

Lecture 5: Sampling Methods

What is sampling?

Is the process of selecting part of a larger group of participants with the intent of generalizing the results from the smaller group, called the <u>sample</u>, to the population. If we are to make valid inferences about the population, we must select the sample so that it is representative of the total population.

Why sampling?

It is not possible to collect data about every member of a population since the population is often too large.

Time, cost and effort constraints do not allow for collection of data about every member of a population.

It is NOT necessary to collect data about all members of a population because valid and reliable generalizations can be made about a population from a sample using inferential statistics.

Sampling Definitions

- Universe: Theoretical and hypothetical aggregation of all elements in a survey. Eg, Americans.
- Population: A more specified theoretical aggregation. Eg, Adult Americans in spring 2001.
- Target Population: The population to which the researcher would like to generalize his or her results.
- Sampling Frame: Actual list of units from which the sample is drawn.
- Sample: A subset of the target population chosen so as to be representative of that population.
- Sampling Unit:: Elements or set of elements considered for sampling. Eg, persons, geographical clusters, churches.

Sampling Definitions

Specified theoretical aggregation of individuals, groups, or events that you are interested in studying

Who you want to generalize to

Population you have access to

Who is in your study



Sampling stages

Identify the target Population	Reconcile
Determine the sampling frame	frame & population
Select a sampling procedure	
Determine the sample size	
Select actual sampling units	
Collect data from respondents	Deal with
Information for decision-making	nonresponse

Types of Sampling

Probability sampling: involves the selection of participants in a way that it is not biased. Every participant or element of the population has a known probability of being chosen to be a member of the sample.

Nonprobability sampling: there is no way of estimating the probability that each participant has of being included in the sample. Therefore bias is usually introduced.

Probability Sampling

When probability sampling is used, inferential statistics enables researchers to estimate the extent to which results from the sample are likely to differ from what we would have found by studying the entire population.

- Simple random sampling
- Systematic random sampling
- Stratified random sampling
- Cluster random sampling

Probability Sampling Methods

Probability sampling methods are those in which the chance that a case is selected is known in advance. Cases are selected by chance, and therefor there is no *systematic bias* in the characteristics of the sample.

Even without systematic bias in the way in which a sample is selected, the sample will still be subject to *sampling error* due to chance.

In general, the amount of sampling error is due to the size of the sample (the larger the sample, the less sampling error) and the homogeneity of the population (if every case is the same, you will not get any sampling error).

Probability Sampling

Simple Random - Each member in the sampling frame has an equal probability of being selected.

Systematic - Every kth member in the sampling frame is chosen.

Stratified - Members in the sampling frame are grouped by forming classes or strata, thereafter selecting a simple random sample from EACH class or stratum.

Cluster - Members in the sampling frame are grouped by forming classes or strata, thereafter simply electing to study all or a sample of members of only some (BUT NOT ALL) of the classes or stratum

A simple random sample (SRS)

A *simple random sample (SRS)* is one in which cases are selected from the population based strictly on chance, and in which the probability of selection is the same for every case in the population.

If there are N cases in the sampling frame (i.e., the population) and n cases are to be selected for the sample, then every case has an n/N chance of being selected.

Simple random sampling is the most straightforward probability sampling techniques, but it does not work well for all research purposes.

For example, imagine that you want to compare Native Americans and Whites in terms of some attribute. If you construct an SRS of 500 people, will you get enough Native Americans in the sample?

Homogeneity and Heterogeneity

- Homogeneous: Having few differences.
 Example: A group of 30-year-old male USA second-year university students majoring in Sociology is a *homogeneous* group.
- Heterogeneous: Having many differences.

- Example: A group of international students.

Cluster Sampling

- A sampling technique in which the population is divided into existing groupings (clusters).
- Then a sample of the cluster is selected randomly from the population.
- The term cluster refers to a natural, heterogeneous, intact grouping of the members of the population.
- The most common variables used in the clustering population are geographical areas, buildings, schools.
- Heterogeneity of the cluster is an important feature of an ideal cluster sample design.
- Sub-types of cluster sampling:
 - Single-stage cluster sampling
 - Two-stage cluster sampling
 - Multistage cluster sampling

Stratified Sampling

- Population is bifurcated into various mutually exclusive, homogeneous and nonoverlapping subgroups known as strata.
- The sample is drawn randomly from each subgroup or stratum
- The sub-types of stratified sampling are:
 - Proportionate Stratified Sampling
 - Disproportionate Stratified Sampling

Stratum refers to a single subgroup or category and Strata refers to several, or all, groups

Stratified Sampling

- Increases efficiency by limiting population variance or increasing sampling accuracy
- Stratifying (grouping) variable such that
 - it is strongly associated with the key dependent variable
 - each group is more homogeneous than the population in terms of the key dependent variable
- Sample size for each stratum will depend on
 - variance in the stratum
 - relative cost of sampling from the stratum
- Proportional (to size)stratified sampling
- Disproportional stratified sampling

Cluster Sampling

- Cluster Sample: a probability sample in which each sampling unit is a collection of elements.
- Effective under the following conditions:
 - A good sampling frame is not available or costly, while a frame listing clusters is easily obtained
 - The cost of obtaining observations increases as the distance separating the elements increases
- Clusters are locations within which an intact group of members of the population can be found
 - Examples
 - Neighborhoods
 - School districts
 - Schools
 - Classrooms

Cluster Sampling

- Increases efficiency by reducing cost
- 2 step process
 - form clusters (each cluster is as heterogeneous as the population)
 - randomly select some clusters (probability proportional to size), collect data from all elements of the selected clusters
- Often used in <u>area sampling</u>

Stratified vs. Cluster sampling:

- Stratified:
 - Subgroups defined by variables under study.
 - Strive for
 homogeneity within
 subgroups,
 heterogeneity
 between subgroups.
 - Randomly choose elements from within each subgroup.

- Cluster:
 - Subgroups defined
 by ease of data
 collection.
 - Strive for heterogeneity within subgroups, homogeneity between subgroups.
 - Randomly choose a number of subgroups, which we study in their entirety.

Nonprobability Sampling

the probability of any population element being selected for the sample is not known a priori:

- convenience sampling
- judgement sampling
- quota sampling
- snowball sampling

Quota Sampling

- <u>Nonprobabilistic</u> attempt to get representation of major sub-groups in the sample
- sub-groups usually defined in terms of easily identifiable demographics
- may match marginal frequencies but difficult to match joint frequencies
- selection of quota basis should have a relationship with the variable of interest
- similar to stratified sampling, but different
 - quota sampling is judgement or convenience based and not probabilistic

Generalizability and Statistics

- Results with non-probability samples are USUALLY not generalizable.
- If generalizability is not required, then <u>neither</u> are statistical tests of significance!
- Important point: If the sample doesn't represent the population, then statistical tests are NOT NECESSARY and may be misleading.

Sample Size

- The larger the sample size, the higher the confidence (smaller values of p)
- The larger the variance (difference amongst subjects), the lower the confidence.
- Larger variance requires a larger sample size to achieve the same level of confidence.

External validity

- The extent to which the results obtained in a given study would also be obtained for different participants and under different circumstances (real world)
- Dependent upon sampling procedure
- Improve primarily by <u>replication</u> under different conditions.

Threats to external validity

Population validity. To what population can you generalize the results of the study? This will depend on the makeup of the people in your sample and how they were chosen. If you have used young, white middle class, above average intelligence, Western students as your subjects, can you really say your results apply to all people?

Ecological validity. Laboratory studies are necessarily artificial. Many variables are controlled and the situation is contrived. A study has higher ecological validity if it generalizes beyond the laboratory to more realistic field settings.

External Validity

POPULATION

- Representativeness of accessible population
- Adequacy of sampling method
- Adequacy of response rate

Medium	High
Some attempt to obtain good sample	Actual sample representative of the population
ECOLOGY Naturalness of setting or condition Adequacy of rapport with observent Naturalness of procedures or task Appropriateness of timing	ons ver sks
Medium	High
Somehow artificial setting, tester, procedure, time	Natural setting, tester procedure, time
	Medium Some attempt to obtain good sample ECOLOGY Naturalness of setting or condition Adequacy of rapport with observe Naturalness of procedures or task Appropriateness of timing Medium Somehow artificial setting, tester, procedure, time



Randomization methods: background

- Most hypotheses generate predictions which specify *certain patterns* in a data set.
- In such cases, the appropriate statistical null hypotheses specify a lack of pattern, i.e., randomness.
- Statistical hypothesis testing is then concerned with estimating the likelihood that the observed pattern is simply due to chance.

Randomization Tests

Randomization Testing: Method of testing hypotheses and sometimes also determining confidence intervals for parameters.

Randomization tests determine the significance level of a test statistic obtained from a set of data by comparing the statistic with randomization distribution.

What is a randomization distribution? It is a distribution of test statistic values that is obtained by randomly reordering the observed data values (without replacement) until all possible permutations are obtained.

Most hypotheses of interest are alternatives to a null hypothesis of randomness. A randomization test tells us how likely it is that a certain pattern in the data arose by chance.

Hypothesis Testing Significance thresholds based on Permutation test



Randomization Test

The basic approach in a randomization test is to calculate a test statistic from the observed data, and then randomly reshuffle the data a large number of times, recalculating the test statistic for each iteration.

Randomization Test

These statistics are used to generate a distribution of values, and the observed value is compared to the distribution to see whether the observed case is an event that was unlikely to have occurred through chance; that is, if it is a tail value from the distribution.



Figure 1. Graphical representation of a test of significance via randomization. The tails of the distribution of test statistics from randomized data provide the tail probabilities for significance tests (one-tailed or two-tailed, as desired).

Randomization Techniques

Randomization techniques allow the determination of the significance of an observed test statistic by comparing it to the distribution of values obtained by randomly reordering the data.

We ask: Is the observed value an unusually large or small value compared to what might occur by chance?

This is typically done in one of two ways:

1) In some cases, random data values (e.g., spatial positions or other measurements on individual subjects) are generated as random values.

2) In comparisons between groups, only the group memberships are randomized while the same set of measurements are maintained.

These tests sometimes are referred to as permutation procedures (the randomization is done by reordering the positions of elements in a array).

Randomization Test

How to do a randomization test:

- 1. Identify hypotheses (null and alternate).
- 2. Choose a test statistic.
- 3. Compute the test statistic for the original data.
- 4. Resample. Recompute the statistic for each of the 'new samples.' When you stop resampling, you will have a randomization distribution.
- 5. Accept or reject the null hypothesis by comparing the original test statistic to the randomization distribution.