CH2: Collecting Data

1. Statistical Methods: Descriptive vs Inferential
2. Basic Terminology
3. First Principle of Statistical Inference
4. Research Studies: Designed Experiment vs Observational Study
5. Sampling methods

1. Statistical Methods

1. **Descriptive Statistical Methods**: collect data and describe them.
2. **Inferential Statistical Methods**: collect data, analyze, interpret, and make conclusions based on the data.

   In this course, we will concentrate on inferential methods but the distinction may not always be clear.

2. Basic Terminology

- **Observation Unit**: The unit upon which data are collected.

  Ex: US Adult
- **Population:** Complete set of units of interest.

  Ex: All US Adults
- **Sample:** A subset of the population that is actually measured.

  Ex: 100 individuals selected based on random sample of SSNs.
- **Census:** when the sample equals the population

  Ex: US Census (?)

- **Variable:** information of interest about each individual item in a population.

  Ex: Height, Weight, Age, Gender
- **Statistic:** numerical descriptive measure for a sample.

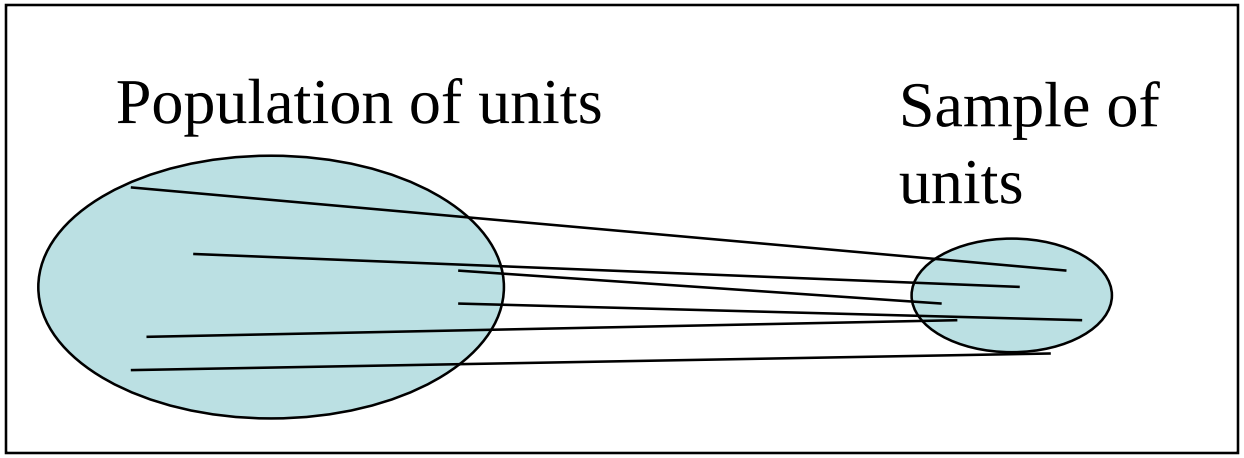
  Ex: Average height of 100 individuals in sample

  Note: Sample average is denoted $\bar{y}$ (y bar).
- **Parameter:** numerical descriptive measure for a population.

  Ex: Average height of all US adults

  Note: Population average is denoted μ (mu).

Return to Statistical Methods



- **Descriptive** statistical methods involve describing the sample.

  Ex: Describe the height values of the 100 people measured.
- **Inferential** statistical methods involve making statements about the population based on the sample.

  Ex: Make statements about the heights of the U.S. population based on the sample.

3. First Principle of Statistical Inference

You make inference to the population from which you sample.

Ex1: Researcher interested in cow grazing behavior. They observe 1 cow, 1 pasture, randomly select 50 one-hour intervals within a 4-week period.
Note: Unreplicated study – bad idea!!!

Ex2: Researcher interested in comparing wheat damage for three application levels of a particular pesticide. They perform a field trial where they inoculate several fields with one strain of pest and apply fertilizer from one batch.

Ex3: A company is interested in implementing a new manufacturing process. Researchers observes several lab runs of the new process. The objective is to find out how the process will work in large scale production.

Ex4: NIH funded study is aimed at developing improved treatments for breast cancer. BABL/c mice are randomly assigned to control and treatment groups and survival time of each mouse is recorded.

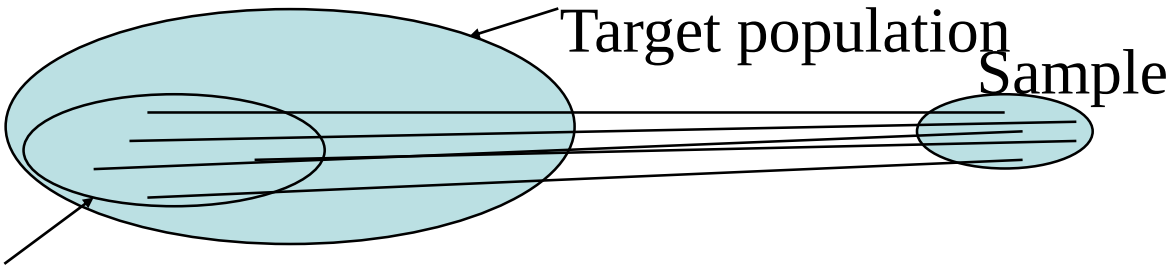
From Wikipedia (08/11/16):
"A **model organism** is a (non-human) species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the organism model will provide insight into the workings of other organisms."

Target vs Sampled Populations

- **Target Population**: the population you would like to sample.
- **Sampled Population**: the population you actually do sample.

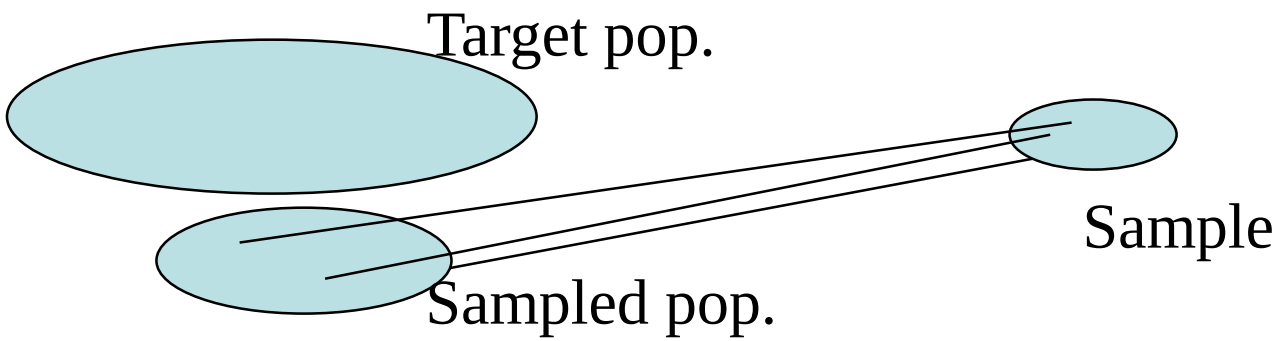
Often, the sampled population is a subset of the target population.

Ex: Pesticide trial (Ex2)



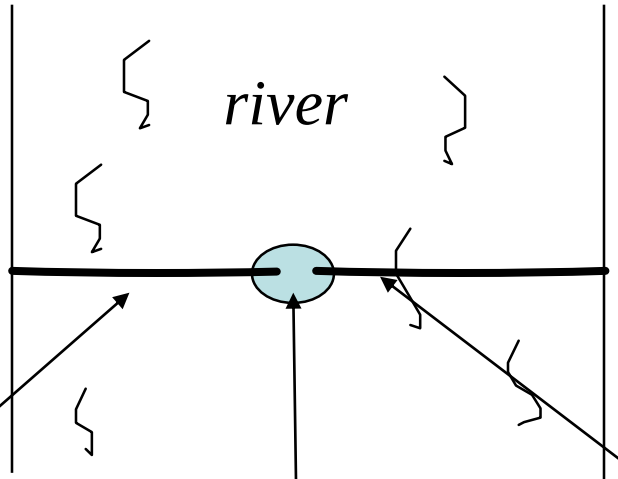
Other times, the Sampled population is not a subset of the Target population.

Ex: Model Organisms (Ex4)



You make <u>statistical inference</u> to the sampled population. You make <u>scientific judgments</u> about whether that is close enough to the target population to satisfy your objectives.

“River diagram” (Cornfield and Tukey)



Measurements taken on the sampled units (the data)

Scientific Objective (a conclusion about the target population)

Use Statistical Inference

Use Scientific Judgment

Statistical conclusions about the sampled population

4. Research Studies

Planning a research study:

1. Start with specific aims or research questions. **Specific aims drive the design and analysis.**
2. Identify the variables of interest.
   - Response variable is the focus of the research question.
   - Predictor variable may be associated with changes in the response.
3. Choose an appropriate design for the study.
   - Experiment vs Observational study
   - Sampling strategy
   - Other issues (including power) will be discussed
4. Collect the data.

Experiments:

Studies in which the researcher manipulates the experimental conditions or "treatments" experienced by the units (or subjects) using a randomization scheme.

Examples:

1. Researcher interested in comparing wheat yield (response) for three fertilizers (treatment/predictor) using a field trial. Fertilizers randomly assigned to plots.
2. Researcher interested in effect of vitamin E supplementation (treatment/predictor) on the rate of beef spoilage (response). Vitamin E or no Vitamin E randomly assigned to individual cows.

Observational studies:

Studies in which the researcher observes or measures, but doesn't manipulate (no randomization) the conditions experienced by the units (or subjects) and does not have control on which units receive which treatment or condition.

Examples:

1. Researcher interested in the association between dietary fat intake (predictor) and heart disease rate (response) in sample of countries.
2. Researcher interested in the association between smoking behavior (predictor) and lung cancer (response) in a population.

A limitation of observational studies is that the response may be affected by variables other than the predictor of interest. These "other" variables are called confounding variables. **Association is not causation!**

A benefit of experiments is that they allow researchers to draw causal conclusions. The reason is that by randomly assigning units to treatments other (confounding) variables should "average out".

In practice, both types of studies may be used together to make a stronger case.
For both types of studies, information about additional variables (covariates) can be collected and used in analysis. (But we can't measure everything.)

Types of observational studies:

1. A **sample survey** is a study that provides information about a population at a particular point in time.
2. A **prospective study** is a study that observes a population in the present and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes.
3. A **retrospective study** is a study that observes a population in the present and also collects information about the subjects in the sample regarding the occurrence of specific outcomes that have already taken place.

5. Sampling Methods

Sampling Frame: list of sampling units from which sampling is done.

1. **Simple Random Sampling**: select units randomly from the sampling frame. Units randomly selected from those remaining.
2. **Systematic Sampling**: take units from the sampling frame according to some method: e.g. every third individual, or every 50 meters along a transect.
3. **Stratified Random Sampling**: Organize the frame into groups (or strata) of like units. Sample independently within each stratum. The objective is to gain efficiency by sampling less intensively in strata that have low variability.

4. **Cluster Sampling:** First sample “clusters” of units, then measure every unit within each cluster in the sample. (Saves time locating units. Used in sampling units that exist in “colonies”)

Note: In stratified random sampling, we take a simple random sample within each group. In cluster sampling, we take a simple random sample of groups and then observe all items within the selected groups.