

Sampling Theory

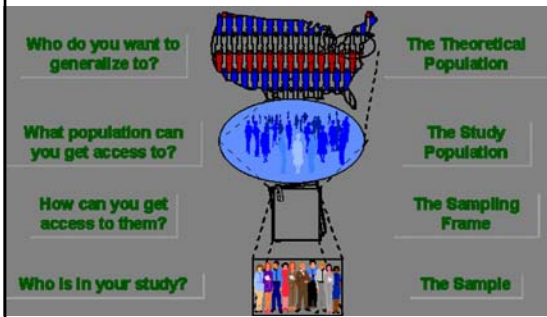
What you need to know (for now)

- **Why** we do sampling
- **How** we do sampling
 - Different methods of random selection
 - Different ways of organizing the population from which you select
 - Different methods of non-random selection
- **When** you would use a particular method
- You **do not** need to know *probability theory* – what are the rules for making generalizations from your sample data

Why Sample?

- Why not study everyone?
- Debate about Census vs. sampling

Key Sampling Concepts



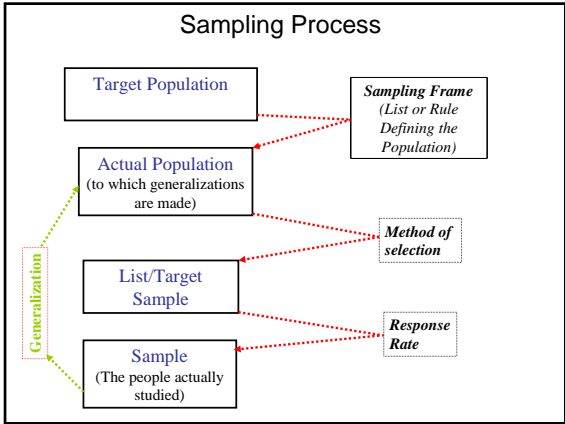
Copyright ©2002, William M.K. Trochim, All Rights Reserved

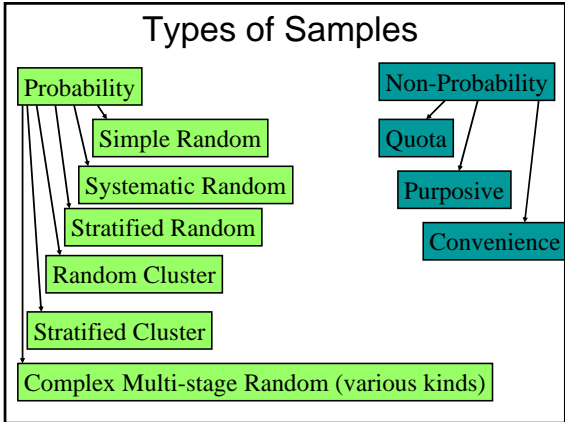
Key Ideas

- Distinction between the population of interest and the actual population defined by the sampling frame
- Generalizations can be made only to the **actual** population defined by the sampling frame
- Understand crucial role of the sampling frame

Sampling Frame

- The list or procedure defining the **SAMPLE POPULATION**. (The population from which the sample will be drawn.)
- It is **NOT** the actual sample.
- Examples:
 - Telephone book
 - Voter list
 - Random digit dialing
- Essential for probability sampling, but can also be defined for nonprobability sampling





Probability Samples

- A probability sample is one in which each element of the population has a **known non-zero** probability of selection.
- **Not** a probability sample if some elements of population cannot be selected (have zero probability)
- **Not** a probability sample if probabilities of selection are not known.

Probability Sampling

- Cannot guarantee “representativeness” on all traits of interest
- The key is to have a sampling plan with known statistical properties
- The properties of the sample will allow you to make certain inferences about the population

Example: Political candidate support

- **Sample finding:** 62 percent of voters in a random sample of 400 registered voters (polled on February 20, 2004) said that they favor John Kerry over George W. Bush for President in the 2004 Presidential election.
- **Generalization:** The probability is .99 that between 57 percent and 67 percent of all registered voters favor Kerry over Bush for President (at or around the time the poll was taken).

Sampling Frame is Crucial in Probability Sampling

- If the sampling frame is a poor fit to the population of interest, random sampling from that frame cannot fix the problem – not all elements have an equal chance of being chosen
- The sampling frame is non-randomly chosen. Elements not in the sampling frame have zero probability of selection.
- Generalizations can be made ONLY to the actual population defined by the sampling frame

Example: UW undergrads

- Your theoretical population is UW undergrads
 - Good sampling frames:
 - University enrollment database
 - Online directory?
 - Bad sampling frames:
 - Phone numbers on bathroom walls at Wando's
 - List of all people who live in dorms

Types of Probability Samples

Simple Random

Systematic Random

Stratified Random

Random Cluster

Stratified Cluster

Complex Multi-stage Random (various kinds)

} How you Select from a List

} How You Organize Your List, Before Selecting

Simple Random Sampling

- Each element in the population has an equal probability of selection AND each combination of elements has an equal probability of selection
 - E.g. Names drawn out of a hat
 - Random numbers to select elements from an ordered list (computer generated)
- This is how most sampling is done today



Systematic Random Sampling-1

- Used before the era of computers
- Each element has an equal probability of selection, but combinations of elements have different probabilities.
- Population size N , desired sample size n , sampling interval $k=N/n$.
- Randomly select a number j between 1 and k , sample element j and then every k^{th} element thereafter, $j+k$, $j+2k$, etc.
- Example: $N=64$, $n=8$, $k=64/8=8$. Random $j=3$. Choose element #s 3, 11, 19, 27, 35, 43, 51, 59



Systematic Random Sampling-2

Benefits:

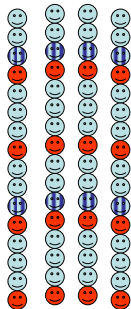
- Has same error rate as simple random sample if the list is in random or haphazard order
- If the list is grouped, provides the benefits of implicit **stratification** (sampling from pre-determined groups)



Systematic Random Sampling-3

Risks:

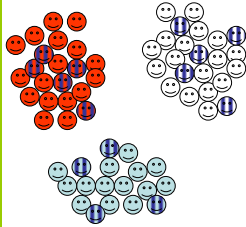
- Runs the risk of error if periodicity in the list matches the sampling interval
- This is rare.
- In this example, every 4th element is red, and red never gets sampled. If j had been 4 or 8, ONLY reds would be sampled.



Stratified Random Sampling-1

How to do it:

- Divide population into groups that differ in important ways
- People in each group are ALIKE on some characteristic
- Basis for grouping must be known before sampling
- Select random sample from within each group



Stratified Random Sampling-2

Benefits

- For a given sample size, reduces error compared to simple random sampling IF the groups really are different from each other
- Probabilities of selection may be different for different groups, as long as they are known
- Oversampling small groups improves inter-group comparisons

Costs

- Tradeoff between the cost of doing the stratification and smaller sample size needed for same error



Example: Amy's Dissertation Case

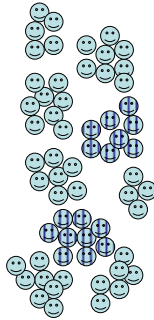
- Randomly-selected citizens making political decisions: how are they chosen?
- Sampling Frame: voter's list
 - Young people are underrepresented on that list
 - Divide into age categories before sampling, and then take the correct proportion from each (based on Census data)
 - Also stratified by gender and geography

When to Use Stratification

- When you know there are specific groups that you want to compare
- When you expect there to be real differences between groups
- When you are unlikely to get enough cases of the relevant characteristic from simple random sampling (ie. One group is small enough proportion of total population)

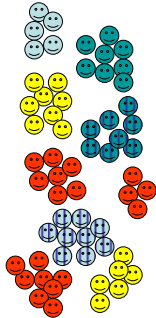
Random Cluster Sampling - 1

- Population is divided into groups
- Some of the groups are randomly selected
- For given sample size, a cluster sample has more error than a simple random sample
- Cost savings of clustering may permit larger sample
- Error is smaller if the clusters are **similar** to each other



Random Cluster Sampling - 2

- Cluster sampling has very high error if the clusters are different from each other – it is NOT desirable if the clusters are different
- It IS random sampling: you randomly choose the clusters
- But you will tend to omit some kinds of subjects



Example: NELS

- National Education Longitudinal Survey wants to be able to generalize to the entire US population of 8th graders
- There is no national list of middle schools
- States DO keep lists
- Chose some schools randomly from within **states**, choose some students randomly from within **schools**

When to Use Clustering

- It is difficult to put together a complete list of a population
- You want to save \$,
 - E.g. Face-to-face interviews – cheaper to do in one location
- There are naturally occurring groups that will resemble each other
 - E.g. states, counties, cities, blocks, schools

Stratification vs. Clustering

Stratification

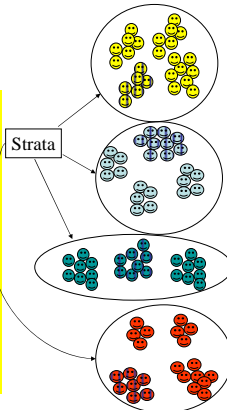
- Divide population into groups **different** from each other: sexes, races, ages
- Sample randomly within each group
- Less error compared to simple random
- More expensive to obtain stratification information before sampling

Clustering

- Divide population into **comparable** groups: schools, cities
- Randomly sample some of the groups
- More error compared to simple random
- Reduces costs to sample only some areas or organizations

Stratified Cluster Sampling

- Reduce the error in cluster sampling by creating strata of clusters
- Sample one or more clusters from each stratum
- Within each cluster, sample everyone, or do a simple random sample
- The cost-savings of clustering with the error reduction of stratification
- E.g. NELS



Multi-stage Probability Samples –1

- Large national probability samples involve several stages of stratified cluster sampling
- Used by many large national surveys
- *How to do it:*
 - The whole country is divided into geographic clusters, metropolitan and rural
 - Some large metropolitan areas are selected with certainty (certainty is a non-zero probability!)
 - Other areas are formed into strata of areas (e.g. middle-sized cities, rural counties); clusters are selected randomly from these strata

Multi-stage Probability Samples –2

- **EXAMPLE, cont'd:**
 - Within each sampled area, more sub-clusters are defined, and the process is repeated, perhaps several times, until blocks or telephone exchanges are selected
 - At the last step, households and individuals within household are randomly selected
- *Random samples make multiple call-backs to people not at home.*

The Problem of Non-Response - 1

- You can randomly pick elements from sampling frame and use them to randomly select people
- But you cannot make people respond
- Non-response destroys the generalizeability of the sample. You are generalizing to people who are willing to respond to surveys.
- If response is 90% or so, not so bad. But if it is 50%, this is a serious problem

Washington Post Survey Response Rates

http://www.washingtonpost.com/wp-srv/politics/polls/poll_response_rate.html

The Lengths to Which Some Researchers Will Go

1996 National Issues Convention in Austin, TX:

- Structured as a "Deliberative Poll": they wanted to be able to generalize from the sample of participants
- Initially surveyed 910 people about their political attitudes (out of 1260 contacted)
- Had to convince respondents to come to Austin by: finding babysitters, getting medical consultations, soothing fears of flying, convincing employers to give time off, finding people to milk cows!
- Still only had 466 people show up (response rate of 37%)

More common practices

- Multiple call-backs are essential for trying to reduce non-response bias
- Samples without call-backs have high bias: cannot really be considered random samples
- BUT : Response rates have been falling
- It is very difficult to get above a 60% response rate
- You do the best you can, and try to estimate the effect of the error by getting as much information as possible about the predictors of non-response.

Non-probability Samples

- Don't control for investigator bias in selection
- Don't know the odds of being included in the sample, so can't make the same kinds of predictions
- Types:
 - Convenience
 - Purposive
 - Quota

When to Use Non Probability Sampling

- When there are very few cases of the phenomenon of interest
- When doing historical research and data are not available
- In the early stages of research, when you are trying to figure out what you will investigate formally
- When your goal is to understand the range of variation/different dimensions of a problem

Convenience Sample

- Subjects selected because it is easy to access them.
- No reason tied to purposes of research.
- Students in your class, people on State Street, friends
- AKA “Haphazard Sampling”

Purposive Samples

- Subjects selected for a good reason tied to purposes of research
- Small samples < 30, not large enough for power of probability sampling.
 - Nature of research requires small sample
 - Choose subjects with appropriate variability in what you are studying
- Hard-to-get populations that cannot be found through screening general population

Quota Sampling

- Pre-plan number of subjects in specified categories (e.g. 100 men, 100 women)
- In uncontrolled quota sampling, the subjects chosen for those categories are a convenience sample, selected any way the interviewer chooses
- In controlled quota sampling, restrictions are imposed to limit interviewer’s choice
- No call-backs or other features to eliminate convenience factors in sample selection

Quota Vs Stratified Sampling

- In Stratified Sampling, selection of subject is **random**. Call-backs are used to get that particular subject.
- Stratified sampling without call-backs may not, in practice, be much different from quota sampling.
- In Quota Sampling, interviewer selects first available subject who meets criteria: is a **convenience sample**.
- Highly controlled quota sampling uses probability sampling down to the last block or telephone exchange

Sample Size

- **Heterogeneity**: need larger sample to study more diverse population
- **Desired precision**: need larger sample to get smaller error
- **Sampling design**: smaller if stratified, larger if cluster
- **Nature of analysis**: complex multivariate statistics need larger samples
- Accuracy of sample depends upon sample size, not ratio of sample to population

Sampling in Practice

- Often a non-random selection of basic sampling frame (city, organization etc.)
- Fit between sampling frame and research goals must be evaluated
- Sampling frame as a concept is relevant to all kinds of research (including nonprobability)
- Nonprobability sampling means you cannot generalize beyond the sample
- Probability sampling means you can generalize to the population defined by the sampling frame
