Section 1.3, Data Collection and Experimental Design

A **variable** is any characteristic that is recorded for subjects in a study. For example, "gender," "major," "age," and "GPA" might be variables for a study about college students.

1 Designing a study

Steps for designing a study:

1. Identify the variable (or variables) of interest and the population of the study.
2. Develop a detailed plan for collecting the data. When using a sample, you need to make sure that the sample is representative of the population.
3. Collect the data.
4. Describe the data using descriptive statistics.
5. Interpret the data and use inferential statistics to make decisions (or assumptions) about the population.
6. Identify any possible errors (some potential problems can be identified earlier in the process).

2 Collecting Data

We will focus on four methods for collecting data: Observational studies, experiments, simulations, and surveys.

- In an **observational study**, a researcher measures and observes the variables of interest without changing existing conditions.
- In an **experiment**, a researcher assigns a **treatment** and observes the **response**. Sometimes, a **control group** (a group receiving no treatment or a **placebo**) may be used to compare the effectiveness of a treatment.
- A **simulation** uses a mathematical, physical, or computer model to replicate the conditions of a process or situation. This is frequently used when the actual situation is too expensive, dangerous, or impractical to replicate in real life.
- A **survey** is used to investigate characteristics of a population. It is frequently used when the subjects are people, and questions are asked of them. When designing a survey, you must be very careful of wording (and sometimes ordering) the questions so that the results are not biased.

Examples
Identify which method for collecting data (observational study, an experiment, a simulation, or a survey) is best in each of the following situations and explain your answer.

1. The effect a severe earthquake would have on the Salt Lake Valley. Simulation, since it is too dangerous to create an actual earthquake (Plus, we cannot control when nature will have an earthquake).
2. Whether or not a certain coupon attached to the outside of a catalog makes recipients more likely to order products from a mail-order company. Experiment, since we are comparing two scenarios and we can control them.

3. Whether or not smoking has an effect on coronary heart disease. Observational study, since we will not be changing a person's behavior (there are ethical and health concerns for deciding whether someone smokes or not)

4. Determining the average household income of homes in Salt Lake City. Survey, since it can be answered with a brief question.

3 Experimental Design

Definitions and Terminology

- A **confounding variable** occurs when an experimenter cannot tell the difference between the effects of different factors on a variable.
- The **placebo effect** occurs when a subject (or "experimental unit") reacts favorably to a placebo when no medicated treatment has been given.
- **Blinding** is a technique used to make the subjects "blind" to which treatment (or placebo) they are being given.
- A **double-blind** experiment is one in which neither the experimenter nor the subjects know which treatment is being given.
- **Randomization** is a process of randomly assigning subjects to treatment groups. There are several different techniques for randomization:
  - A **completely randomized design** assigns subjects to different treatment groups through random assignment.
  - A **randomized block design** is sometimes used to make sure that subjects with certain characteristics are assigned to each treatment. For example, when testing a certain medication, you might first want to split subjects in groups according to either gender or age (or both), then randomly assign each of these groups to the different treatments.
- A **matched pairs design** pairs up subjects according to similarities. One subject in the pair receives one treatment, while the other receives a different treatment.
- **Sample size** is the number of participants in the experiment. The larger the sample, the more representative of the population the results will be, but the costs of the experiment will also be higher.
- **Replication** is the ability to reproduce the experiment (and results) under similar conditions.

Examples

1. For the following experiment, determine the experimental units, treatments, and sample size. Indicate whether this experiment is blind, double-blind, and/or randomized. Also identify any potential problems with the design.

   A study with 233 low-income adult smokers evaluated the effectiveness of usual care (physician advice and follow-up) for smokers wishing to quit to the usual care enhanced by computer-assisted telephone counseling sessions. Each subject was assigned randomly either to the usual care or to the usual care plus counseling, and their smoking status (still smoking or quit smoking) was observed after 3 months. The percentage who had quit smoking was higher for the group receiving counseling. (from *Journal of Family Practice* 2000;50:138-144)
   The experimental units are the 233 adult smokers.
   The treatments are "usual care" and "usual care plus counseling."
   The sample size is 233.
   The experiment is not blind or double-blind (experimental units will know if they receive counseling or not), but it is randomized since the subjects were assigned randomly.

2. How would you design a placebo-controlled double-blind experiment with a randomized block design for the following situation: A veterinarian wants to test a strain of antibiotic on calves to determine their resistance to a common infection, and if their gender plays a role. In a pasture, there are 22 newborn calves (11 males and 11 females). There is enough vaccine for 10 calves, but blood tests to determine their resistance to infection can be done on all calves.

For the randomized block design, we can choose 5 males and 5 females randomly to receive the antibiotic, and the remaining 6 males and 6 females will receive a "placebo" (in this case, the "placebo" could be no treatment since the calves won't be telling anyone if they received a shot or not). In order to make the experiment double-blind, we need to make the calves unaware of what treatment they are receiving (not difficult), and for the person carrying out the blood tests to be unaware of the treatment (to make this possible, another person will be assigning which of the calves receive the antibiotic and which do not).

4 Sampling Techniques

Ideally, we would take a **census**, that is, use every member of a population as a subject since the descriptive statistics would be sufficient. However, this is often too costly and difficult. Instead, we **sample** part of the population. With sampling, we need to make sure that the sample is representative of the population and large enough to be meaningful.

Definitions and Terminology

- A **sampling error** is the difference between the results of the sample and those of the population. Even with the best sampling techniques, this is possible.
- A **biased sample** is one that is not representative of the entire population. We want to avoid bias.
- A **random sample** is one in which every member of the population has an equal chance of being chosen.
- A **simple random sample** (SRS) is a sample in which every possible sample of the same size has the same chance of being collected. Normally, we will start by using a simple random sample.
- A **stratified sample** is used when it is important to have members from multiple segments of the population. First, the population is split into segments (called "strata"), then a predetermined number of subjects is chosen from each of the strata.
- **Cluster sampling** can be used when the population naturally falls into subgroups with similar characteristics. First, determine the clusters, then select all the members of one or more of the clusters.
- **Systematic sampling** first involves assigning a number to each member of the population and ordering them in some way. Sample members are selected by choosing the first member randomly, then selecting subsequent members at regular intervals after the starting number (for example, every 7th person). This method is fairly simple to use, but should be avoided if there are regularly occurring patterns in the data.
- A **convenience sample** consists only of available members of the population, but this often leads to biased studies.
- A **volunteer sample** is a kind of convenience sample in which only volunteers participate.

To choose subjects for a SRS, first determine the size of the population and number everyone on the list. Then find a table of random numbers (such as Table 1 in Appendix B of your textbook). The size of your population tells you how many digits to read at once (for example, if there are 132 members, you will need 3 digits, but with 32, you would only need 2. Note that with 100 members, you will only need 2 digits if you number the members of your population from 00–99). To select the first subject, read the first "few" digits of the table ("few"= the number of digits that you determined you needed), then find that number in your numbered list of the population. If the number you selected is larger than any number on your list, ignore that number and move onto the next "few" digits and try again. Continue until you have as many subjects as you need. Note that you need to make sure that subjects are not repeated.

Examples

1. Determine which kind of sampling was used in each of the following scenarios:

(a) To determine the quality of on-campus housing, 20 residents from each dorm were chosen to complete a survey. Stratified sampling

(b) To evaluate employee compensation, choose a random sample of 10 zip codes in the state, then survey all businesses within each chosen zip code about their benefits package. Cluster sampling

(c) Those who participate to a survey linked to from cnn.com. Volunteer sampling

(d) To determine the quality of education at the University of Utah, a UNID number is chosen at random, then every 1000th student is evaluated until 30 students are selected. Systematic sampling

(e) Interested in only one neighborhood, you walk door-to-door to ask residents questions. Since your time is limited, you do not have a chance to revisit houses where no one answered the door. Convenience sampling

(f) Interested in only one neighborhood, you walk door-to-door to ask residents questions. Everyone was home and willing to participate, so you have survey results from every household in the neighborhood. Census

(g) Chosen at random, 300 people who received care at the University Hospital participated in a survey. Simple random sampling

2. Using Table 1, determine which 5 students from this class (numbered 1-26) should be selected to participate in a survey.
We only need to read 2 digits at a time, so we start reading Table 1 as:
92 63 07 82 40 19 26 79 54 57 53 49 72 38 94 37 70 87 98 62 76 47 16 64 18
Going from left to right on the list, and ruling out numbers that are larger than 26, we are left with #7, 19, 26, 16, and 18.