked fish remaining in the population at the time of the recapture run normally is not the same, for a variety of reasons, as the number marked. The problem in this latter case is that our estimator becomes biased because it contains a value for M (number originally marked) that is no longer relevant to the situation existing at

---

1/ The conventional Petersen formula (Ricker 1958)

2/ Modification proposed by Bailey (1951)

3/ Modification proposed by Chapman (1951) recapture time.

The two sources of estimator bias, then, that we will investigate in this section are sample size and change in number of marked fish between mark time and recapture time. Since changes in total population size between mark time and recapture time can also be expected in the normal course of events, this generalization is also included in our investigation.

In the derivation of the results, all assumptions required for unbiasted estimates, except those relating to the estimator, are made. In particular, these include recapture sampling equivalent to simple random sampling without replacement. The applicability of the derivations in any given situation depends heavily on the validity of these assumptions.

We begin by introducing the following notation:

N_m = population size at time of mark run

M_m = number marked at time of mark run

N_r = population size at time of recapture run

M_r = number of M_m remaining at time of recapture run

C = number caught at time of recapture run (sample size)

R = number of recaptures

$\hat{N} = (N_m+1)(C+1)/(R+1) - 1$

$P(R=0)$ = probability that R equals zero, given N_r, M_r, and C

Clearly, if we assume an open population, there is no constraint on N_r, i.e., it may be greater than, equal to, or less than N_m. On the other hand, M_r can never be greater than M_m.

Under the conditions listed, it can be shown that the expected value of the estimate $\hat{N}$ is

$$E(\hat{N}) = (N_r+1)(M_m+1)/(M_r+1) - 1 - (N_r-M_r-C)P(R=0)(M_m+1)/(M_r+1) \tag{5}$$

for $M_r+C \leq N_r$, and

$$E(\hat{N}) = (N_r+1)(M_m+1)/(M_r+1) - 1 \tag{6}$$

for $M_r+C \geq N_r$.

A feature of these expressions worth noting is that they explicitly show that expectation is in terms of N_r rather than N_m. In other words, the population being estimated at recapture time is the population extant at that time. The population existing at mark time (N_m) has no relevance to the problem, except in the special situation described later.

If $M_m = M_r = M$, and $N_m = N_r = N$, expression (5) reduces to

$$E(\hat{N}) = N - (N-M-C)P(R=0) \tag{7}$$

This is equivalent to the formula derived by Robson and Regier (1964, page 216), viz.,

$$E(\hat{N}) = N - (N-M)h(0;C,M+1,N)$$

Using expressions (5) and (6) and the definition of bias, we find that

$$B = (N_r+1)(M_m-M_r)/(M_r+1) - (N_r-M_r-C)P(R=0)(M_m+1)/(M_r+1) \tag{8}$$

for $M_r+C \leq N_r$, and

$$B = (N_r+1)(M_m-M_r)/(M_r+1) \tag{9}$$

for $M_r+C \geq N_r$.

Estimator bias, therefore, is the sum of two principal components, one positive, the other negative. The positive component results from bias caused by the difference between M_m and M_r. The negative component results from two sources of bias: that caused by the relative difference between M_m and M_r, and that caused by the difference between N_r and M_r+C.

There are two sets of conditions under which this bias is zero:

(1) $M_m = M_r$ and $N_r \leq M_r+C$, or

(2) the positive component equals the negative component.

Condition (2) is not worth further consideration because knowledge about this event is outside the domain of information available to the investigator. On the other hand, supplementary data or educated guesses are often available concerning condition (1). In particular, it is often assumed that $M_m = M_r$. Furthermore, it is almost always true that $M_r+C < N_r$. Thus, under these assumptions, any estimator bias in the estimate of N_r results from the excess of N_r over M_r+C. In this regard, Robson and Regier (1964) point out that such bias is less than two percent if $M_rC/N_r > 4$. Fortunately, this latter condition is rather easily met, particularly if the graphs provided by these authors for determining sample size are used.

A frequent alternative assumption concerning M_r is that $M_r/M_m = N_r/N_m$. This assumption implies that the proportional change in marked members of the population between mark and recapture times is the same as the proportional change in unmarked members. It is particularly reasonable for closed populations, but is not generally valid for open populations such as those found in streams.

Under this assumption, if bias due to the negative component of expression (8) is negligible, it is easily shown that estimator bias is approximately $N_m - N_r$. This is equivalent to confirming the well-known fact that the expectation of our estimate is N_m, rather than N_r. Thus, whenever it is valid to assume the same rates of change in number of both marked and unmarked members of a population, estimation may be thought of in terms of the mark time population. Otherwise, estimation is in terms of the recapture population.

It remains to note that estimator bias decreases as M_r approaches M_m, and/or M_r+C approaches N_r.

VARIANCE OF PETERSEN ESTIMATES

Although the exact form of the variance of Petersen estimates is not known, Chapman (1951) has derived an approximate formula that is applicable to "large" samples, viz.,

$$V(\hat{N}) \approx N^2[(N/MC) + 2(N/MC)^2 + 6(N/MC)^3]$$

The implication of this expression is that variance is inversely related to sample size (C) and number of marked members (M) in the population (N). Unfortunately, this implication is not valid for all sample sizes. That is, there are some sets of values of M and C for which variance of the estimates actually increases as M and/or C increases. Since such aberrant behavior is contrary to the logic of conventional sampling theory, it is worthwhile investigating its cause.

The simplest way to do this is to look at ranges of estimates. There are two reasons for this. First, the range of a set of values is often closely correlated to the variance of those values. Second, in the case of the Petersen estimator, ranges are easily calculated.

In deriving the results, we must consider estimates in two different groups based on the values of M and C.

1. $M+C \leq N$

   $C \leq M$

   In this case, the minimum value that R can take is zero, and the maximum value is C. Thus, the maximum value that the estimate can take is

$$\hat{N}_{max} = (M+1)(C+1)/(0+1) - 1$$
$$= MC + M + C$$

and the minimum value that the estimate can take is

$$\hat{N}_{min} = (M+1)(C+1)/(C+1) - 1$$
$$= M$$

The range of the estimates is thus

$$\hat{N}_{max} - \hat{N}_{min} = MC + M + C - M$$
$$= C(M+1)$$

Since this is an increasing function of M and C, we should not be unduly surprised that variance of the estimates also increases. If we look at the ratio of extreme estimates, we find that

$$\hat{N}_{max}/\hat{N}_{min} = (MC+M+C)/M$$
$$= C + 1 + C/M$$

Since this also is an increasing function of C (but not of M), the basis for an increasing variance is even more firmly established. The foregoing results are symmetric in M and C. That is, if $M < C$, then

$$\hat{N}_{max} - \hat{N}_{min} = M(C+1)$$

and,

$$\hat{N}_{max}/\hat{N}_{min} = M + 1 + M/C$$

2. $N < M + C$

   $C \leq M$

Under these conditions the minimum value that R may take is M+C-N, the maximum value C. Then,

$$\hat{N}_{max} = (M+1)(C+1)/(M+C-N+1) - 1$$
$$= (MC+N)/(M+C-N+1)$$

and,

$$\hat{N}_{min} = (M+1)(C+1)/(C+1) - 1$$

$$= M$$

so that,

$$\hat{N}_{max} - \hat{N}_{min} = (MC+N)/(M+C-N+1) - M$$

$$= (N-M)(M+1)/(M+C-N+1)$$

Since this is a decreasing function of both M and C under the conditions imposed, variance of the estimates must also decrease. Again, this result is symmetric in M and C, so that, if $M < C$,

$$\hat{N}_{max} - \hat{N}_{min} = (N-C)(C+1)/(M+C-N+1)$$

In summary then, for $M + C \leq N$, the range and ratio of extreme values of Petersen estimates increases with increasing M and/or C, whereas for $N < M + C$, the opposite is true. This is why we find that variance of the estimates shows similar behavior. It is important to note, however, that the correspondence is far from exact. This is because the variance depends not only on the deviations of estimates from expectation, but also on their probabilities of occurrence. As it turns out, larger deviations caused by more extreme values tend to be offset by much smaller probabilities.

Unfortunately, it is not mathematically possible to show combinations of M and C, as a function of N, such that any values larger than these would result in reduced variance of estimates. However, a very satisfactory alternative is to make use of the aforementioned graphs published by Robson and Regier (1964). These provide combinations of M and C, for any N (guessed in advance), such that if the guessed value of N is reasonably accurate, the resulting estimate will have limits that contain the true population value with predetermined confidence. Anyone using these graphs may be confident that the values of M and C shown for any given N are in a region where variance of estimates decreases with increasing M and/or C.

TESTS OF HYPOTHESES

In our applications of the Petersen method in Montana, there are two sources of error with which we are particularly concerned: (1) nonuniform ratios of marked to unmarked fish throughout the distribution of the population; and, (2) loss of marked fish from the population. There are two ways of trying to cope with these problems. First, we can take steps to minimize them. Second, we can try to measure the extent of their occurrence.

In regard to ratios of marked to unmarked fish, it has already been recommended that, where feasible, the best way to minimize nonuniformity is to blanket the entire study section on the mark run, concentrating especially on those portions of the habitat where shocking efficiency is low. The objective is to mark as many fish as possible, and release them as close to their original territory as possible.

Marked fish may be lost as a result of death and/or movement out of the study section. Loss due to death is essentially eliminated by conducting studies at times of the year when water temperatures are below 65 °, and by properly capturing and handling the fish. Loss due to movement can be minimized by blocking off the study section, although this is usually impractical for the time span of studies as conducted in Montana.

Measurement of Uniformity of Ratios

The degree of uniformity of ratios of marked to unmarked fish may be measured by sampling at several sites within the study section during the recapture run. If the ratios are relatively uniform, and other requirements of the method are satisfied, the implication is that the estimate is a reasonable one.

A statistical test that may be used in this situation is χ^2. If the study section has been divided into n subsections, and recapture samples have been selected in each, then an n x 2 contingency table may be constructed (Table 1).

Table 1. OBSERVED AND EXPECTED VALUES OF MARKED AND UNMARKED FISH SELECTED IN n RECAPTURE SAMPLES TO TEST THE HYPOTHESIS OF UNIFORM RATIOS OF MARKED TO UNMARKED FISH

		SUBSECTION				
		1	2	...	n	Total
Marked	OBSERVED	R_{o1}	R_{o2}		R_{on}	R
	EXPECTED	R_{e1}	R_{e2}		R_{en}	
Unmarked	OBSERVED	U_{o1}	U_{o2}		U_{on}	U
	EXPECTED	U_{e1}	U_{e2}		U_{en}	
Total		C_1	C_2		C_n	C

Expected values are calculated as

$$R_{ei} = RC_i/C$$

and

$$U_{ei} = C_i - R_{ei}$$

The test is calculated as

$$\chi^2_{n-1} = \sum_{i=1}^{n} (R_{oi}-R_{ei})^2/R_{ei} + \sum_{i=1}^{n} (U_{oi}-U_{ei})^2/U_{ei}$$

If $\chi^2_{\alpha,n-1}$, is tabled χ^2 at the α probability level with n-1 degrees of freedom, and $\chi^2_{\alpha,n-1} < \chi^2_{n-1}$, then with 100(1-α) percent confidence we reject the hypothesis of uniform ratios of marked to unmarked fish.

In applying this test, the investigator should keep in mind that it is only valid in the situation, during the mark run, where all the fish are uniformly catchable, or alternatively, where fishing effort is inversely proportional to catchability.

It is recommended that n be chosen so that it is no larger than $\widetilde{MC}/(10\tilde{N})$, where $\tilde{N}$, $\tilde{M}$, and $\tilde{C}$ are advance estimates of N, M, and C respectively. If good estimates of these quantities are made, then the average value of R_{oi} will be approximately 10 or larger, thus ensuring a valid test of the hypothesis.

In regard to defining the location of the subsections, this is best done on the basis of habitat type or some other relevant consideration, rather than merely dividing the stream along the direction of flow. Such a procedure will provide a more meaningful test of the hypothesis.

Measurement of Random Mixing

In those situations where it is desired to learn the actual extent of random mixing of marked and unmarked fish, a slightly different test may be applied. As in the foregoing, the study section is divided into n subsections, but in this case fish are marked uniquely in each. During the recapture run, marked fish are recorded according to both subsection of marking and subsection of recapture. An (n+1) x n contingency table may then be constructed (Table 2).

Table 2. OBSERVED AND EXPECTED VALUES OF MARKED AND UNMARKED FISH IN n RECAPTURE SAMPLES TO TEST THE HYPOTHESIS OF RANDOM MIXING OF MARKED AND UNMARKED FISH

		Recaptured in Subsection 1	Recaptured in Subsection 2	...	Recaptured in Subsection n	Total
Marked in Subsection 1	OBS EXP	R_{o11} R_{e11}	R_{o12} R_{e12}		R_{o1n} R_{e1n}	R_1
Marked in Subsection 2	OBS EXP	R_{o21} R_{e21}	R_{o22} R_{e22}		R_{o2n} R_{e2n}	R_2
. . .						
Marked in Subsection n	OBS EXP	R_{on1} R_{en1}	R_{on2} R_{en2}		R_{onn} R_{enn}	R_n
Unmarked	OBS EXP	U_{o1} U_{e1}	U_{o2} U_{e2}		U_{on} U_{en}	U
Total		C_1	C_2		C_n	C

Expected values are calculated as

$$R_{eij} = R_i C_j / C$$

and

$$U_{ej} = C_j - \sum_{i=1}^{n} R_{eij}$$

The test is calculated as

$$\chi^2_{n(n-1)} = \sum_{i=1}^{n} \sum_{j=1}^{n} (R_{oij} - R_{eij})^2 / R_{eij} + \sum_{j=1}^{n} (U_{oj} - U_{ej})^2 / U_{ej}$$

If $\chi^2_{\alpha,n(n-1)}$ is tabled χ^2 at the α probability level with n(n-1) degrees of freedom, and $\chi^2_{\alpha,n(n-1)} < \chi^2_{n(n-1)}$, then with 100(1-$\alpha$) percent confidence we reject the hypothesis of random mixing of marked and unmarked fish.

Here again, this test should be applied only to those fish that, during the mark run, are uniformly catchable or against which the fishing pressure is inversely proportional to catchability. Also, in this case, n should be chosen so that it is no larger than $\widetilde{MC}/(100\tilde{N})$, thus ensuring that, on the average, R_{oij} will be 10 or larger.

Measurement of Loss of Marked Fish

The extent to which marked fish leave the study section can be monitored by establishing n subsections of equal length, linearly arranged from the upstream end (subsection 1) to the downstream end (subsection n) of the section. If fish are marked uniquely in each, then during the recapture run marked fish may be recorded according to subsection of marking and subsection of recapture. The objective is to estimate the total number of marked fish that leave the section by migrating upstream, and the total number that leave by migrating downstream (assuming that, in general, these migration rates differ).

Now the total number of fish that migrate upstream out of the section can be divided into n groups, representing the n subsections where the fish were marked. Each of these groups can be further divided into subgroups representing the different distances migrated. Assuming that distance of movement is independent of originating subsection, and that catchability of fish is the same in all subsections, the estimated number of fish that leave the section by migrating upstream from subsection j a distance represented by the distance between subsections n-i+1 and 1 is M_jP_i. M_j is the number of fish marked in subsection j ($j \leq n-i$), and P_i is the estimated proportion of fish that leave the section by migrating upstream a distance represented by the distance between subsections n-i+1 and 1, where i is the number of pairs of subsections n-i subsections apart. That is,

$$P_i = (1/i)\sum_{k=1}^{i} R_{n-i+k,k}/R_{n-i+k,n-i+k} \qquad (10)$$

where R_{pq} = number of fish marked in subsection p and recaptured in subsection q.

Thus, the estimated number of fish that leave the section by migrating upstream is

$$\hat{M}_u = \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} M_j P_i \qquad (11)$$

This estimator is negatively biased by an amount representing the number of marked fish that migrated upstream a distance equal to or greater than the length of the entire section. However, by making the section long enough, this bias can be made as small as desired. Also, the more subsections that are used, the more accurate should be the estimate.

By logic similar to the above, we find that the estimated number of marked fish that leave the section by migrating downstream is

$$\hat{M}_d = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} M_j P_i \qquad (12)$$

where, in this case,

$$P_i = (1/i)\sum_{k=1}^{i} R_{k,n-i+k}/R_{k,k} \qquad (13)$$

= estimated proportion of fish that leave the section by migrating downstream a distance represented by the distance between subsections 1 and n-i+1.

The estimated total loss of marked fish from the study section is then

$$\hat{M}_l = \hat{M}_u + \hat{M}_d$$

and an estimate of the number of marked fish remaining at recapture time is

$$\hat{M}_r = M_m - \hat{M}_l$$

Thus, a lower bound to the relative bias of our population estimate, due to loss of marked fish, may be approximated as

$$\begin{aligned}\hat{B}_{rel} &= (M_m - \hat{M}_r)/\hat{M}_r \\ &= \hat{M}_l/(M_m - \hat{M}_l) \qquad (14)\end{aligned}$$

EXAMPLES

The objective in this section is to illustrate the important results discussed previously. In particular, we shall see that (1) even though bias is present in a sampling and estimation scheme, individual estimates do not necesssarily reflect this bias; (2) bias in Petersen estimates is at least partially a function of sample size; and, (3) variance of Petersen estimates increases with sample size over a significant range of values of M and C.

In evaluating the implications of these examples, the reader should bear in mind that the purpose here is to illustrate qualitative characteristics of bias and variance that are independent of population size. Quantitative characteristics that are a function of the unrealistically small population that is used as a basis for the examples cannot be reliably extrapolated to larger populations.

Also illustrated are a χ^2 test of uniformity of ratios, and estimation of loss of marked fish due to migration.

EXAMPLE I

Conditions: $N = 8$

$M = 2$

$C = 2$

Total number of samples = 28

Estimator is $\hat{N} = (M+1)(C+1)/(R+1) - 1$

Expectation formula (7) is applicable.

Number of Recaptures (R)	Number of Samples	Probability of Occurrence	Estimate ($\hat{N}$)
0	15	15/28	$(3)(3)/(1) - 1 = 8$
1	12	12/28	$(3)(3)/(2) - 1 = 3\frac{1}{2}$
2	1	1/28	$(3)(3)/(3) - 1 = 2$

By the definition of expectation (which is valid for any estimator),

$$E(\hat{N}) = (8)(15/28) + (3\tfrac{1}{2})(12/28) + (2)(1/28)$$

$$= 164/28 \simeq 5.9$$

By formula (7), which is valid for the Petersen estimator,

$$E(\hat{N}) = 8 - (8-2-2)(15/28)$$

$$= 164/28$$

By the definition of bias,

$$B = 164/28 - 8$$

$$= -60/28 \simeq -2.1$$

The sole source of this bias is sample size.

The interesting feature of this example is that, even with a negative bias of approximately 25 percent, over 50 percent of the estimates are identical to the attribute being estimated. We know from previous discussion involving expression (1) that the reason for this is that the bias is exactly balanced by sampling variability (Table 3).

Table 3. RECONSTRUCTION OF INDIVIDUAL COMPONENTS OF ESTIMATES

ESTIMATE ($\hat{N}$)	=	ATTRIBUTE (N)	+	BIAS (B)	+	DEVIATION FROM EXPECTATION (d)
8		8		-60/28		60/28
3½		8		-60/28		-66/28
2		8		-60/28		-108/28

EXAMPLE II

Conditions: $N = 8$

$M = 3$

$C = 3$

Total number of samples = 56

Estimator is $\hat{N} = (M+1)(C+1)/(R+1) - 1$

Expectation formula (7) is applicable.

Number of Recaptures (R)	Number of Samples	Probability of Occurrence	Estimate ($\hat{N}$)
0	10	10/56	15
1	30	30/56	7
2	15	15/56	4 1/3
3	1	1/56	3

$$E(\hat{N}) = (15)(10/56) + (7)(30/56) + (4\ 1/3)(15/56) + (3)(1/56)$$

$$= 428/56$$

$$= 8 - (8-3-3)(10/56)$$

$$B = 428/56 - 8$$

$$= -20/56$$

The sole source of this bias is sample size.

This example demonstrates further the relative effects of sampling variability and bias. Thus, even though there is a negative bias of approximately 5 percent, sampling variability in 10 of the 56 samples resulted in overestimation of the attribute by nearly 100 percent.

EXAMPLE III

Conditions: $N = 8$

$M = 4$

$C = 4$

Total number of samples = 70

Estimator is $\hat{N} = (M+1)(C+1)/(R+1) - 1$

Expectation formula (7) is applicable.

Number of Recaptures (R)	Number of Samples	Probability of Occurrence	Estimate ($\hat{N}$)
0	1	1/70	24
1	16	16/70	11 1/2
2	36	36/70	7 1/3
3	16	16/70	5 1/4
4	1	1/70	4

$$E(\hat{N}) = (24)(1/70) + (11\ 1/2)(16/70) + (7\ 1/3)(36/70) + (5\ 1/4)(16/70) + (4)(1/70)$$

$$= 8$$

$$= 8 - (8-4-4)(1/70)$$

$$B = 8 - 8$$

$$= 0$$

This example verifies that sample size bias in the Petersen estimation method becomes zero when $N \leq M+C$, even though $P(R=0) > 0$. Furthermore, it shows that even though the bias is zero, no individual estimate is equal to the population attribute.

These first three examples illustrate two general features of Petersen estimates discussed previously, both of which relate to increases in M and C in the range $M+C \leq N$. First, sample size bias decreases. Second, the spread between maximum and minimum estimates increases. As noted earlier, this is responsible for an increase in variance of the estimates.

EXAMPLE IV

Conditions: $N_m = 8$

$M_m = 4$

$N_r = 7$

$M_r = 3$

$C = 4$

Total number of samples = 35

Estimator is $\hat{N} = (M_m+1)(C+1)/(R+1) - 1$

Expectation formula (6) is applicable.

Number of Recaptures (R)	Number of Samples	Probability of Occurrence	Estimate ($\hat{N}$)
0	1	1/35	24
1	12	12/35	11 1/2
2	18	18/35	7 1/3
3	4	4/35	5 1/4

$$\begin{aligned} E(\hat{N}) &= (24)(1/35) + (11\ 1/2)(12/35) + (7\ 1/3)(18/35) + (5\ 1/4)(4/35) \\ &= 315/35 \\ &= 9 \\ &= (7+1)(4+1)/(3+1) - 1 \\ B &= 9 - 7 \\ &= 2 \end{aligned}$$

This example demonstrates the bias that results from a biased estimator because of a change in number of marked fish occurring between times of mark and recapture.

EXAMPLE V

Conditions: $N = 8$

$M = 4$

$C = 4$

Total number of samples = 70

Estimator is $\hat{N} = (M+1)(C+1)/(R+1)$[1/]

[1/] This is formula 3.9 of Ricker (1958).

Number of Recaptures (R)	Number of Samples	Probability of Occurrence	Estimate $(\hat{N})$
0	1	1/70	25
1	16	16/70	12 1/2
2	36	36/70	8 1/3
3	16	16/70	6 1/4
4	1	1/70	5

$$E(\hat{N}) = (25)(1/70) + (12\ 1/2)(16/70) + (8\ 1/3)(36/70) + (6\ 1/4)(16/70) + (5)(1/70)$$

$$= 630/70$$

$$= 9$$

$$B = 9 - 8$$

$$= 1$$

This example demonstrates another kind of estimator bias, one that results simply from the form of the estimator, rather than from inadequate sample size or loss of marked fish. In this case, the bias is always 1 regardless of population size.

EXAMPLE VI

The purpose of this example is to demonstrate the bias in population estimation that results when geographic sampling is used in an attempt to overcome the problem of nonuniform association of marked and unmarked fish.

In this case, we assume a population of size $N = 8$, with $M = 4$ marked members, $C = 4$ recapture sample size, and estimator $\hat{N} = (M+1)(C+1)/(R+1) - 1$. Since the investigator suspects that the marked and unmarked fish are not randomly associated with each other, he attempts to overcome this problem by dividing the geographic range of the population into two subsections, one of which is to be randomly selected for sampling. This division results in ratios of marked to unmarked fish in the two subsections of 1:3 and 3:1, respectively.

Under these conditions, two possible samples may be selected, one in each subsection, and both of which contain four fish. In subsection 1, one fish is marked and in subsection 2, three fish are marked. The total population estimates that result from these samples are:

$\hat{N} = (5)(5)/(2) - 1 = 11\ 1/2$ (based on the sample from subsection 1);

and,

$\hat{N} = (5)(5)/(4) - 1 = 5\ 1/4$ (based on the sample from subsection 2).

Since each of the two samples is equally likely, the expectation of the estimate is:

$$E(\hat{N}) = (11\ 1/2)(1/2) + (5\ 1/4)(1/2)$$
$$= 8\ 3/8$$

with bias of

$$B = 8\ 3/8 - 8$$
$$= 3/8$$

EXAMPLE VII

In this example we examine the effect of changes in M and C on variance and mean square error of the resulting estimates (Table 4). Estimates were calculated using Chapman's estimator, i.e., $\hat{N} = (M+1)(C+1)/(R+1) - 1$. Since both the variance and mean square error are symmetric in M and C, those entries below the main diagonal have been omitted.

Table 4. VARIANCE AND MEAN SQUARE ERROR OF ESTIMATES FOR VARIOUS COMBINATIONS OF M AND C WHEN N = 8

M \ C	1	2	3	4	5	6	7
1	0.44 1/ 28.00 2/	1.69 15.75	3.75 10.00	6.25 8.50	8.44 9.00	9.19 9.25	7.00
2		5.37 9.96	9.57 10.71	12.24 12.43	11.88 11.89	8.25	3.00
3			13.21 13.33	12.54 12.55	8.57	4.17	1.67
4				8.64	4.61	2.42	1.00
5					2.63	1.43	0.60
6						0.78	0.33
7							0.14

1/ Variance

2/ Mean square error (MSE); for those combinations of M and C for which MSE = variance (bias is zero), MSE is not shown.

This table illustrates how variance increases for most values of M and C in the range $M + C \le N$, but decreases for $N < M + C$, as implied by the results derived previously in terms of ranges of estimates. Mean square error is somewhat erratic.

EXAMPLE VIII

This example presents more data showing the relationship between mean square error of Petersen estimates and various combinations of M, C, and N (Table 5), where $\hat{N} = (M+1)(C+1)/(R+1) - 1$. The value of this example lies in the fact that the population and sample sizes cover a range of values commonly encountered in real life situations. The data imply that, for combinations of M and C below a limit somewhere below that of MC/N = 5, mean square error (hence, variance) increases with increasing M and C.

Table 5. MEAN SQUARE ERROR OF PETERSEN ESTIMATES FOR VARIOUS VALUES OF N, M, AND C[1/]

N	M	C	MC/N	MSE
10,000	100	1,000	10	.1102N
	100	5,000	50	.0107N
	500	5,000	250	.0021N
100,000	100	1,000	1	.2165N
	100	5,000	5	.2996N
	500	5,000	25	.0407N
	1,000	10,000	100	.0091N
1,000,000	100	10,000	1	.2152N
	500	5,000	2.5	.3104N
	1,000	10,000	10	.1132N
	1,000	20,000	20	.0544N

[1/] Adapted from Table 3 of Chapman (1951).

EXAMPLE IX

The purpose of this example is to illustrate the χ^2 test of the hypothesis of uniform ratios of marked to unmarked fish. For this test, we assume that the study section has been divided into 6 subsections at recapture time, with the recaptures as recorded in Table 6.

Table 6. OBSERVED AND EXPECTED VALUES OF MARKED AND UNMARKED FISH SELECTED IN SIX RECAPTURE SAMPLES TO TEST THE HYPOTHESIS OF UNIFORM RATIOS OF MARKED TO UNMARKED FISH

		SUBSECTIONS						
		1	2	3	4	5	6	Total
MARKED	OBSERVED	25	55	50	30	50	100	310
	EXPECTED	35.43	53.14	44.29	17.71	70.86	88.57	
UNMARKED	OBSERVED	175	245	200	70	350	400	1440
	EXPECTED	164.57	246.86	205.71	82.29	329.14	411.43	
TOTAL		200	300	250	100	400	500	1750

Expected values are calculated according to the formulas shown in the section describing this test. For example,

$R_{e1} = RC_1/C = (310)(200)/1750 = 35.43$

$U_{e1} = C_1 - R_{e1} = 200 - 35.43 = 164.57$

χ^2 with 5 degrees of freedom is calculated as

$$\chi^2_5 = (25-35.43)^2/(35.43) + (55-53.14)^2/(53.14) + (50-44.29)^2/(44.29) + (30-17.71)^2/(17.71) + (50-70.86)^2/(70.86) + (100-88.57)^2/(88.57) + (175-164.57)^2/(164.57) + (245-246.86)^2/(246.86) + (200-205.71)^2/(205.71) + (70-82.29)^2/(82.29) + (350-329.14)^2/(329.14) + (400-411.43)^2/(411.43)$$

$$= 24.32$$

Since $\chi^2_{.005,5} = 16.7$, we reject the hypothesis with 99.5 percent confidence.

EXAMPLE X

The purpose of this example is to illustrate how to calculate the estimated loss, due to migration, of marked fish from the study section, and consequently, provide a lower bound to this particular source of bias in the population estimate. In this case, we assume that our stream section is divided into four equal subsections (numbered 1 to 4 from upstream end to downstream end). Also, we assume numbers marked and recaptured by subsection, as shown below.

		Rpq			
		Recapture Subsections			
Marking Subsections	Mj	1	2	3	4
1	150	56	10	0	2
2	400	4	150	12	2
3	100	2	4	30	5
4	200	1	5	3	80

M_m = 850

Loss due to upstream migration is estimated as follows. By expression (10),

$$P_1 = R_{4,1}/R_{4,4} = 1/80 = .0250$$

$$P_2 = (1/2)(R_{3,1}/R_{3,3} + R_{4,2}/R_{4,4}) = (1/2)(2/30 + 5/80) = .0646$$

$$P_3 = (1/3)(R_{2,1}/R_{2,2} + R_{3,2}/R_{3,3} + R_{4,3}/R_{4,4})$$
$$= (1/3)(4/150 + 4/30 + 3/80) = .0658$$

By expression (11),

$$\hat{M}_u = (M_1 + M_2 + M_3)P_1 + (M_1 + M_2)P_2 + M_1P_3$$

$$= (150 + 400 + 100)(.0250) + (150 + 400)(.0646) + (150)(.0658)$$

$$= 61.65$$

Loss due to downstream migration is estimated as follows. By expression (13),

$$P_1 = R_{1,4}/R_{1,1} = 2/56 = .0357$$

$$P_2 = (1/2)(R_{1,3}/R_{1,1} + R_{2,4}/R_{2,2}) = (1/2)(0/56 + 2/150) = .0067$$

$$P_3 = (1/3)(R_{1,2}/R_{1,1} + R_{2,3}/R_{2,2} + R_{3,4}/R_{3,3})$$

$$= (1/3)(10/56 + 12/150 + 5/30) = .1417$$

By expression (12),

$$\hat{M}_d = (M_2 + M_3 + M_4)P_1 + (M_3 + M_4)P_2 + M_4P_3$$

$$= (400 + 100 + 200)(.0357) + (100 + 200)(.0067) + (200)(.1417)$$

$$= 55.34$$

Thus, the total loss is estimated as

$$\hat{M}_l = \hat{M}_u + \hat{M}_d$$

$$= 61.65 + 55.34$$

$$= 117$$

and the bias in our population estimate due to this loss is estimated as at least

$$\hat{B}_{rel} = \hat{M}_l/(M_m - \hat{M}_l)$$

$$= 117/(850 - 117)$$

$$= .1596, \text{ or about 16 percent.}$$

CONCLUSIONS

On the basis of experience and the analyses contained in this paper, we feel that the following conclusions and recommendations are justified:

1. Chapman's estimator of population size, used in conjunction with the Petersen mark-recapture method, is the best one available. However, it exhibits two unusual properties that investigators should be aware of:
    a. It is potentially unbiased only for certain sample sizes of M and C;
    b. Variance of the estimates is not a decreasing function of sample size over the entire range of values of sample size.

    Sample sizes recommended by Robson and Regier for various population sizes essentially eliminate any undesirable consequences of these two properties.
2. There are several sources of bias in the Petersen method. Those attributable to lack of probability sampling are:
    a. Nonrandom selection by the fishing method of marked and unmarked fish; <u>or</u>,
    b. Nonrandom association of marked and unmarked fish.

    If the investigator himself can establish reasonably uniform ratios of marked to unmarked fish throughout the population, then the need to depend on probability sampling is eliminated.

    Those biases attributable to the estimator itself are:
    a. Use of any estimator other than Chapman's;
    b. Sample size, i.e., when $M + C < N$; bias due to this source is negligible when $MC/N > 4$;
    c. Loss of marked fish between mark time and recapture time.

Biases attributable to inaccurate recording of data are not discussed in this paper. For information about these, the section on the Petersen method by Ricker (1958) may be consulted.

3. The Petersen mark-recapture method is a reliable tool for estimating the size of fish populations under the following conditions:
   a. The graphs provided by Robson and Regier are used to estimate required sample sizes;
   b. Reasonably uniform ratios of marked to unmarked fish are established throughout the distribution of the population;
   c. Loss of marked fish between mark time and recapture time is negligible (or is balanced by a proportionate loss of unmarked fish).

It is recommended that tests of conditions b and c be made as regular a feature of mark-recapture studies as practicable, because failure of these conditions is the most serious source of error in the method. The benefits gained from this practice are twofold. First, it helps determine the reliability of any individual recapture sample. Second, it provides a technique for refining the mark and recapture sampling process in general.

REFERENCES

Bailey, N. T. J. 1951. On estimating the size of mobile populations from recapture data. Biometrika, 38:293-306.

Chapman, D. G. 1951. Some properties of the hypergeometric distribution with applications to zoological sample censuses. Univ. Calif. Publ. Stat., 1(7):131-159.

Ricker, W. E. 1958. Handbook of computations for biological statistics of fish populations. Fish. Res. Bd. Can., Bull. 119. 300 pp.

Robson, D. S. and H. A. Regier. 1964. Sample size in Petersen mark-recapture experiments. Trans. Am. Fish. Soc., 93(3):215-226.

Vincent, R. 1971. River electrofishing and fish population estimates. Prog. Fish-Cult., 33(3):163-169.