Sampling and Data Collection
Chapter 2

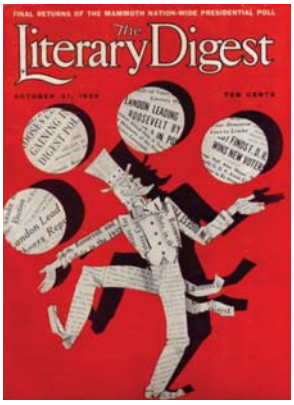
Learning Outcomes

By the end of this lesson, you should be able to define the following vocabulary terms:

- Observational study
- Designed experiment
- Categorical variable
- Quantitative random variable
  - Discrete quantitative random variables
  - Continuous quantitative random variables
- Population
- Sample
- Parameter
- Statistic
- Sampling schemes
  - Simple random sample (SRS)
  - Systematic sample
  - Cluster sample
  - Stratified sample
  - Convenience sample
- Bias
- Undercoverage
- Nonresponse
- Voluntary response samples
- Poorly worded questions
- Subjects (experimental units)
- Factors
- Treatment groups
- Three principles of experimental design
  - Control
  - Randomization
  - Replication

Observational Studies Versus Designed Experiments

Observational Study: Literary Digest Poll



Cover from the famous *Literary Digest* magazine from 31 Oct. 1936.

The Democratic Party and Republican Party are the two major political groups in the United States of America. In the 1936 presidential election, President Franklin Roosevelt, a democrat, was challenged by a republican named Alf Landon. In 1920, 1924, 1928, and 1932, the popular magazine *Literary Digest* used polls to correctly predict the outcome of four presidential elections. In 1936, *Literary Digest* carried out another poll, what was then the most extensive public opinion poll in history, mailing out questionnaires to over 10 million people in the United States.

More than 2.4 million people responded, with 43% of the respondents indicating that they would vote for Franklin Roosevelt in the upcoming election [3]. The *Literary Digest* published the results under the headline "Landon, 1,293,669; Roosevelt, 972,897: Final returns in the Digest's poll of ten million voters."

In the accompanying article, the chairman of the Democratic National Convention, James Farley, says, "The *Literary Digest* poll is an achievement of no little magnitude. It is a poll fairly and correctly conducted [it is the] most extensive straw ballot in the field-the most experienced in view of its twenty-five years of perfecting-the most unbiased in view of its prestige-a poll that has always previously been correct" [3].

Incredibly, on election day Franklin Roosevelt easily won, carrying 63% of the popular vote! Although *Literary Digest* had predicted a huge victory for Alf Landon, the landslide was for Roosevelt. What went wrong? How could such a large sample give such inaccurate results?

There are many reasons for the poll's failure [4, 5], and some background vocabulary will help you understand them. Let's call the total number of people who voted in the election the **population**, and the true proportion of people who voted for Roosevelt the **population parameter**, which is simply a number describing a population. In most cases, we never know the true parameter, so we have to estimate it using a **statistic**.

Memory Aid:
A sample statistic is any number that describes a sample.
A population parameter is any number that describes a population.
Usually, we do not know what the values of the parameters are, so we collect a sample and use the statistics to estimate the parameters.
For example, *Literary Digest* did not know the true proportion of people who would vote for Roosevelt. To estimate this parameter, they collected a sample and estimated the true proportion with their sample proportion.

Because it would be unreasonable to survey the entire population, the *Literary Digest* researchers surveyed a **sample**, or selection of the population, which in this case included 10 million voters. From this sample the researchers gathered their sample statistics, which are any numbers that describe a sample.

Typically, the sample proportion will be close to the population proportion, but in this study, the sample proportion (43%) was a low estimate of the true population proportion (63%). With such a large sample size, we would have expected the sample proportion to be closer to the population proportion. Clearly, the estimator was biased. **Bias** occurs when there is a difference in the long run average of a statistic compared to the true parameter it is estimating. Bias introduces a systematic error into a study. In this study, the sample proportion was systematically lower than it should have been, due to the sampling procedures.

Another reason for the poll's inaccurate prediction lies in the method of data collection, which in this case was a **convenience sample**, or a sample drawn using individuals who are easily accessed. The magazine's researchers used lists that were conveniently available, such as subscription lists, telephone records, and automobile registrations. Samples like these are easy to collect, but may not always represent the entire population. For example, the *Digest*'s sample tended to include

a disproportionate number of wealthy Americans, those who could afford magazine subscriptions, telephones, and motor vehicles during the Great Depression-and who traditionally favor Republican candidates like Alf Landon.

As a result, the poor and middle class voters were not sufficiently represented in the *Digest*'s study, which was consequently characterized by **undercoverage**, or a failure to include certain elements of the population appropriately. If the responses of the unrepresented individuals differ from the responses of the sample, undercoverage will result in a biased estimate of the true parameter.

Moreover, out of over 10 million questionnaires sent out, only 2.4 million were returned, resulting in less than a 24% response rate. We do not know anything about the opinions of the over 7.6 million people. Members of a sample, invited to participate but failing to do so, create VocabNonresponse.

Nonresponse can lead to a **biased estimator**. For example, people who were unhappy with President Roosevelt were more likely to express their feelings and send in a survey than those who are either satisfied with him or just neutral. So, the probability that those who favor Roosevelt will respond to the survey is not the same as the probability that those who favor Landon will respond to the survey. Clearly, James Farley's claim that the poll conducted by the *Literary Digest* was "the most unbiased" of any other poll is based on a poor understanding of the importance of response.

Although the following two ideas are not necessarily the cause of the failed survey conducted by the *Literary Digest*, they are useful ideas that will be referenced throughout this textbook. For example, in a **voluntary response sample**, the subjects determine whether or not they will participate in the study. The T.V. show *American Idol* chooses their winners based on the number of times viewers call or text their votes for each of the contestants. These types of polls tend to be biased, since only those with strong feelings will make the effort to be included in the study. Sometimes there is even a fee for participating, leading to even more bias.

Many surveys and questionnaires are weakened by **Poorly worded questions**, which leave respondents confused or misdirected. Poorly worded questions also lead to bias in the survey results and incorrect conclusions. The wording of questions has been shown to influence the responses in surveys [6]. "A poorly worded question," explain statisticians Linda Del Greco and Wikke Walop, "may yield a wide variety of responses that do not relate to the question, and this leaves the researcher with unusable information" [7].

Researchers must take precautions against these kinds of errors when undertaking observational studies. Had the pollsters at the *Literary Digest* been more careful in interpreting the public's response to their survey, they might not have published such an inaccurate prediction. In fact, this significant error was the primary factor contributing to the magazine's demise over the next two years.

Designed Experiment: Gratitude Study

The preceding example of data collection-the study undertaken by the *Literary Digest* magazine-is an example of an observational study, or a study in which the researchers do not assign the participants into treatment or control groups, but rather simply observe the responses of the individuals.

Whereas an observational study gathers statistics by simply counting and organizing existing data without directly manipulating participants, a **designed experiment** is an experiment which researchers invent and carry out, often dividing participants into groups and manipulating their environments. In a designed experiment, the researchers randomly assign subjects to treatment and control groups and observe the effect of the treatment. The following study is an example of a designed experiment.

In 2003 Professors Robert Emmons and Michael McCullough scientifically "examine[d] the influence of grateful thinking on psychological well-being in daily life" [8]. To do so, the researchers recruited 192 undergraduate participants and divided them into three experimental groups.

Each student was asked to complete a weekly journal in which they recorded five specific things and then answered some questions about their feelings about life in general. The students in Group 1 were asked to record five things that they were grateful for that week, the students in Group 2 were instructed to record five hassles they had experienced during the week, and the students in Group 3 were told to record any five events-good or bad-that had occurred in the previous week. The researchers wanted to know if there would be a difference in the perceptions of life as a whole among the three groups. People in the "gratitude" group tended to be more satisfied with life in general than people in the other group.

A **subject** (or an **experimental unit** or simply "**unit**") is one individual in the population that was selected for participation in the study. In this experiment, the subjects are the 192 students who participated in the experiment. In a research study, the **population** is defined as the collection of all the experimental units that could have been selected.

The population is the group to which the results can be generalized. In this study, the population is all college students. (Remember that in the previous example, the population included all registered American voters.) However, the researchers were willing to generalize their definition to include the general public. The **sample** is the subset of the population selected for inclusion in the study.

A **treatment** is any manipulation that is applied to the subjects or experimental units. The students were randomly assigned to one of three **treatment groups**, the treatments being the things the students were told to record (gratitude, hassles, or events). It is important that the students were randomly assigned to the groups, so we can assume that each group is similar in every respect except for the treatment they received. Any differences that occur in the responses of the groups can be attributed to the treatment. That is, if we see there is a difference in the feelings about life between the three groups, then we would conclude that this was due to the things they were

asked to write.

Although controversial, it is well documented that the condition of individuals tends to improve when they receive attention or treatment, even if the treatment is an inert **placebo** or a **control** [9]. This is known as the **Placebo Effect**. An essential part of a good experimental design is the use of a control. The control is a treatment applied to some of the subjects so that they think they are receiving the same treatment as the rest of the participants. In the gratitude study, Group 3 is the control. The people in this group wrote five events that occurred in their life. Imagine if the events group did not write anything. A critic could claim that differences in the attitudes of the students toward life in general are attributable to the *act* of writing five things down, not *what* they wrote.

A **factor** is a variable(?) that is studied in an experiment. In the gratitude study, there was one main factor that was addressed: the expression of feelings (whether people expressed gratitude, complained about hassles, or merely recorded events.) Other factors that could have been included are: gender, age, family background, religious affiliation, etc. The possible number of factors

A **random variable** is anything that is measured where the outcome is not predetermined. In the *Literary Digest* poll, the response of each person was a random variable. The respondent could either state that they favor the Republican or the Democratic candidate. When a random variable can result in one of several categories, we call it a **categorical variable**. If the survey asked the age of the respondent, the age would have been quantitative. We say that a variable is a **quantitative variable** if the values that it can assume are given on a numerical scale.

A quantitative random variable can either be discrete or continuous. A **discrete random variable** is a random variable where the possible outcomes can be counted. For example, the number of stars visible on any given night is a discrete random variable. The number of sands upon the sea shore is a discrete random variable. There could be an infinite number of possibilities, but as long as there is a way to number them, the random variable is discrete. A **continuous random variable** is an outcome that is measured. The height of a person is an example of a continuous random variable. Note that when we think of height here, we are not rounding it to the nearest inch or centimeter. If a person is six feet tall, they have had to grow continuously through every possible height from zero to their present height.

Sampling Schemes

There are several common ways to get a random sample. To avoid introducing bias into the results and to allow the observations from a sample to represent a population, it is important that every sample be chosen randomly when possible.

Simple Random Sample (SRS)

All the statistical procedures in a typical introductory statistics course assume that a **simple random sample** or **SRS** was conducted. A simple random sample is obtained when every unit in the population has an equal chance of being chosen and every group of size n has an equal chance of being chosen. The classical example of a simple random sample is to put names in a "hat" and to draw out n names at

random. One advantage of a simple random sample is that it is easy to understand. This sampling method is the basis of the calculations for the inferential procedures in the second half of this course. The major limitation of this method is that it requires the researcher to have a list of all the units in the population in advance of selecting the sample. In practice, this is hard to do. Consider the difficulty of taking a simple random sample of the wages of people in your state or country. In order to do this, you would need to obtain contact information for every person in the boundaries. Then, you would need to draw a sample from this list. Due to the difficulty in obtaining the list of all individuals in the population, simple random samples are seldom done in practice.

Convenience Sample

A **convenience sample** is one of the most common sampling schemes used, but it is one of the worst. In a convenience sample, a researcher collects data from subjects that are readily available. An example is a person who asks their friends' opinion on a political matter. In many cases, the responses of volunteers who are easily contacted may not represent the entire population. For example, there is a good chance that many of your friends' political views are similar to your own. In almost all cases, a convenience sample is not random at all.

"Teach ye diligently and my grace shall attend you, that you may be instructed more perfectly in theory, in principle, in doctrine, in the law of the gospel, in all things that pertain unto the kingdom of God, that are expedient for you to understand;

Of things both in heaven and in the earth, and under the earth; things which have been, things which are, things which must shortly come to pass; things which are at home, things which are abroad; the wars and the perplexities of the nations, and the judgments which are on the land; and a knowledge also of countries and of kingdoms-

That ye may be prepared in all things..."

(D&C 88:78-80.)

Systematic Random Sample

A **systematic random sample** is a convenient way to choose a random sample when the experimental units can be ordered in a unique way and the order is not related to the responses. A systematic random sample requires that all the items be ordered. There must be a way to identify a "first" and "last" unit, and there should be a logical sequence for every unit in between these two. For example, if you wanted to take a systematic random sample of breakfast cereals at your local grocery store, you might mentally break up the aisle into 8-foot lengths of shelving. Starting on the top row, work across and move to the next row when you get to the end of the first 8-foot length of shelves. When you have finished the last row, move up to the top left corner of the next set of shelves, and continue the process. The exact order of the units is not important, the point is to be able to place the units in the population in some order. If you are working with people, you can order the people alphabetically.

Next, you need to know the approximate number of units in the population. This does not need to be exact, but it should be close. For example, if you are working with breakfast cereals, you would count how many types of cereals are for sale. For our purposes, let's call the total number of units you estimated to be T (for total).

Suppose we want a sample of n units. Take the total number of items you estimated, T, and divide this number by n. Round the result to the nearest whole number. This is the "sequence" number, k, for a systematic random sample.

$$k = ROUND\left(\frac{T}{n}\right)$$

In a systematic sample, data are collected on every k^{th} unit. If we always start

counting with the first unit, then there are many units that can never be chosen. To avoid this problem and to make it possible for every unit to be selected in the sample, we start counting from a random location. To choose the starting point, randomly select a number between 1 and k. Go to the item with that number and record data on that individual. Then, skip ahead k units and record data on every k^{th} item from then on.

As an example, suppose you estimate that there are 690 items in all in your population. If you want a sample with $n = 50$ observations, you divide 690 by 50: $690/50 = 13.8$. Rounding 13.8 to the nearest whole number, we get a sequence number of k=14. Next, randomly choose a starting point, a number between 1 and $k = 14$. You could do this by writing the numbers 1, 2, 3, , 14 on pieces of paper and putting them in a hat. Draw out one of the papers. This is the starting point. For our example, suppose that the randomly selected starting point was 5. Finally, starting with the first item, count until you find the fifth item. This is the first unit in our sample. Now progress along the list of units in the population until you come to the $5 + 14 = 19^{th}$ unit. This is the second unit in the sample. The next unit will be the $19 + 14 = 33^{th}$ unit. This continues until you reach the end of the population. You do not begin again at the start once you pass through the population one time. You should have about 50 or so observations when you finish.

Stratified Sample

If you were to conduct a study of residents' opinion on immigration reform, you might expect that the attitudes would be related to the economic status of the respondent. To be sure that opinions from individuals from various economic backgrounds are represented, you might conduct a stratified sample. In a **stratified sample**, the population is subdivided into subgroups called "strata" and a simple random sample is drawn from each subgroup (or stratum). In the study on attitudes toward immigration reform, the strata might be defined by grouping people based on their total household income. A simple random sample would be drawn from each group. If the University wants to poll students and be sure that a good cross-section of the population is obtained, they might, for example, survey a simple random sample of 100 freshmen, 100 sophomores, 100 juniors and 100 seniors. This is an example of a stratified sample design.

Note: *Strata* is the plural form of the word *stratum*.

A stratified sample is particularly helpful when there are clearly defined strata, and the observations within a stratum tend to be similar but the observations tend to differ significantly from stratum to stratum.

Cluster Sample

Conducting research in a public school is difficult. First, a researcher must obtain permission from their sponsoring institution (i.e. BYU-Idaho) to conduct the research. Then, they must obtain permission from the superintendents of the school districts or school corporations where they hope to conduct the research. Assuming that is granted, they must obtain the consent and cooperation of each teacher who will be involved. Lastly, they must receive permission from the parents of the children who will be involved in the research. It is costly to obtain permission to work with a child in a particular classroom. Once you have permission to conduct the research on one

child, it is relatively inexpensive to obtain permission to work with other children in the class. A cluster sample is particularly well suited to this setting. In a **cluster sample**, the population is divided into subgroups, called clusters. In the example, the cluster would be the classroom. A simple random sample of clusters is taken, and then every unit in the randomly selected clusters is surveyed.

Another example that illustrates the use of a cluster sample would be if a researcher wanted to survey residents of a nursing home. It is costly to get permission to conduct the research and to travel to a nursing home to collect data. It would make sense to use a cluster sample to randomly select nursing homes (clusters) to survey and then to conduct the research with every person in the randomly selected facilities.

A cluster sample is particularly effective if the units within a cluster are fairly representative of the population in general. This method is used frequently when the cost to include a cluster in the study is high, but the cost to conduct research within the cluster is relatively low.

Principles of Experimental Design

There are three basic principles of experimental design: control, randomization, and replication.

The first principle is to use a **control**. The only way to know if an observed effect can be attributed to a treatment is if there is a control. Without a control, it is impossible to know if the treatment caused the outcome or if it is due to other sources. In any designed experiment, it is essential that **randomization** is used to assign the subjects to the treatments and where possible, to select the subjects. The scientific method highlights that results should be repeatable. In designing an experiment, a researcher should use **replication** to make sure the results are repeatable. Ideally, the treatments should be repeated independently on several different units to help eliminate the possibility that the observed outcomes are due to extraneous factors.

Build Your Understanding (BYU)

Answer the following questions.

1. List the five processes of statistics.

2. The BYU-Idaho learning model includes many processes involving active learning. Paul H. Kvam wrote an article on the impact of active learning on student learning [10]. His abstract is given below. Explain how each of the five processes of statistics are evidenced or implied in his research.

> An experiment was carried out to investigate the long-term effects of active learning methods on student retention in an introductory engineering statistics class. Two classes of students participated in the study–one class was taught using traditional lecture-based learning, and the other class stressed group projects and cooperative learning-based methods. Retention was measured by examining the students immediately after the course finished, and then again eight months later. The findings suggest that active learning can help to increase retention for students with average or below average scores. Graphical displays

of the data, along with standard statistical analyses, help explain the observed difference in retention between students in the two different learning environments.

3. Read the following abstract from an article written by Aryn C. Karpinski and Adam Duberstein [11]. Explain how each of the five processes of statistics were used by the authors.

> Facebook has gained popularity in the college student population since its inception in 2004. Studies have evidenced the broad implications of technology on academic performance, but researchers had not previously examined differences between Facebook users and non-users in undergraduate and graduate student populations in relation to academic performance. Surveys were administered to undergraduate ($n = 71$) and graduate students ($n = 43$) in the Summer quarter of 2008. Significant differences were found between Facebook users and non-users for GPA and study time, [where users of Facebook have a lower mean GPA and shorter study times,] with these differences persisting in the undergraduate and graduate student samples. University administrators may consider using Facebook as a learning tool to enhance academic performance, or find ways to deter recreational Facebook use and promote better time-management skills.

4. Cheating is a problem on college campuses and can occur in many different ways. Design a study to assess how common cheating is

Learning Checklist

- distinguish between an observational study and an experiment
- identify the difference between a categorical and a quantitative variable
- classify quantitative random variables as either discrete or continuous
- distinguish between a population and a sample
- distinguish between a parameter and a statistic
- design a simple observational study or experiment
- discuss the characteristics of the following sampling schemes:
  - simple random sample (SRS)
  - systematic sample
  - cluster sample
  - stratified sample
  - convenience sample
- identify specific instances in which each type of sampling scheme should be used
- explain the importance of using a random sample
- discuss important considerations in designing a sample survey including:
  - bias
  - undercoverage
  - nonresponse

  - voluntary response samples
  - convenience samples
  - poorly worded questions
- identify the following when presented with the description of an experiment:
  - subjects (or experimental units)
  - factors
  - treatment groups
- explain the three principles of experimental design: control, randomization, and replication
- correctly utilize the basic vocabulary associated with data collection