SAMPLING METHODS FOR BIOLOGICAL POPULATIONS

by Kimberly Sinclair

Masters Writing Project, April 1993
Montana State University
Bozeman, Montana 59715

PURPOSE: to summarize sampling methods of biological populations for individuals with minimal knowledge of statistics.

TABLE OF CONTENTS

1. INTRODUCTION

Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population. Estimating a population total, such as the total number of animals, is often of interest in biological sampling. Although there are many other inferences that can be made about a population we will refer only to population totals. Several different methods exist for sampling biological populations, more than can be discussed here. The following methods will be reviewed in detail with respect to assumptions and implementation: simple latin square sampling + 1 (SLSS+1), capture-recapture sampling, inverse sampling, line transect sampling, line intercept sampling, and adaptive cluster sampling.

2. REVIEW OF TERMINOLOGY

In this section, statistical terminology will be defined to allow for more detailed discussions of the methodologies in further sections.

i) *parameter* - a numerical value that describes a population. Parameters are usually unknown values.

ii) *statistic* - a numerical value that describes a sample which is a subset of the population. Statistics are used to estimate parameters.

iii) *sampling units* - nonoverlapping collections of elements from the population that cover the entire population. A sampling unit is defined relative to the type of sampling procedure. Some examples of a sampling unit are a plot of land (quadrat), a herd of animals, an individual animal and possibly a habitat region.

iv) *independent* - choosing any one sampling unit will not affect whether or not another sampling unit is chosen.

v) *simple random sampling without replacement (SRSWOR)* - a classical sampling design defined as removing a unit from the population once it has been sampled and sampling until n distinct units are chosen. At each stage of sampling the units in the population have an equal probability of being sampled. The n selections are not independent because the probability of selecting a unit depends on how many units have already been removed from the population (i.e. sampled).

vi) *simple random sampling with replacement (SRSWR)* - a classical sampling design defined as selecting a unit from the population and replacing it into the population before selecting the next unit. A sample of size n obtained in this manner may include repeat selections. The n selections are independent and each unit has the same probability of being selected for the sample. Every possible combination of n units, distinguishing order of selection and including repeat selections, has equal probability of being the chosen sample.

vii) *stratified random sampling* - a classical sampling design in which the population is divided into groups, or strata, based on an auxiliary influencing variable. Within each strata a simple random sample of the sampling units is taken and the total number of objects from the population is counted. This design is used to control variability and obtain more information from a study. Some advantages of stratification include: possibly obtaining a smaller bound on the error of estimation (i.e. estimates will be more accurate), particularly if the units within a group are homogeneous; the cost per observation may be reduced by stratifying units into convenient groupings; and estimates of the population total may be desired for subgroups of the population in which case the subgroups should be the strata.

viii) *systematic sampling* - a classical sampling design in which a sample is obtained by randomly selecting one unit from the first k units and then selecting every kth unit afterwards. This is often called a 1-in-k systematic sample with a random start.

ix) *latin square sampling* - a sampling design in which the sampling units form a perfect n x n square (e.g. a square region of land that can be divided into n rows and n columns so there are n^2 quadrats which are the sampling units). The design is based on sampling n units such that only one unit is selected from each row and each column.

x) *detectability or sightability* - the probability that an object (e.g. bird, plant, seal) in a particular sampling unit or region is observed. Detectability is a nonsampling error. When sampling a given unit it is assumed that all variables of interest are recorded without error. In most situations it is inevitable that objects are "missed" and detectability compensates for this error.

Although assumptions do not have a direct definition it is important to discuss their importance with respect to sampling methods. For each method the assumptions necessary for the method to be valid are stated. One should examine these assumptions in detail and determine whether or not they are realistic for the particular sampling situation. If they are not realistic then attention should be given to what influence they have on estimators and whether or not adjustments can be made.

3. SIMPLE LATIN SQUARE SAMPLING + 1 (SLSS+1)

SLSS+1 is used to increase coverage of the population. SLSS+1 ensures that the sample covers the entire population which will provide more information per unit cost than simple random sampling. It has been shown to be superior to systematic samples of the same size because an unbiased estimate of the variance can be obtained. Systematic samples will give unbiased variance estimates only if the design is replicated but this requires additional cost and effort in most cases. SLSS+1 is often used when sampling quadrats for presence/absence of a species.

i) *Design*

The first step is to arrange the population of units in a square of dimension n so the number of units in the population is $N = n^2$. A simple latin square sample of size n is drawn from the population. Finally one additional unit is randomly drawn from the remaining units not included in the initial latin square sample. The total number of objects observed in each unit selected for the SLSS+1 is measured and used to estimate the total number of objects in the population. Choosing an additional unit allows for an unbiased estimator of the variance to be obtained.

ii) *Assumptions*

The main assumption for this model to be valid is that the population of units can be arranged in a square of dimension n so that a latin square sample can be obtained.

iii) *Comments*

A SLSS without the additional unit will give biased or conservative estimators of the variance, therefore by applying SLSS+1 these problems can be overcome and unbiased estimators can be obtained. This method is only possible if the population can be arranged into a perfect square. For a situation where the researcher is more interested in obtaining maximum cover over the population simple latin square sampling + 1 is ideal.

iv) *References*

For more details of SLSS+1 contact Dr. Patricia Munholland or Dr. John Borkowski. The methodology has been submitted for publication and will be accessible in the future at which time the reference will be updated.

4. CAPTURE - RECAPTURE SAMPLING

The method involves obtaining two independent samples from the population as described below. The data obtained from these samples can be used to estimate population size. The above method can be repeated many times to obtain even more information such as estimates of survival rate and recruitment for each sampling period. Here we will discuss the simple method of selecting only two samples. It is evident from the method described here that the sampling unit is defined as the individual animal.

i) *Design*

An initial sample of size n_1 is obtained and the units in the sample are marked or otherwise identified. A second sample of size n_2, independent of the first sample, is obtained and the number of marks in the second sample is recorded as m_2. If the second sample is representative of the whole population, then the proportion of marked animals in the second sample should be proportional to the proportion of marks in the population.

ii) *Assumptions*

The assumptions necessary for this model to be valid are: 1) the population is closed during the study which means no emigration or immigration occurs; 2) the first and second samples are independent simple random samples; 3) tags are not lost between the two samples; and 4) all marks are accurately identified. Assumption 1 ensures that N, the total number of animals in the population, remains constant throughout the study. If N is not constant then it becomes a random variable and the analysis becomes more complex. Assumption 2 is stating that marked and unmarked animals behave in a similar manner. For this assumption to be valid animals must not be "trap-happy" or "trap-shy". If assumption 3 is violated then N will be overestimated which may have serious consequences.

iii) *Comments*

After collecting the data by the method described above, the model that describes this situation is that of the hypergeometric distribution. The hypergeometric distribution defines the probability of observing m_2 marked animals in the second sample. In estimating a population total it is necessary to assume that the proportion of marks in the sample (i.e. m_2 divided by n_2) is a reasonable estimate of the proportion of animals in the population (i.e. n_1 divided by N which is the population total). We can then set the two proportions equal and solve for N which is the estimate of the total number of animals in the population. It can be shown that the estimate of N is equal to $(n_1 * n_2)/m_2$. This method is also called single mark release.

iv) *References*

For more details of the capture-recapture method the reader is referred to the text Sampling by Steven K. Thompson, the monograph Statistical Inference for Capture-Recapture Experiments by Pollock et al, and the text The Estimation of Animal Abundance by Seber. Complete references are given at the end of this article.

5. INVERSE SAMPLING

The method is very similar to capture-recapture but the theory differs. Inverse sampling is used mainly for rare and elusive populations. Recall from capture-recapture, the size of the second sample is preset. However, when sampling a rare population it would be possible to observe no marks in the second sample. Observing no marks would invalidate our estimator for N, the population size. Inverse sampling accommodates this situation by predetermining the number of marks to observe in the second sample thus leaving the sample size as the unknown (i.e. sample until m_2 marks are observed then n_2 will be known).

i) *Design*

An initial sample of size n_1 is obtained and the units in the sample are marked or otherwise identified. A second sample, independent of the first sample, is obtained by sampling until m_2 marks are observed. Once m_2 marks are observed then the total number of animals (marked and unmarked) observed in the second sample will be known which is n_2. Once again, if the second sample is representative of the whole population, then the proportion of marked animals in the second sample should be proportional to the proportion of marks in the population.

ii) *Assumptions*

The assumptions for this model to be valid are: 1) the population is closed during the study; 2) the two samples are independent simple random samples; 3) marks are not lost between samples; 4) all marks are correctly identified; 5) the number of marks to be observed in the second sample is greater than zero; and 6) sampling without replacement is approximately equal to sampling with replacement (i.e. the second sample size is very small relative to the population size N). Assumption 6 ensures that the second sample size does not grow infinitely large.

iii) *Comments*

After collecting the data by the method described above, the model that describes this situation is that of the negative binomial distribution. The distribution differs from that of the capture-recapture model because n_2 is unlimited in size. The negative binomial distribution defines the probability of observing a second sample of size n_2. Through statistical theory regarding the negative binomial distribution it can be shown that the estimate of the population total, N, is equal to $(n_1 * n_2)/m_2$ which is the same as the estimator for the capture-recapture method.

iv) *References*

For more details of the inverse sampling method the reader is referred to the text The Estimation of Animal Abundance by Seber. Complete references are given at the end of this article.

6. LINE TRANSECT SAMPLING

The method involves traversing a selected line and noting the location of any observed units with respect to the line. Line transect sampling is used for many types of populations some of which are birds, plants and mammals. It is reasonable to conclude that animals closer to the line are more likely to be observed than animals at a distance from the line. This is not to say that there are more animals closer to the line but that they are easier to detect when near the line. Because we are interested in estimating animal abundance we must account for the change in detectability when animals vary in distance from the transect line. A line transect is characterized by a detectability function giving the probability that an animal at a given location is detected. In this design the variable of interest is the number of animals observed from a particular transect. However, the sample size n refers to the number of transects selected and not to the variable of interest. The theory of transect sampling can be adapted to circular plots which is useful for sampling bird populations. In this situation the method is called variable area circular sampling.

i) *Design*

A random sample of n transects in the study area will be selected as follows. A straight baseline of length W is drawn across (or below) the study region on a map. The study area need not be regular in shape. The width of the baseline is 0 to W and is perpendicular to every object in the population. A random sample of n transect locations is selected from the uniform distribution on the interval [0,W]. The uniform distribution refers to every point in 0 to W having the same probability of being selected (i.e. every point is equally likely to be chosen). For each of the n locations chosen a transect is extended perpendicular to the baseline across the study region. The transects are traversed and the number of objects observed is recorded.

ii) *Assumptions*

The assumptions for this model to be valid are: 1) points on the line will never be missed, they are seen with probability one; 2) points are fixed at the initial sighting position or they do not move before being detected and no points are counted twice; 3) distances and angles are measured accurately; 4) there exists a detectability function g(y) that is the conditional probability of observing a point given its right angle distance y from the line; 5) sightings are independent; and 6) the transect line is randomly placed in the region.

iii) *Comments*

Some methods of traversing the selected line include foot, snowmobile, and skiing. Line transect methods have also been applied to aerial surveys, surveys from research vessels, and sightings of animals from automobiles. Since the transect is equally likely to be anywhere in [0,W], the objects are equally likely to be any distance in 0 to W from the transect. Therefore the distance a randomly selected point is from the transect is also distributed as a uniform distribution. This fact along with the detectability function can be used to estimate the density of animals in the population. There are several methods for determining the detection function some of which include the narrow-strip method and the smooth-by-eye method.

iv) *References*

For details of this method the reader is referred to the text Sampling by Thompson which gives a detailed discussion for estimating detectability. Also refer to the text The Estimation of Animal Abundance by Seber, the article A Review of Estimating Animal Abundance by Seber, and the monograph Estimation of Density from Line Transect Sampling of Biological Populations by Burnham, Anderson and Laake. Complete references are given at the end of this article.

7. LINE INTERCEPT SAMPLING

The method involves randomly placing n transect lines in the study area. Each line is traversed by ground or by air. Once an object of the population intersects the line, the variable of interest is recorded for the object. Line intercept sampling differs from line transect sampling in that there is no detectability function to estimate. In line intercept sampling every object is observed with probability one because it is only observed if it crosses the line. One use of line intercept sampling is the study of snow tracks to estimate wildlife populations. It is also quite common in botany studies.

i) *Design*

Determining the area to be studied is the first step. This area is usually defined as a rectangular region but may also be irregular in shape. Next, randomly select n points from the interval 0 to W where W is the length of the baseline representing the width of the region. The transect lines will start at each of these points and extend perpendicular to the baseline across the study area (i.e. the lines will be parallel to each other). Traverse the lines in whichever manner is appropriate and record the data from each object that intersects the line.

ii) *Assumptions*

The assumptions necessary for this model to be valid are: 1) no animals or objects move during sampling; 2) the transects are randomly placed in the study region; 3) all objects crossing the transect line are sighted. When sampling snow tracks some additional assumptions that are necessary are 4) all animals move after snowfall (i.e. there are tracks to be sampled), and 5) a 1-to-1 correspondence between track and animal can be established.

iii) *Comments*

Although this method is similar to line transect sampling, points that are seen but do not intersect the line are not allowed to be sampled. Line transect sampling will usually give more information but if it is impossible or difficult to determine the detectability function then line intercept sampling would be more appropriate.

iv) *References*

For more details of line intercept sampling the reader is referred to the text Sampling by Thompson, and the text The Estimation of Animal Abundance by Seber.

8. ADAPTIVE CLUSTER SAMPLING

The method takes advantage of the population characteristics to obtain more precise estimates of population abundance or density, for a given sample size or cost, than is possible with conventional methods. Adaptive cluster sampling methods give lower variance than conventional methods for populations that are rare and clustered. This method utilizes the clusters encountered when sampling by obtaining as much information as possible from them.

i) *Design*

The method requires dividing the study region into the appropriate number of units to form a grid, usually quadrats of land. An initial random sample of size n is selected from the population. Adjacent neighboring units in the grid are added to the sample whenever one or more of the objects of the population is observed in a selected unit. If any of these additional units contains an object of the population then more units are added. This is repeated until no objects are present in adjacent units.

ii) *Assumptions*

The assumptions necessary for this model to be valid are 1) the original sample is a simple random sample, and 2) that every unit *i* in the population has a neighborhood of units that is defined around and including *i*.

{NOTE: include a picture or something}

iii) *Comments*

Conventional estimators (e.g. sample mean) may be biased when used with adaptive cluster sampling. However, unbiased estimators of the population mean and its variance have been derived. Adaptive cluster sampling can also be extended to stratified and systematic sampling situations. Whether an adaptive design is more efficient or less efficient than a classical design such as simple random sampling depends on the type of population being sampled. Adaptive cluster sampling is most appropriate when sampling populations that are rare or clustered.

iv) *References*

For more details of adaptive cluster sampling the reader is referred to the article Adaptive Cluster Sampling by Thompson and the text Sampling by Thompson. For details on stratified and systematic adaptive cluster sampling refer to the text Sampling by Thompson. Complete references are given at the end of this article.