Why Sample?

Often researchers are interested in answering questions about a particular population. They might be interested in the density, species richness, or specific life history parameters such as survival or fecundity. They often want to know how the population is responding to particular management actions. It is usually not feasible to make a complete census of a population because of time and budget constraints. Therefore, a sample of the population is used to make inferences about the whole population. The goal of this type of sampling is to collect data that are representative of the entire population of interest.

Four aspects of a sampling strategy necessary to make accurate and precise inferences about populations are:

1) Well-defined, specific objectives that allow you to answer your question(s) of interest.
2) A probabilistic sampling framework to account for spatial variability.
3) A sampling method that allows you to estimate detection probabilities (i.e. the proportion of individuals present, but not detected during your survey).
4) Sufficient sample sizes to make statistically valid inferences about your population.

Well-defined Objectives:

It is important that your sampling design address the questions of interest. The first step is to explicitly state your objectives and hypotheses, and then tailor your sampling framework around those specific questions. For instance: it would not make sense to ask questions about broad habitat associations of birds but only sample a small number of locations.

A Probabilistic Sampling Framework:

It is virtually impossible to collect information on every individual in a population. With this in mind, you need to design a sampling framework that is representative of the entire population. Historically, many researchers simply used their experience to select representative sampling sites, or they simply selected sites that were logistically convenient, for example along roads or trails. We now understand that this approach results in biased estimates and inappropriate inferences about populations. A large branch of biostatistics has developed to address these types of problems using probabilistic sampling designs. Several representative approaches include:

- Simple random sampling – all sampling units are independent of one another and have equal probability of being selected.
- Systematic sampling – sampling units allocated uniformly over area of interest or according to some other systematic method.
Stratified random sampling – sampling units are first stratified by homogenous features (e.g., habitat) and then sample units are randomly selected within each strata.

A probabilistic sampling design will reduce sampling bias caused by spatial variation. For example, if our parameter of interest is population size ($N$), a probabilistic sampling design will allow us to adjust our estimate of $N$ ($\hat{N}$) by the amount of area sampled.

If $M =$ total area of interest, and
$m =$ the area actually sampled in area of interest
Then, $m/M =$ the proportion of the total area sampled

Thus:

If $C =$ our count of individuals in the area sampled
Then our population estimate $\hat{N} = C/(m/M)$

Detectability:

As we saw in lab last week, it is very important when you are sampling populations to account for individuals present, but not detected. We do this by estimating detection probabilities for the parameters of interest. We will illustrate this with three common methods of estimating detection probabilities with avian count data:

1) Distance Sampling – estimates detection probability as a function of distance from the observer. The method is applicable to both transect and point sampling.
2) Independent Observer Method – uses mark-recapture theory to estimate detection probability for observer conducting point counts.
3) Patch Occupancy Models – estimate the probability that a patch is occupied by a given species. The method is applicable to a variety of sampling methods.

Sample Size Requirements:

All of the methods that we use are statistical in nature, so it is important to ensure that your sample sizes are sufficient to make statistically valid inferences about your population. Sample size requirements will vary depending on your objectives, hypotheses, and sampling methods (see references), but generally speaking, a sample of at least 30 is required for testing most hypotheses. This requirement is often difficult to meet in ecological studies, thus it is very important to consider sample size requirements when you are formulating the objectives of your study. When it is not feasible to obtain sufficient independent samples there are several options to consider:

1) Use representative areas or habitats that are larger, but have the same characteristics as the habitat or area of interest.
2) Relax statistical assumptions (e.g. independence amongst sampling units), but acknowledge how this compromise might have weakened your conclusions.
3) Change your study plan to address a more feasible question.

**Methods Of Estimating Detection Probability:**

**Distance Sampling:**

Distance sampling is a unified approach that incorporates sampling design, data collection and statistical theory to estimate density and detection probability as a function of the distance between observers and the animals they are counting (Buckland et al. 2001). Distance sampling is used for line transects, point transects (point counts), cluster counts or cue counts (e.g. marine mammals breaching the water). The probability of detection is modeled by the function \( g(r) \) where \( r \) is the radial distance of the subject of interest from the observer, such that as distance from the observer increases the detection probability decreases (Fig. 1).

![General detection function, \( g(r) \), where detection probability decreases as distance increases.](image)

**Figure 1.** General detection function, \( g(r) \), where detection probability decreases as distance increases.

Statistical inference based on distance sampling relies on three basic assumptions:

1. Objects directly on the line or point are always detected (they are detected with probability 1, or \( g(0) = 1 \)).
2. Objects are detected at their initial location, prior to any movement in response to the observer.
3. Distances are measured accurately.

In addition to these three explicit assumptions, there are several quasi-assumptions that relate study design to the effectiveness of distance sampling:

1) the survey should follow a probability based sampling plan such as random sampling or systematic random sampling,
2) the point locations should represent the area of interest and therefore be located in homogenous, uniform habitat away from edges and other sampling points,

3) the objects of interest are spatially distributed in the study area according to a natural process.

Program Distance (Thomas et al. 1997) was developed to analyze data collected from distance sampling. Program Distance estimates density as:

\[ \hat{D} = \frac{n}{k \pi r^2 \hat{p}} \]

where \( \hat{D} \) is the density estimate, \( n \) is the number of objects counted, \( \pi r^2 \) is the area of the point, \( k \) is the number of points and \( \hat{p} \) is the estimated detection probability. The detection probability is modeled by using one of four possible functions, called "key functions." These functions (uniform, half-normal, negative exponential, and hazard-rate) are simply different models of how detection probability changes with distance from an observer (different shapes of the curve in Figure 1). The uniform model represents a constant detection probability of 1 from the center of the point to the effective detection radius. The half-normal is simply half of the bell-shaped curve from the standard normal distribution (Fig. 1). The negative exponential and hazard-rate models are similar to the half-normal except that each has a different slope and “shoulder” length. The “shoulder” refers to the portion of the curve where the detection probability is constant with increasing distance from the point. Beyond the “shoulder”, detection probability decreases according to the characteristics of the detection function. Key function selection is based on a visual examination of a histogram of the data.

After a key function is selected, a series expansion is used to add flexibility to the model by increasing the number of estimated parameters. The number of parameters needed for the series expansion varies for each dataset. Three series expansions are available; the cosine, simple polynomial, and the hermite polynomial. All three expansions have different properties, and each modify the key function in different ways to obtain a better fit of the model to the data. A large amount of literature exists describing the properties of the key functions in combination with different series expansions; see Buckland et al. (2001) for further information.

Model selection in Program Distance is based on an information-theoretic approach (Burnham and Anderson 2002) using Akaike’s Information Criterion, AIC. A model is selected from an a priori set of candidate models to optimize the balance between precision and bias (see Burnham and Anderson 2002).

Double Independent Observer Method:

Mark-recapture methods provide another tool for estimating detection probabilities (Williams et al. 2001). Traditionally, animals are caught, marked with a unique mark, and recaptured at a later date. Resampling the population one or more times generates a capture history for each animal that can be used to estimates overall detection (i.e. capture) probabilities.

General theory of mark-recapture:

If

\[ n_1 = \text{Total # of animals observed and marked in the first sample} \]
n_2 = Total # animals observed in the second sample
m_2 = Total # of marked animals in the second sample (new captures)
N = population size

Then

\[
m_2/n_2 \text{ is approximately equal to } n_1/N
\]

(sample) \quad = \quad (population)

And the estimate of the population size (N) is \( (n_1 \cdot n_2) / m_2 \)

This is called a Lincoln-Peterson closed population estimate and it relies on three basic assumptions:

1) the population is closed (no births, deaths, immigration, emigration)
2) all the animals in the population have an equal probability of capture
3) no animals lose their marks

Mark-recapture concepts can be applied to avian point counts to provide estimates of population size and detection probability (p). Birds are not physically marked as they are in traditional mark-recapture studies. They are considered "marked" when they are mapped by one observer, and "recaptured" when they are also mapped by another observer. One approach is called the independent double-observer method. It requires two observers to identify and map the locations of all the birds they hear within a fixed radius (usually 100 – 200 meters) during a 3 – 5 minute count. It is important that each observer is independent and not influenced by the other observer. At the end of the count the two observers compare their mapped locations of birds detected during the count, and they match up individual birds that both observers identified (equivalent to \( m_2 \) in the previous example). The method assumes that observers correctly match individual birds that they both detected.

For each species:

n_1 = The number of individuals seen by the first observer
n_2 = The number of individuals seen by the second observer
m_2 = The number of individuals seen by both observers

The estimated population size (Chapman’s modification) within the fixed radius is;

\[
\hat{N} = \frac{((n_1+1)\cdot(n_2+1))}{m_2+1} - 1
\]

and the probability of detection for each observer is

\[
\hat{p}_1 = m_2/n_2 \quad \text{and} \quad \hat{p}_2 = m_2/n_1
\]
Notice that the detection probability is already incorporated into the estimate of \( N \) so it is not necessary to adjust the estimate by the detection probability.

It is important to also present a measure of variation represented by the Variance of \( N \).

\[
\hat{\sigma}^2(\hat{N}) = \frac{(n_1+1)*(n_2+1)*(n_1-m_2)*(n_2-m_2)}{(m_2+1)^2 * (m_2+2)}
\]

We can use this information to compare detection probabilities for different species among observers and to estimate population size. If multiple points are sampled, the population estimates (\( N \)) for each point are added together to estimate the total population size for the study area. The variance estimates are also added together to obtain the total variance in the estimate of \( N \). Remember to multiply the total estimate of \( N \) by the sampling proportion to obtain an estimate of the total population in your study area. To obtain a density estimate divide by the total area, \( D = \frac{N}{A} \). For example: if four 200 meter fixed radius point counts are conducted producing population estimates of 20, 19, 21, and 22, the total population estimate is 82 and the estimated density is 82 divided by the total area of the study area (\( 4 \pi *200^2 = 502654.82 \) meters ). In this case density would be \( \frac{82}{502654.82} = 1.631 \times 10^{-4} \) birds per square meter (\( 1.631 \times 10^{-8} \) birds per hectare).

The assumptions of this method are:

1) The population is closed.
2) There are no matching errors.
3) The observers are independent.

It is also possible to use this method for species richness where:

\( n_1 = \) The number of species identified by observer 1
\( n_2 = \) The number of species identified by observer 2
\( m_2 = \) The number of species identified by both observers

You can then use the same estimates for \( N \) and detection probability except that \( N \) will represent a measure of species richness and not population size.

**Patch Occupancy:**

Patch occupancy models are useful when you are interested in the proportion of the sample area that is occupied by a particular species or the temporal dynamics of species occurrence (Mackenzie et al. 2006). Occupancy modeling relies on mark-recapture theory by accounting for a species present in a patch, but not detected (detection probability) Patch occupancy modeling is simpler than abundance modeling because it is based on simple presence/absence data. Because the data contain less information, inferences from patch occupancy modeling are usually more limited than those from models based on abundance estimates. We will not go into the details of

**Important References:**


**Important websites with software:**

There has been a considerable amount of recent work on software to perform population analysis, particularly in terms of estimation of abundance, and both survival and recruitment rates (using both capture-recapture and recovery models). The following is a list of websites that contain available programs and related information.


[http://www.warnercnr.colostate.edu/~gwhite/software.html](http://www.warnercnr.colostate.edu/~gwhite/software.html)