2.2 Sampling Design

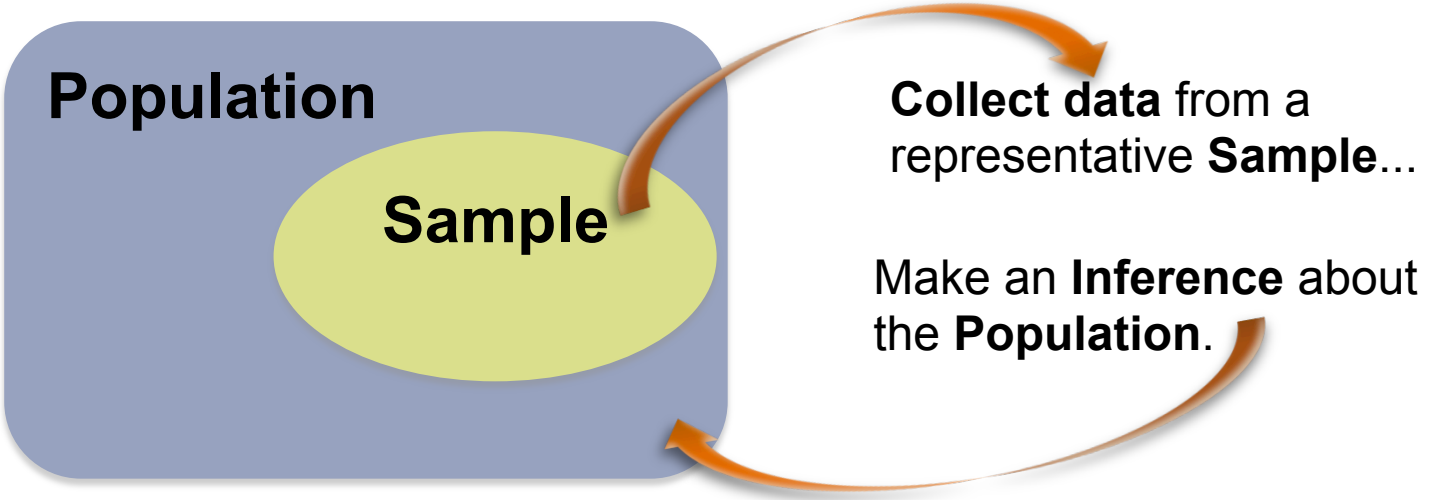
- Population and Sample
- Voluntary Response Sample
- Simple Random Sample
- Stratified Samples
- Undercoverage and Non-Response

Population and Sample

The distinction between population and sample is basic to statistics. To make sense of any sample result, you must know what population the sample represents.

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is the part of the population from which we actually collect information. We use information from a sample to draw conclusions about the entire population.



How to Sample Badly

The design of a sample is **biased** if it systematically favors certain outcomes.

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples show bias because people with strong opinions (often in the same direction) are most likely to respond.

Simple Random Samples

Random sampling, the use of chance to select a sample, is the central principle of statistical sampling.

A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected.

In practice, people use random numbers generated by a computer or calculator to choose samples. If you don't have technology handy, you can use a **table of random digits.**

How to Choose an SRS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these properties:

- Each entry in the table is equally likely to be any of the 10 digits 0–9.
- The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

How to Choose an SRS Using Table B

Step 1: Label. Give each member of the population a numerical label of the *same length.*

Step 2: Table. Read consecutive groups of digits of the appropriate length from Table B.

Your sample contains the individuals whose labels you match with the random digits.

SRS Example

Use the random digits provided to select an SRS of four hotels.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 Banyan Tree	11 Diplomat	18 Sandpiper	25 Tradewinds
05 Beach Castle	12 Holiday Inn	19 Sea Castle	26 Tropical Breeze
06 Best Western	13 Lime Tree	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda

69051 64817 87174 09517 84534 06489 87201 97245

Our SRS of four hotels for the editors to contact is: 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.

Other Sampling Designs

The basic idea of sampling is straightforward: Take an SRS from the population and use your sample results to gain information about the population.

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Sometimes, there are statistical advantages to using more complex sampling methods. One common alternative to an SRS involves sampling important groups (called strata) within the population separately. These "sub-samples" are combined to form one stratified random sample.

To select a **stratified random sample,** first classify the population into groups of similar individuals, called **strata.** Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Multistage Sampling Designs

Many national sampling designs are multistage:

divide the US into geographical areas that don't cross state lines

select a stratified sample that includes largest poplulations and others at random

then divide the selected areas into "blocks" – stratify using important variables and sample within the strata

cluster the housing units in each "block" and use a random sampling method to choose households.

Notice that in each stage, random sampling is the key to making the choices…

HW: Read sections 2.2 and 2.3, go over the summaries; try #2.46, 2.49, 2.51, 2.54, 2.59

Cautions about Sample Surveys

Good sampling technique includes the art of reducing all sources of error.

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

A systematic pattern of incorrect responses in a sample survey leads to **response bias.**

The **wording of questions** is the most important influence on the answers given to a sample survey.

Example: see this Gallup survey from October, 2012:

http://www.gallup.com/poll/165539/first-time-americans-favor-legalizing-marijuana.aspx

2.3 Ethics

- Basic Data Ethics
- Institutional Review Boards
- Informed Consent
- Confidentiality
- Clinical Trials
- Behavioral and Social Science Experiments

Basic Data Ethics

The most complex issues of data ethics arise when we collect data from people.

Basic Data Ethics

The organization that carries out the study must have an **institutional review board** that reviews all planned studies in advance in order to protect the subjects from possible harm.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential.** Only statistical summaries for groups of subjects may be made public.

Institutional Review Boards

- The organization that carries out the study must have an institutional review board that reviews all planned studies in advance in order to protect the subjects from possible harm.
- The purpose of an institutional review board is “to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities.”
- The institutional review board:
  - ✓ Reviews the plan of study
  - ✓ Can require changes
  - ✓ Reviews the consent form
  - ✓ Monitors progress at least once a year

Informed Consent

- All subjects must give their informed consent before data are collected.
- Subjects must be informed in advance about the nature of a study and any risk of harm it might bring.
- Subjects must then consent in writing.
- Who can't give informed consent?
  - ✓ Prison inmates
  - ✓ Very young children
  - ✓ People with mental disorders

Confidentiality

- All individual data must be kept confidential. Only statistical summaries may be made public.
- Confidentiality is not the same as **anonymity.** Anonymity prevents follow-ups to improve non-response or inform subjects of results.
- Separate the identity of the subjects from the rest of the data immediately!

Example: Citizens are required to give information to the government (tax returns, social security contributions). Some people feel that individuals should be able to forbid any other use of their data, even with all identification removed.

Clinical Trials

- Clinical trials study the effectiveness of medical treatments on actual patients—these treatments can harm as well as heal.
- Points for a discussion:
  - ✓ Randomized comparative experiments are the only way to see the true effects of new treatments.
  - ✓ Most benefits of clinical trials go to future patients. We must balance future benefits against present risks.
  - ✓ The interests of the subject must always prevail over the interests of science and society.
- In the 1930s, the Public Health Service Tuskegee study recruited 399 poor blacks with syphilis and 201 without the disease in order to observe how syphilis progressed without treatment. The Public Health Service prevented any treatment until word leaked out and forced an end to the study in the 1970s.

Behavioral and Social Science Experiments

- Many behavioral experiments rely on hiding the true purpose of the study.
- Subjects would change their behavior if told in advance what investigators were looking for.
- The “ethical principles” of the American Psychological Association require consent unless a study merely observes behavior in a public space.
- **HW: Read Section 2.3, go over each of the Examples (2.28-2.35), the Summary and try #2.70, 2.73, 2.76, 2.77, 2.80.**