Lecture 1: Introduction to Sampling

Maochao Xu

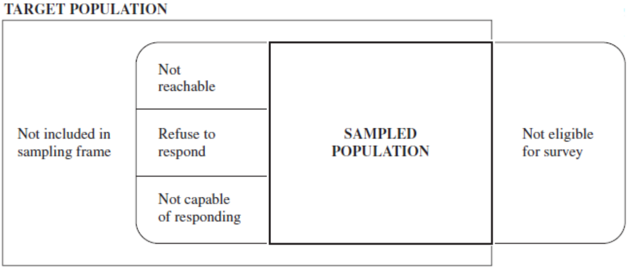
Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Ideally, a perfect sample would be a **scaled-down** version of the population, mirroring every characteristic of the whole population. A good sample will be representative in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.

1. **Observation unit** An object on which a measurement is taken. This is the basic unit of observation, sometimes called an element. In studying human populations, observation units are often individuals.
2. **Target population** The complete collection of observations we want to study. Defining the target population is an important and often difficult part of the study.
3. **Sampled population** The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.
4. **Sampling unit** A unit that can be selected for a sample. We may want to study individuals, but do not have a list of all individuals in the target population. Instead, households serve as the sampling units, and the observation units are the individuals living in the households.
5. **Sampling frame** A list, map, or other specification of sampling units in the population from which a sample may be selected. For a telephone survey, the sampling frame might be a list of all residential telephone numbers in the city. For a survey using in-person interviews, the sampling frame might be a list of all street addresses.

Example

Target population and sampled population in a telephone survey of likely voters. Not all households have telephones, so a number of persons in the target population of likely voters will not be associated with a telephone number in the sampling frame. In some households with telephones, the residents are not registered to vote and hence are not eligible for the survey. Some eligible persons in the sampling frame population do not respond because they cannot be contacted, some refuse to respond to the survey, and some may be ill and incapable of responding.



Selection Bias

Selection bias occurs when some part of the target population is not in the sampled population, or, more generally, when some population units are sampled at a different rate than intended by the investigator.

For example, if a survey designed to study household income omits transient persons, the estimates from the survey of the average or median household income are likely to be too large.

A sample of convenience is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the harder-to-select or nonresponding units.

The following examples indicate some ways in which selection bias can occur.

- Using a sample selection procedure that, unknown to the investigators, depends on some characteristic associated with the properties of interest.

  For example, investigators took a convenience sample of adolescents to study how frequently adolescents talk to their parents and teachers about AIDS. But adolescents willing to talk to the investigators about AIDS are probably also more likely to talk to other authority figures about AIDS. The investigators, who simply averaged the amounts of time that adolescents in the sample said they spent talking with their parents and teachers, probably overestimated the amount of communication occurring between parents and adolescents in the population.

- Deliberately or purposively selecting a **representative** sample.
- Misspecifying the target population.
- Failing to include all of the target population in the sampling frame, called **undercoverage**.
- Including population units in the sampling frame that are not in the target population, called **overcoverage**.
- Having multiplicity of listings in the sampling frame, without adjusting for the multiplicity in the analysis.

In its simplest form, random digit dialing prescribes selecting a random sample of 10-digit numbers. Households with more than one telephone line then have a higher chance of being selected in the sample.

- Substituting a convenient member of a population for a designated member who is not readily available.
- Failing to obtain responses from all of the chosen sample. **Nonresponse** distorts the results of many surveys, even surveys that are carefully designed to minimize other sources of selection bias.
- Allowing the sample to consist entirely of volunteers.

  Such is the case in radio and television call-in polls, and in most online surveys. The statistics from such surveys **cannot be trusted***. At best, they are entertainment; at worst, they mislead, particularly when statistics from polls with self-selected respondents are cited in policy debates without any mention of their unscientific nature.*

Measurement Error

When a response in the survey differs from the true value, **measurement error** has occurred.

- In a study area, the presence of a bird's nest is missed and 'absence' is recorded. This also happens when studying elusive populations.
- The measuring device is uncalibrated or its user is improperly trained.
- People lie when asked about sensitive issues.
- People give different answers to different interviewers.
- People may say what they think an interviewer wants to hear or what they think will impress the interviewer.

✓ Discussion-1

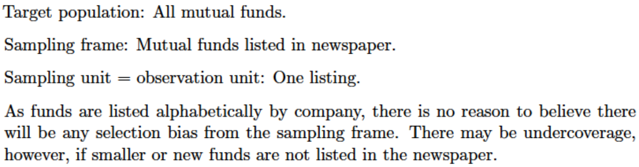
A student wants to estimate the percentage of mutual funds whose shares went up in price last week. She selects every tenth fund listing in the Mutual Fund pages of the newspaper, and calculates the percentage of those in which the share price increased.

Please specify the following items.

- Target population
- Sampling frame
- Sampling unit
- Observation unit

Discuss any possible sources of selection bias or inaccuracy of responses.

Answer:



Target population: All mutual funds.

Sampling frame: Mutual funds listed in newspaper.

Sampling unit = observation unit: One listing.

As funds are listed alphabetically by company, there is no reason to believe there will be any selection bias from the sampling frame. There may be undercoverage, however, if smaller or new funds are not listed in the newspaper.

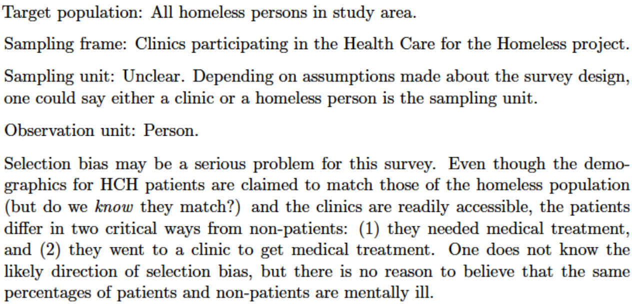
✓ Discussion-2

Many scholars and policy makers are interested in the proportion of homeless people who are mentally ill. Wright (1988) estimates that 33% of all homeless people are mentally ill, by sampling homeless persons who received medical attention from one of the clinics in the Health Care for the Homeless (HCH) project. He argues that selection bias is not a serious problem because the clinics were easily accessible to the homeless and because the demographic profiles of HCH clients were close to those of the general homeless population in each city in the sample. Do you agree?

Please also specify the following items.

- Target population
- Sampling frame
- Sampling unit
- Observation unit

Answer:



Target population: All homeless persons in study area.

Sampling frame: Clinics participating in the Health Care for the Homeless project.

Sampling unit: Unclear. Depending on assumptions made about the survey design, one could say either a clinic or a homeless person is the sampling unit.

Observation unit: Person.

Selection bias may be a serious problem for this survey. Even though the demographics for HCH patients are claimed to match those of the homeless population (but do we *know* they match?) and the clinics are readily accessible, the patients differ in two critical ways from non-patients: (1) they needed medical treatment, and (2) they went to a clinic to get medical treatment. One does not know the likely direction of selection bias, but there is no reason to believe that the same percentages of patients and non-patients are mentally ill.

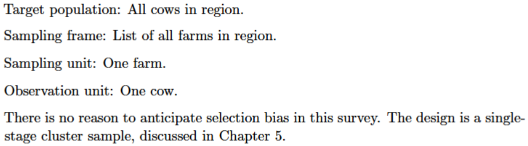
✓ Discussion-3

A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.

Please also specify the following items.

- Target population
- Sampling frame
- Sampling unit
- Observation unit

Answer:



Target population: All cows in region.

Sampling frame: List of all farms in region.

Sampling unit: One farm.

Observation unit: One cow.

There is no reason to anticipate selection bias in this survey. The design is a single-stage cluster sample, discussed in Chapter 5.

Sampling Errors

Sampling error is the error that results from taking one sample instead of examining the whole population.

Selection bias and measurement error are examples of **nonsampling errors**, which are any errors that cannot be attributed to the sample-to-sample variability.