Lecture 1 - Key

Introduction to Sampling

Sampling is a process that selects a part of the population (via some mechanism or plan) for observation. Typically, the goal is to estimate one of more characteristics about the population based on information contained in the sample.

Two important sampling questions:

1. What is the best way to obtain a sample?

2. How do we use the information contained in the sample to estimate the population characteristics of interest?

Question (1) is the *design* problem and (2) is the *analysis* problem. The design problem involves questions about the method of selecting sampling units, the sample size, the information to be recorded for each sampling unit, and the observational methods used to collect data from a sampling unit.

Example: Suppose the goal is to find the average age of students enrolled at MSU.

- Can we use the collected ages from our course survey, why or why not?

- Now devise a strategy to collect this information. Consider how students will be selected, how many students to question, and how to ask students their age.

By properly addressing the design problem, the analysis problem will often not be difficult. In other words, given the design, the researcher should know the format of the analysis prior to data collection.

Populations and Sampling Units

A **survey** is a sample selected from a finite population of interest to represent the whole population.

A **census** is a complete enumeration where every member of the population is observed.

An **observation unit** is the object on which a measurement is taken, with humans the observation unit is typically referred to as an individual.

The **target population** is the set of individuals about which information is desired (i.e., the set of individuals the researcher would like to study).

The **sampled population** or study population is the set of individuals a researcher intends to study (i.e., the set of individuals that could possible included in the sample).

Now back to the age of MSU students. What is the target population? Discuss a situation where the target population does not match the sampled population.

A random sample of students attending the university who are currently on the MSU campus is taken. If students happen to be studying abroad, then the sampled population does not match the target population.

It is certainly desirable for the target and study populations to match. However, this is often not the case. When they do not match, it may not be possible to make statements about the target population from data collected on the study population. That is the scope of inferences or conclusions will be restricted to the study population.

Our responses to the course survey will not extend to the entire MSU community as a whole.

Similarly the *scope of inference* is restricted to the study population in our hypothetical data collection studies.

The potential members or units of a sample are the *sampling units*.

A **sampling frame** is a complete specification of the sampling units from a population.

Consider three sampling schemes: one that calls dorm rooms and asks the person answering the phone their age, one administered in classes, and one that interviews students based on a master list from the registrar. What are the sampling units and sampling frame in each situation?

Note: At this point we clarify the distinction between *individuals in a population* and *sampling units in the sampling frame.*

If the sampling frame consists of individuals in the study population then the sampling units and the individuals in the study population are the same.

However, a sampling plan could consist of sampling subgroups of individuals. In this case, the sampling units are subgroups of individuals. For example, it is common to sample households (which often contain more than one person per household). The household is the sampling unit and the set of all households forms the sampling frame. This is an example of 'cluster sampling.'

For many populations, the sampling unit is obvious. It is necessary to conceptually form the sampling frame of population units. Often it is necessary to record additional information (e.g. age, gender,. . .) that allows different classification of sampling units (for 'stratification').

For other populations, the sampling unit may not be obvious. For example, when surveying a geographical region, you may have to use a map to identify what is the basic sampling unit. Consider an experiment designed to assess the effect of pine beetle on Montana forests, how might you conduct this study?

This type of sampling introduces a number of problems, namely:

1. The numerous alternative size and shape combinations for a sampling unit.

2. A discrepancy between a sampling frame and the researcher's inability to access certain sampling units (e.g., too costly, remote, or dangerous to access).

3. A complete list of units or classification information is not available.

Estimates vs. Estimators

The goal of sampling is to make conclusions about some characteristics of interest for one or more populations of interest based on the data collected. This process of making conclusions is called **statistical inference.**

Using our class data on the number of hours spent at Bridger Bowl this winter answer the following questions:

1. What is the sampled population in this case?

2. What is the sampling frame for this problem?

3. How would you estimate the average number of hours for the class?

A **parameter** is a value which describes some characteristics of a population (or possibly describes the entire population). Examples: the population mean μ or the population variance σ^2.

A **statistic** is a value that can be computed from the data without knowing the values of any parameters.

In general, the value of the population parameter is unknown. Statistics computed from data can provide information about the unknown parameter.

The process of estimating a population parameter by a statistic derived from survey or experimental data is called **point estimation.** However, as statisticians and scientists, I'd encourage you to think 'distributionally.'

Prior to data collection, a sample statistic is a random variable and is called a *point estimator* of a parameter. For example $\bar{Y} = \frac{Y_1+Y_2+...Y_n}{n}$ is a point estimator for μ where $Y_1, Y_2, ..., Y_n$ are random variables.

After collecting a sample and conditional on the data, a sample statistic is no longer a random variable but is a realization of the point estimator and is called a *point estimate* of the population parameter. For example, $\hat{y} = \frac{y_1+y_2+...+y_n}{n}$ is called a point estimate for μ where $y_1, y_2, ..., y_n$ are observed data values.

Using data from our class survey, what is the point estimator and point estimate for the number of hours spent at Bridger Bowl this winter?

Later in the course we will discuss interval estimation in the form of confidence intervals to express uncertainty in point estimates.

The researcher's sampling goal is to collect a sample that is *representative*. Specifically parameters of interest can be estimated from the sample with a known degree of accuracy and precision.

Accuracy is related to *bias* and precision is related to *variability*.

Sampling and Estimation Concepts

In one of the most common sampling situations, we assume the population consists of a finite number N of sampling units. The units in the population are identifiable and can be labeled $1, 2, ..., N$.

Associated with each of the N units is a measurable value related to the population characteristic of interest (often referred to as y).

Each y-value is considered a fixed quantity representing that unit. In other words, we assume the sequence of population y-values $(y_1, y_2, ..., y_N)$ is fixed.

For each sampled unit there will be a unit label, its y-value, and any other recorded and potentially useful auxiliary variables (e.g., elevation, temperature, precipitation when studying environmental or ecological problems, or, age, income and gender when studying human populations.)

A **sampling design** is the procedure by which a sample of units is selected from the population.

The classical sampling designs (e.g., simple random, stratified, cluster, systematic) require that randomness be built into the sampling design so that its estimators can be assessed *probabilistically.* For example, we can make statements like `Our estimate is unbiased'' or`We are 95% confident that our estimate will be within 2 percent of the true population.

Sampling designs that are based on planned randomness are called **probability samples.** More formally, the design is determined by assigning to every possible sample $\mathcal{S}$ a sample probability $P(\mathcal{S})$ that equals the probability of actually selecting the sample.

When taking a simple random sample (SRS) of size n, the possible samples consist of n distinct units selected from the population of N units, and $P(\mathcal{S})$ is the same for every possible sample $\mathcal{S}$. Thus $P(\mathcal{S})= 1/$(the total number of unique samples of size n)

Sketch out pseudocode to sample 5 members of the class to give their course presentations on Friday.

The typical inference problems in sampling are

1. the estimation of some population characteristic based only on the sample data (point estimation)

2. an assessment of the variability associated with estimates.

This variability assessment is often an **interval estimate** expressed in terms of a confidence interval.

Ideally, we would like a sampling strategy which will yield samples that produce estimates with small variability that are centered around the true value. In other words, we want `high precision' and`high accuracy' (or little or no bias).

Thus, by choosing an appropriate sampling design and estimation method, the researcher can often obtain unbiased estimates without making additional assumptions about the population.

Selection by use of probability samples removes intentional or unintentional human sources of bias (such as the tendency to select units with larger or smaller than average values). Use of probability samples to generate a representative sample is especially desirable when there are parties with conflicting interests (e.g., a fish study that will be used by fishery management, commercial users, and environmental groups).