SECTION 1.1: SAMPLING

OBJECTIVES

1. Construct a simple random sample
2. Determine when samples of convenience are acceptable
3. Describe stratified sampling, cluster sampling, systematic sampling, and voluntary response sampling
4. Distinguish between statistics and parameters

TERMINOLOGY

Statistics is the study of procedures for collecting, describing, and drawing conclusions from information.

A population is the entire collection of individuals about which information is sought.

A sample is a subset of a population, containing the individuals that are actually observed.

OBJECTIVE 1
CONSTRUCT A SIMPLE RANDOM SAMPLE

A simple random sample of size *n* is a sample chosen by a method in which each collection of *n* population items is equally likely to make up the sample. It is analogous to a lottery. Suppose that 10,000 lottery tickets are sold and 5 are drawn as the winning tickets. Each collection of 5 tickets than can be formed is equally likely to make up the group of 5 that is drawn.

EXAMPLE: A physical education professor wants to study the physical fitness levels of 20,000 students enrolled at her university. She obtains a list of all 20,000 students, numbered from 1 to 20,000 and uses a computer random number generator to generate 100 random integers between 1 and 20,000, then invites the 100 students corresponding to those numbers to participate in the study. Is this a simple random sample?

SOLUTION: Yes, this is a simple random sample since any groups of 100 students would have been equally likely to have been chosen.

EXAMPLE: The professor in the last example now wants to draw a sample of 50 students to fill out a questionnaire about which sports they play. The professor's 10:00 am class has 50 students. She uses the first 20 minutes of class to have the students fill out the questionnaire. Is this a simple random sample?

SOLUTION: No, this does not meet the critierion

OBJECTIVE 2
DETERMINE WHEN SAMPLES OF CONVENIENCE ARE ACCEPTABLE

SAMPLES OF CONVENIENCE

In some cases, it is difficult or impossible to draw a sample in a truly random way. In these cases, the best one can do is to sample items by some convenient method. A sample of convenience is a sample that is not drawn by a well-defined random method.

EXAMPLE: A construction engineer has just received a shipment of1000 concrete blocks. The blocks have been delivered in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks. Explain why it might be difficult to draw a simple random sample of blocks.

SOLUTION: To draw a simple random sample would require removing blocks from center to bottom of the pile.

One way to draw a sample of convenience could be to simply take 10 blocks off the top of the pile.

PROBLEMS WITH SAMPLE OF CONVENIENCE

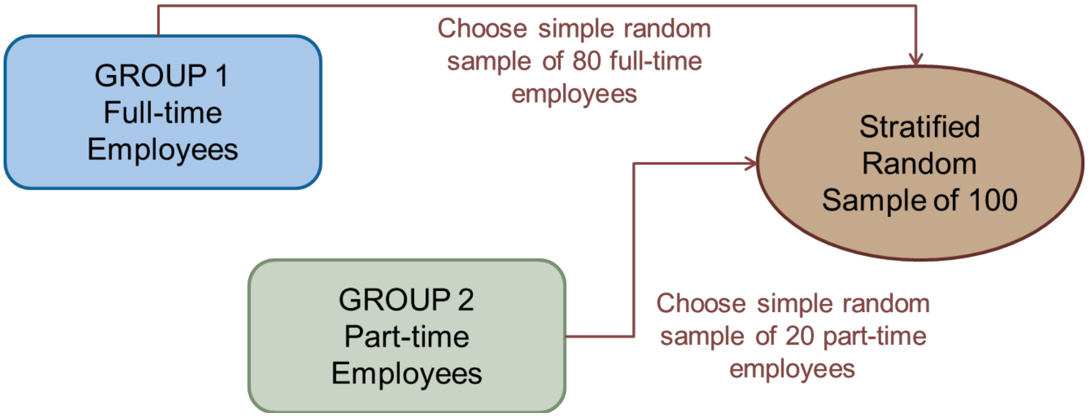
The problem with samples of convenience is that they may differ systematically in some way from the population. If it is reasonable to believe that no important systematic difference exists, then it is acceptable to treat the sample of convenience as if it were a simple random sample.

OBJECTIVE 3
DESCRIBE STRATIFIED SAMPLING, CLUSTER SAMPLING, SYSTEMATIC SAMPLING, AND VOLUNTARY RESPONSE SAMPLING

STRATIFIED RANDOM SAMPLING

In stratified random sampling, the population is divided up into groups, called strata, then a simple random sample is drawn from each stratum. Stratified sampling is useful when the strata differ from one another, but the individuals within a stratum tend to be alike.

EXAMPLE: A company has 800 full-time and 200 part-time employees. To draw a sample of 100 employees, a simple random sample of 80 full-time employees is selected and a simple random sample of 20 part-time employees is selected.

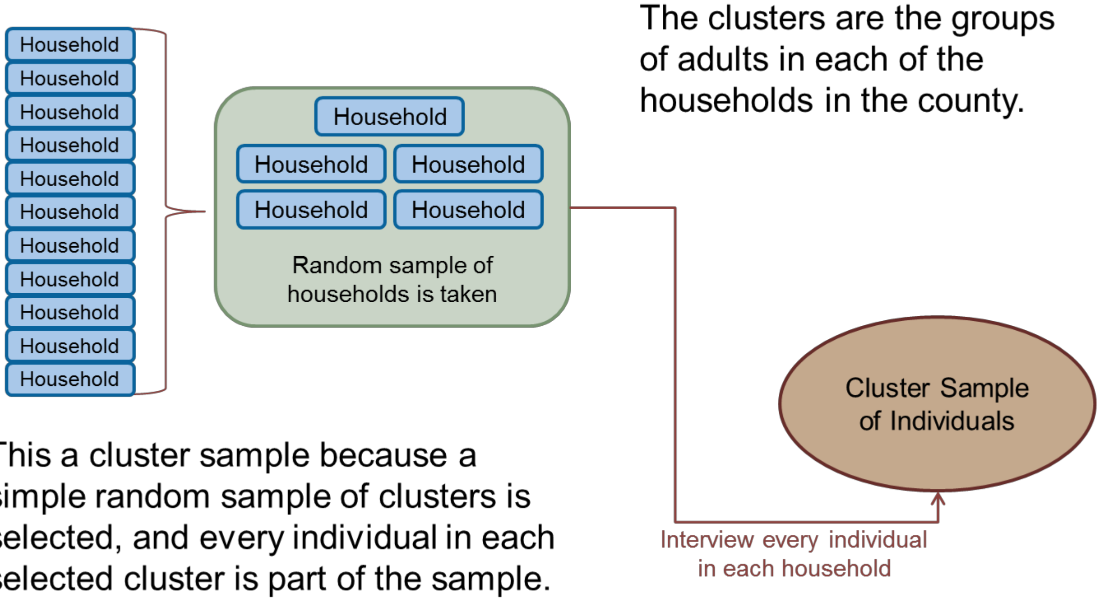

GROUP 1
Full-time Employees
Choose simple random sample of 80 full-time employees
Stratified Random Sample of 100
GROUP 2
Part-time Employees
Choose simple random sample of 20 part-time employees

CLUSTER SAMPLING

In cluster sampling, items are drawn from the population in group, or clusters. Cluster sampling is useful when the population is too large and spread out for simple random sampling to be feasible.

EXAMPLE: To estimate the unemployment rate, a government agency draws a simple random sample of households in a county. Someone visits each household and asks how many adults live in the household, and how many of them are unemployed. What are the clusters? Why is this cluster Sample?

The clusters are the groups of adults in each of the households in the county.



This a cluster sample because a simple random sample of clusters is selected, and every individual in each selected cluster is part of the sample.

Note: After groups are selected randomly, we choose all members from the selected cluster.

SYSTEMATIC SAMPLING

In systematic sampling, items are ordered and every k^{th} item is chosen to be included in the sample. Systematic sampling is sometimes used to sample products as they come off an assembly line, in order to check that they meet quality standards.

EXAMPLE: Automobiles are coming off an assembly line. It is decided to draw a systematic sample for a detailed check of the steering system. The starting point will be the third car, then every fifth car after that will be sampled. Which cars will be sampled?

SOLUTION: The sample will consist of cars numbered 3, 8, 13, 18, 23, 28, and so on

VOLUNTARY RESPONSE SAMPLING

Voluntary response samples are often used by the media to try to engage the audience. For example, a radio announcer will invite people to call the station to say what they think. Voluntary response samples are never reliable for the following reasons:

- People who volunteer an opinion tend to have stronger opinions than is typical of the population.
- People with negative opinions are often more likely to volunteer their response.

OBJECTIVE 4
DISTINGUISH BETWEEN STATISTICS AND PARAMETERS

A statistic is a number that describes a sample. A parameter is a number that describes a population.

EXAMPLE: Which of the following is a statistic and which is a parameter?

- 57% of the teachers at Central High School are female
- In a sample of 100 surgery patients who were given a new pain reliever, 78% of them reported significant pain relief

SOLUTION:
The quantity "57%" is a parameter.
The quantity "78%" is a sample.

YOU SHOULD KNOW …

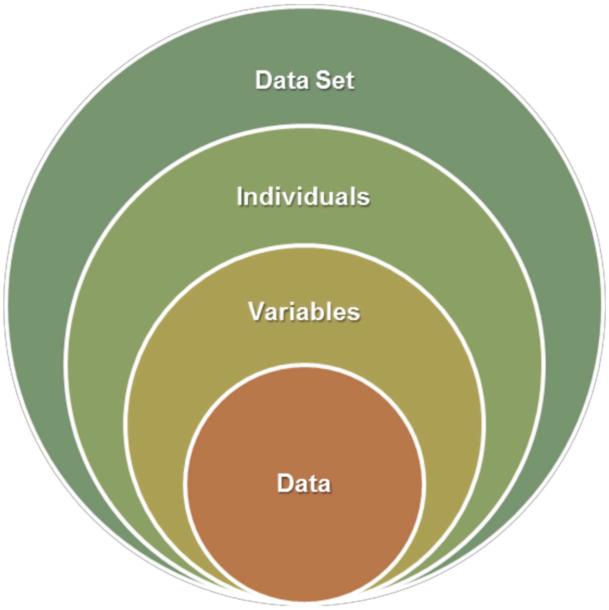
- What is Statistics
- The difference between a population and a sample
- What is a simple random sample
- When samples of convenience are acceptable
- The differences among:
  - Stratified sampling
  - Cluster sampling
  - Systematic sampling
  - Voluntary response sampling
- The difference between a statistic and a parameter

OBJECTIVES

1. Understand the structure of a typical data set
2. Distinguish between qualitative and quantitative variables
3. Distinguish between ordinal and nominal variables
4. Distinguish between discrete and continuous variables

OBJECTIVE 1
UNDERSTAND THE STRUCTURE OF A TYPICAL DATA SET

The values of the variables that we obtain are the ___data___. The characteristics of the individuals about which we collect information are called ___variable___. Information is collected on ___individuals___. The information collected is called a ___data set___.



OBJECTIVE 2
DISTINGUISH BETWEEN QUALITATIVE AND QUANTITATIVE VARIABLES

QUALITATIVE AND QUANTITATIVE VARIABLES

Variables can be divided into two types:

Qualitative: ___classify individuals into categories___

Quantitative: ___tell how much or how many of something___ there is

EXAMPLE: Which of the following variables are qualitative and which are quantitative?

a) A person's age

SOLUTION: The variable is quantitative

b) A person's gender

SOLUTION: The variable is qualitative

c) The mileage of a car

SOLUTION: The variable is quantitative

d) The color of a car

SOLUTION: The variable is qualitative

OBJECTIVE 3
DISTINGUISH BETWEEN ORDINAL AND NOMINAL VARIABLES

ORDINAL AND NOMINAL

Qualitative variables can be further divided into ordinal and nominal variables:

Ordinal variables: have natural ordering

Nominal variables: do not have natural ordering

EXAMPLE: Which of the following variables are ordinal and which are nominal?

a) State of residence

SOLUTION: This variable is nominal

b) Gender

SOLUTION: This variable is nominal

c) Letter grade in a class (A, B, C, D, or F)

SOLUTION: This variable is ordinal

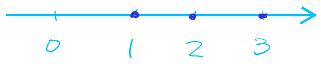
d) Size of a soft drink ordered at a fast-food restaurant

SOLUTION: This variable is ordinal

OBJECTIVE 4
DISTINGUISH BETWEEN DISCRETE AND CONTINUOUS VARIABLES

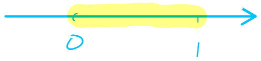
DISCRETE AND CONTINUOUS

Quantitative variables can be further divided into discrete and continuous variables:



Discrete variables: possible values can be listed 1, 2, 3, ...

Continuous variables: can take any value in some interval



EXAMPLE: Which of the following variables are discrete and which are continuous?

a) Age of a person at his or her last birthday

SOLUTION: This variable is discrete

b) Height of a person

SOLUTION: This variable is continuous

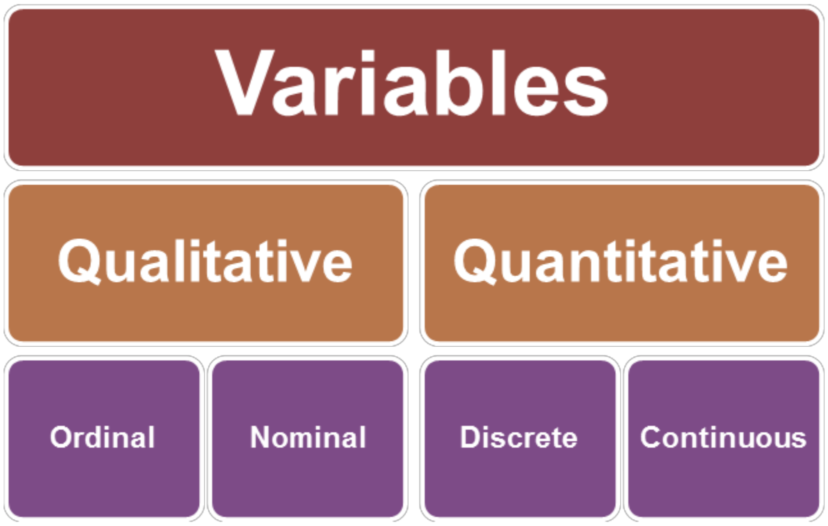
c) Number of siblings a person has

SOLUTION: This variable is discrete

d) Distance a person commutes to work

SOLUTION: This variable is continuous

SUMMARY



YOU SHOULD KNOW …

- The structure of a data set
- How to distinguish between
  - Qualitative and quantitative variables
  - Ordinal and nominal variables
  - Discrete and Continuous variables

OBJECTIVES

1. Distinguish between a randomized experiment and an observational study
2. Understand the advantages of randomized experiments
3. Understand how confounding can affect the results of an observational study
4. Describe various types of observational studies

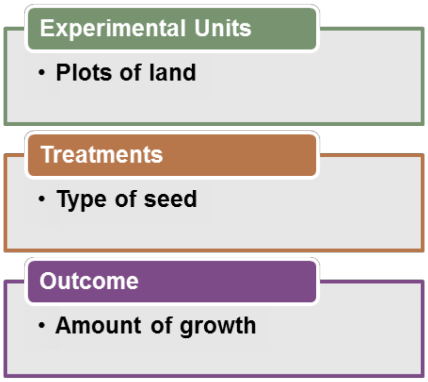
OBJECTIVE 1
DISTINGUISH BETWEEN A RANDOMIZED EXPERIMENT AND AN OBSERVATIONAL STUDY

TERMINOLOGY

Experimental units are individuals that are studied. These can be people, animals, plants, or things. When the experimental units are people, they are sometimes called subjects. The outcome, or response, is what is measured on each experimental unit. treatments are the procedures applied to each experimental unit.

EXAMPLE: Suppose that scientists want to determine which of three types of seed will result in the largest wheat yield. The study is conducted as follows:

- Prepare three identically sized plots of land, with similar soil types.
- Plant each type of seed on a different plot, choosing the plots at random.
- Water and fertilize the plots in the same way.
- Harvest the wheat, and measure the amount grown on each plot.
- If one type of seed produces substantially more (or less) wheat than the others, then that one is clearly better (or worse) than the others.



Experimental Units
- Plots of land

Treatments
- Type of seed

Outcome
- Amount of growth

RANDOMIZED EXPERIMENT

A randomized experiment is a study in which the investigator assigns treatments to the experimental units at random.

EXAMPLE: To assess the effectiveness of a new method for teaching arithmetic to elementary school children, a simple random sample of 30 first graders were taught with the new method, and another simple random sample of 30 first graders were taught with the currently used method. At the end of eight weeks, the children were given a test to assess their knowledge. What are the treatments and why is this randomized experiment?

SOLUTION: The treatments in this experiment are the two methods of teaching
This is a randomized experiment because children were assigned to treatments groups randomly

OBSERVATIONAL STUDY

An observational study is one in which the assignment to treatment groups is not made by the investigator.

EXAMPLE: A study is performed to determine how smoking affects people's health. A group of smokers and a group of nonsmokers are observed for several years. Scientists observe differences in health outcomes between the groups of smokers and nonsmokers. Why is this observational study?

SOLUTION: THIS is observational study because the assignment of treatment (smoking or non smoking) is not made by the investigator

OBJECTIVE 2
UNDERSTAND THE ADVANTAGES OF RANDOMIZED EXPERIMENTS

WHY RANDOMIZE?

In a perfect study, treatment groups would not differ from each other in any important way except that they receive different treatment. In practice, it is impossible to construct treatment groups that are exactly alike, but randomization does the next best thing.

- In a randomized experiment, small differences among treatment groups are likely to be due only to chance.
- If there are large differences in outcomes among the treatment groups, we can conclude that the differences are due to the treatments.

EXAMPLE: In July 2008, scientists reported the results of a study to determine whether a new drug called Raltegravir is effective in reducing levels of virus in patients with HIV. These patients were divided into two groups where one group was given Raltegravir and the other group was given a placebo. In the Raltegravir group, 62% of the subjects had reduced levels of virus, but only 35% of the placebo group did. Because this study was a randomized experiment, it is reasonable to conclude that the difference was actually due to Raltegravir.

DOUBLE-BLIND EXPERIMENTS

An experiment is **double-blind** if neither the investigators nor the subjects know who has been assigned to which treatment.

When investigators or subjects know which treatment is being given, they may tend to report the results differently. Therefore, experiments should be double-blinded whenever possible.

OBJECTIVE 3
UNDERSTAND HOW CONFOUNDING CAN AFFECT THE RESULTS OF AN OBSERVATIONAL STUDY

OBSERVATIONAL STUDIES ARE LESS RELIABLE

Imagine an observational study that is intended to determine whether smoking increases the risk of heart attack. A group of smokers and nonsmokers are observed for several years, and during that time a higher percentage of the smoking group experiences a heart attack.

One problem with this type of study is that the smoking group will differ from the nonsmoking group in many ways other than smoking. For example, smoking is more prevalent among men.

So, the smoking group will contain a higher percentage of men than the nonsmoking group. Men generally are at higher risk of heart attack than women. Therefore, the higher rate of heart attacks in the smoking group may be due to the fact that there are more men in the smoking group, and not to the smoking itself.

CONFOUNDING

The preceding example illustrates a major problem with observational studies. It is difficult to tell whether a difference in the outcome is due to the treatment or to some other difference between the treatment and control groups. This is known as confounding.

EXAMPLE: In an observational study of the effects of blood pressure on health, a large group of people of all ages were given regular blood pressure checkups for a period of one year. It was found that people with high blood pressure were more likely to develop cancer than people with lower blood pressure. Explain how this result might be due to confounding.

SOLUTION: In this example, age is a likely confounder. Older people tend to have higher blood pressure, and older people are more likely to get cancer than younger people.

OBJECTIVE 4
DESCRIBE VARIOUS TYPES OF OBSERVATIONAL STUDIES

TYPES OF OBSERVATIONAL STUDIES

There are two main types of observational studies: cohort studies and cax - cohort studies. Cohort studies can be further divided into prospective, cross - sectional, and retrospective studies.

COHORT STUDIES: PROSPECTIVE

In a cohort study, a group of subjects is studied to determine whether various factors of interest are associated with an outcome.

A propective study is one where the subjects are followed over time.

One of the most famous prospective cohort studies is the Framingham Heart Study. This study began in 1948 with 5209 men and women from the town of Framingham, Massachusetts. Every two years, these subjects are given physical exams and lifestyle interviews, which are studied to discover factors that increase the risk of heart disease.

COHORT STUDIES: CROSS-SECTIONAL

A cross - sectional is one where measurements are taken at one point in time.

An example of a cross-sectional cohort study is a study published in the Journal of the American Medical Association by I. Lang and colleagues. They studied the health effects of Bisphenol A, a chemical found in the linings of food and beverage containers. They measured the levels of Bisphenol A in urine samples from 1455 adults and found that people with higher levels of Bisphenol A were more likely to have heart disease and diabetes.

COHORT STUDIES: RETROSPECTIVE

In a retrospective, subjects are sampled after the outcome has occurred.

For example, in a study published in The New England Journal of Medicine, T. Adams and colleagues sampled 9949 people who had undergone gastric bypass surgery between 5 and 15 years previously, along with 9668 obese patients who had not had bypass surgery. They looked back in time to see which patients were still alive. They found that the survival rates for the surgery patients were greater than for those who had not undergone surgery.

CASE-CONTROL STUDIES

In a case-control study, two samples are drawn. One sample consists of people who have the disease of interest (the cases), and the other consists of people who do not have the disease (the controls).

The investigators look back in time to determine whether a factor of interest differs between the two groups.

S.S. Nielsen and colleagues conducted a case-control study to determine whether exposure to pesticides is related to brain cancer in children.

They sampled 201 children who had been diagnosed with brain cancer, and 285 children who did not have brain cancer. They interviewed the parents of the children to estimate the extent to which the children had been exposed to pesticides. They did not find a clear relationship between pesticide exposure and brain cancer.

YOU SHOULD KNOW …

- The difference between a randomized experiment and an observational study
- The advantages of randomized experiments
- What it means for an experiment to be double-blind
- How confounding can affect the results of an observational study
- The various types of observational studies

OBJECTIVES

1. Define bias
2. Identify sources of bias

OBJECTIVE 1
DEFINE BIAS

BIASED AND UNBIASED STUDIES

Imagine that you were to draw a simple random sample of students to estimate the percentage who are Democrats. By chance, your sample would probably contain a somewhat larger or smaller percentage of Democrats than the entire population of students. However, imagine drawing many simple random samples. Some would have too many Democrats and some would have too few. But on the average, they would balance out. On the average, the percentage of Democrats in a simple random sample will be the same as in the population. A study conducted by a procedure that produces the correct result on the average is said to be unbiased.

Now imagine that you tried to estimate the percentage of Democrats in the population by selecting students who attended a speech made by a Democratic politician. On the average, studies conducted in this way would overestimate the percentage of Democrats in the population. Studies conducted with methods that tend to overestimate or underestimate the true value are said to be biased.

OBJECTIVE 2
IDENTIFY SOURCES OF BIAS

VOLUNTARY RESPONSE BIAS

A voluntary response surveys is one in which people are invited to log onto a website, send a text message, tweet, or call a phone number, in order to express their opinion on an issue. In many cases, the opinions of the people who choose to participate in these surveys do not reflect the population as a whole. In particular, people with strong opinions are more likely to participate. In general, voluntary response surveys are highly biased.

SELF INTEREST BIAS

Many advertisements contain data that claim to show that the product being advertised is superior to its competitors. The advertiser, however, may not report any data that tends to show that the product is inferior. People who have an interest in the outcome of an experiment have an incentive to use biased methods.

SOCIAL ACCEPTABILITY BIAS

People may be reluctant to admit to behavior that may reflect negatively on them which affects many surveys. For example, a pollster may ask “Did you vote in the last presidential election?” The problem with this direct approach is that people are reluctant to answer “No,” because they are concerned that not voting is socially less acceptable than voting.

LEADING QUESTION BIAS

Sometimes questions are worded in a way that suggest a particular response. Consider the difference in the following questions:

- “Do you favor decreasing the heavy tax burden on middle-class families?”
- “What is your opinion on decreasing taxes for middle-class families? Choices: Strongly disagree, Somewhat disagree, Neither agree or disagree, Somewhat agree, Strongly agree.”

NON-RESPONSE BIAS AND SAMPLING BIAS

People cannot be forced to answer questions or to participate in a study. In any study, a certain proportion of people who are asked to participate refuse to do so. These people are called

non - responders.

In many cases, the opinions of non-responders tend to differ from the opinions of those who do respond. As a result, surveys with many non-responders are often biased. Sample bias occurs when some members of the population are more likely to be included in the sample than others.

YOU SHOULD KNOW …

- The difference between a biased and an unbiased study
- The various sources of bias including:
  - Voluntary Response Bias
  - Self-Interest Bias
  - Social Acceptability Bias
  - Leading Question Bias
  - Non-Response Bias
  - Sampling Bias