

Stat 549 Applied Sampling - Overview

The intent of this handout is to provide a brief overview to some of the topics in applied sampling which will be covered this semester.

What is sampling?

The purpose of sampling is to estimate some unknown population parameter from a subset (sample) of the population and to estimate the accuracy of the estimate. An estimate without a measure of accuracy is of limited use.

Definitions:

- Parameter - any numerical characteristic of a population.
- Sampling unit - the basic unit in the population on which we record the variable(s) of interest. We get one value from each sampling unit.
- Population - the set of all sampling units

Example 1: we want to estimate the average length of grass blades on the oval.

Parameter:

Sampling unit:

Population:

Example 2: we want to estimate the number of grass blades on the oval.

Parameter:

Sampling unit:

Population:

Throughout most of this course, we assume a finite population of size N . The infinite population case can be treated by simply letting N be very large, so the finite population case is the most general setting.

Basic Sampling Designs

1. Census - sampling the entire population. This means that we know the parameter of interest, so that the use of statistics is not necessary.
2. Simple random sample (SRS) (n = sample size) - random selection of n sampling units without replacement.
3. Systematic random sample - selecting every m^{th} member of the population. Where does the “randomness” come in here?
4. Unequal probability sample - any sampling plan where some units have a higher probability of being chosen than others. Does this introduce bias?

- One type of unequal probability sampling is PPS sampling - sampling with Probability Proportional to Size.

5. Stratified random sampling - sampling where the population is broken into different strata, and an SRS is taken within each stratum.

- Stratified random sampling can be thought of as a two-stage sampling plan, where the stratification is one stage, and the SRS is the second stage. In fact, any of sampling designs 1-4 above could have been used at the second stage.

6. Cluster sampling - sampling where the population is first broken into clusters, and then an SRS of *clusters* is taken. Typically, every individual within a cluster is then sampled.

- When is cluster sampling appropriate?

7. Convenience or haphazard sampling

Two-Stage Sampling Plans

Example: UM faces the prospect of significant increases in tuition over the next biennium depending on what happens in the legislature. Suppose we want to sample University of Montana first-year students regarding their opinions on the tuition increases.

Population?

Sampling unit?

What are some possible parameters we might be interested in?

There are currently 9 freshman dormitories on campus. So we might take:

- An SRS of 4 dorms, and then
- An SRS of 10 students in each of these 4 dorms.

At each stage of a sample, we need to be able to identify the following:

1. What is the sampling unit?
2. What is the population?
3. What is the sampling plan?

Back to the Example - 1st Stage: Primary sampling unit?
Population?
Sampling Plan?
2nd Stage: Secondary sampling unit?
Population?
Sampling Plan?

Example: Suppose we want to estimate the number of people who enter the Griz Market in the UC over the 15 weeks of a semester, including weekends. Observers will monitor the entrance and count the number of people entering. For simplicity, suppose we decide that the sampling unit is a day and that we will monitor the entrance on a sample of 21 days during the semester.

SRS: Randomly select 21 days from the 105 days of the semester.

Stratified: Stratifying by day of the week, we randomly select 3 Sundays, 3 Mondays, etc. over the 15 week semester.

1st Stage: Primary sampling unit?
Population?
Sampling Plan?
2nd Stage: Secondary sampling unit?
Population?
Sampling Plan?

Cluster: Use the 15 weeks as clusters of 7 days each. Randomly select 3 weeks and monitor the entrance during every day of each of these weeks.

1st Stage: Primary sampling unit?
Population?
Sampling Plan?
2nd Stage: Secondary sampling unit?
Population?
Sampling Plan?

There are many other possible designs: we could do a two-stage design with an SRS at each stage: e.g., an SRS of 7 weeks with an SRS of 3 days within each week. If a day is considered too long a period to monitor and we want the sampling unit to be an hour (or a longer period), we could add a third stage where we sample hours within each day. Such designs are common when attempting to estimate visitation to recreation sites, for example. These designs can also be used as the initial stages of a design to select a sample of visitors and the parameter(s) of interest are characteristics of all visitors to a site.

What are the relative advantages/disadvantages of the above designs in terms of cost, convenience and accuracy? (These issues will be addressed as we discuss these designs).

Use of Auxiliary Information

- Auxiliary information generally consists of other variables (other than the response variable) we can measure on the sampling units which help in the estimation of some parameter of interest. The hope is that this auxiliary information is related to the variable of interest.
 - These auxiliary variables can be utilized in either the design phase or in the estimation phase of the sampling scheme. Common uses of auxiliary information are given below.
1. Ratio or regression estimation - we use the relationship between an auxiliary and response variable in the estimation phase. Any examples?
 2. Double sampling - This consists of taking an SRS of n sampling units and measuring the auxiliary variable, and then taking a subsample of these n units to measure the response variable. This is done generally because the response variable may be difficult or expensive to measure, whereas the auxiliary variable may be easier to measure. Any examples?
 3. Stratified random sampling - With stratified random sampling, the strata themselves are the auxiliary variable.
 - Here, auxiliary information is incorporated in the “design phase” of the sample, so that the auxiliary variable affects how we *sample* instead of how we estimate.
 4. Ranked set sampling - Here, we take groups of sampling units, visually rank the units within each group according to the size of the response variable, and then randomly choose groups where we sample the unit of rank 1 in the first group chosen, rank 2 in the second group chosen, etc.
 5. Spatial sampling - sampling method which incorporates spatial proximity information into the sampling plan.
 - The main idea here is that since variables tend to be more alike in value the closer they are to one another spatially, optimal sampling plans purposely avoid sampling units which are spatially “close.” Sampling two units close together provides redundant information if there is spatial correlation present.

6. Adaptive sampling - sampling method generally used in sampling rare or “difficult to find” species.
- As an example, consider sampling some rare plant species in some area to determine how much of that species there is in the area.
 - We could take an SRS of say, 1 meter square areas, but might miss plants we can see in areas not scheduled for sampling.
 - Adaptive sampling allows you to “adapt” your sampling plan to account for what you actually see, by sampling these units with the rare species and all units in some vicinity of these units. Does this introduce bias?
- Note: In these final two sampling strategies (spatial, adaptive), we embrace two of the most troublesome facts of life in statistics: DEPENDENCE and BIAS!