

1 | SAMPLING AND DATA



Figure 1.1 Photo by [Markus Spiske](#) on [Unsplash](#)

Introduction

Chapter Objectives: By the end of this chapter, the student should be able to

- Understand basic statistical terminology
- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. Think about the Covid-19 pandemic. Think about the political elections. As well, the fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are some basic ideas and terms used in Statistics. We will learn how data are gathered and how to distinguish "good" data from "bad."

1.1 | Basic Definitions

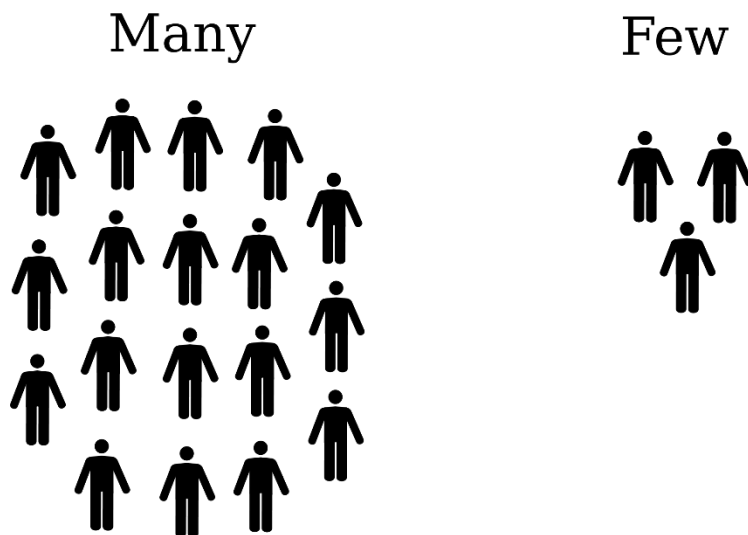
The science of **Statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. There are two basic branches of Statistics: Descriptive and Inferential.

- **Descriptive Statistics** involves organizing and summarizing data. Two general ways to summarize data are with graphs and numerical values, such as an average.
- **Inferential Statistics**. Involved learning formal methods for drawing conclusions from data, which we will learn later in the text. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. We will encounter what may seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer; the understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life and in your chosen career.

Key Terms

In Statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. For logistical reasons, it is usually not possible to gain access to all of the information from the entire population. So when we want to study a population, we usually select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.



<https://openclipart.org/detail/216524/many-and-few>

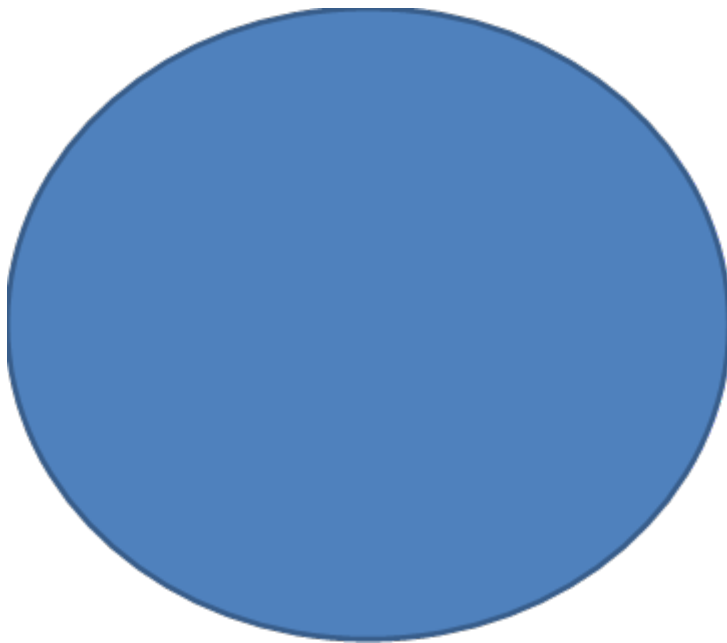
Because it takes a lot of time and resources to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential

elections, samples of between 1,000 and 2,000 prospective voters are used for opinion polls. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that represents a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

Population:

all members being studied



Parameter:

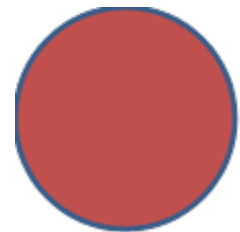
Measurements found using the population. μ is the average of the population. Therefore μ is a parameter.

Statistic:

Measurement found using the sample. $\bar{x}$ is the average of the sample. Therefore $\bar{x}$ is a statistic.

Sample:

subset of
population



One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, denoted by capital letters such as X and Y , is a characteristic of interest for each person or thing in a population. Variables may be numerical or categorical.

- **Numerical variables** take on numerical values with equal units such as weight in pounds and time in hours.
- **Categorical variables** place the person or thing into a category.

If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then Y is a categorical variable, and some possible values of Y would be Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

The actual values of a variable are called **data** (a single value is a **datum**); these values may be numbers or words.

Example 1.1

Determine what the key terms refer to in the following study: We want to know the average amount of money first-year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first-year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution 1.1:

The **population** is all first year students attending ABC College this term.

The **sample** is the 100 first year students surveyed at the college (although this sample may not represent the entire population).

The **parameter** is the average amount of money spent (excluding books) by first year college students at ABC College this term. This average would be represented by μ .

The **statistic** is the average amount of money spent (excluding books) by first year college students in the sample. This average would be represented by $\bar{x}$.

The **variable** is the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** would be the actual dollar amounts spent by the first-year students. Examples of the data would be \$150, \$200, and \$225.

Try It

1.1 Determine what the key terms refer to in the following study. We want to know the average amount of money spent on school uniforms each year by families at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Example 1.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. ___ Population 2. ___ Statistic 3. ___ Parameter
 4. ___ Sample 5. ___ Variable 6. ___ Data
- a) the cumulative GPA of a student who graduated from the college last year
 - b) the average cumulative GPA of students surveyed who graduated from the college last year
 - c) 3.65, 2.80, 1.50, 3.90
 - d) a group of students who graduated from the college last year, randomly selected
 - e) all students who graduated from the college last year
 - f) the average cumulative GPA of all students in the study who graduated from the college last year

Solution 1.2: 1. e; 2. b; 3. f; 4. d; 5. a; 6. c

Example 1.3

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution 1.3

The **population** consists of all cars containing dummies in the front seat.

The **sample** is the set of 75 randomly selected cars.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries measured in the sample.

The **variable** X = whether or not the driver dummies (if they had been real people) have suffered head injuries.

The possible **data** values would be either: yes, had head injury, or no, did not.

Example 1.4

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution to 1.4:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

1.2 | Data and Sampling

Data can come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the two major categories of **qualitative** or **quantitative**.

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, type of car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+.

Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

- **Quantitative discrete data** is data that are the result of counting. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.
- **Quantitative continuous data** is all data that are the result of measuring assuming that we can measure accurately. Measuring angles in radians might result in such numbers as π , $\pi/3$, $5\pi/6$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks would be discrete data and the weights of the backpacks would be continuous data.

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Try It

1.5 The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type data is this?

Try It

1.6 The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. ft, 190 sq. ft, 180 sq. ft, and 210 sq. ft. What type of data is this?

Example 1.7

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and 32 ounces chocolate chip cookies).

Identify data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution 1.7: Answers will vary, but one possibility is:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) and weights of desserts are quantitative continuous data because we measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Example 1.8

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

NOTE: You may collect data as numbers and report it categorically. For example, exam scores for students are recorded throughout the term. At the end of the term, letter grades are reported as A, B, C, D, or F.

Example 1.9

Work collaboratively to determine the data is quantitative or qualitative. For quantitative data, indicate whether they are continuous or discrete.

- a. the number of pairs of shoes you own
- b. the type of car you drive
- c. where you go on vacation
- d. the distance it is from your home to the nearest grocery store
- e. the number of classes you take per school year
- f. the tuition for your classes
- g. the type of calculator you use
- h. movie ratings (PG, R, ...)
- i. political party preferences
- j. weights of sumo wrestlers
- k. amount of money (in dollars) won playing poker
- l. number of correct answers on a quiz

- m. peoples' attitudes toward the government
- n. IQ scores (This may cause some discussion.)

Solution to 1.9:

Quantitative Discrete (a, e, l); Quantitative Continuous (d, f, j, k, n); Qualitative (b, c, g, h, i, m)

Try It Σ

1.9 Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. If quantitative indicate whether it is continuous or discrete.

Example 1.10

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. Figure 1.1 summarizes the data she collected.

What type of data does this graph show?

Classification of Statistics Students

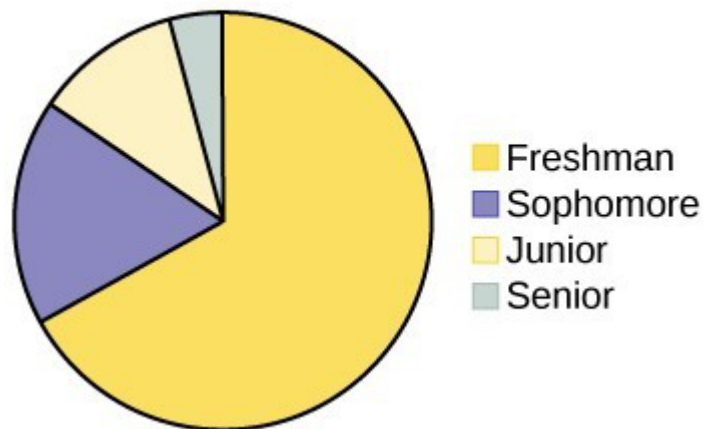


Figure 1.1

Solution 1.10 The chart shows the students in each year, which is **qualitative data**.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we usually use data from a sample of the population.

A sample should have the same characteristics as the population it is representing.

Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common sampling methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons.

The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in **Table 1.2**:

ID	NAME	ID	NAME	ID	NAME
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Statcher
06	Fontecha	17	Pate;	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Table 1.2

Lisa can then use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers. The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14),

the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Macierz, Cuningham, and Cuarismo.



Using the TI-83, 83+, 84, 84+ Calculator

To generate random numbers:

- Press MATH.
 - Arrow over to PRB.
 - Press 5: randInt(: Enter 0, 30.
 - Press ENTER for the first random number.
 - Press ENTER two more times for the other 2 random numbers.
- If there is a repeat press ENTER again.

```
randInt(0,30) 29
randInt(0,30) 28
randInt(0,30) 4
```

Note: We can provide third input get a specified number of random values. E.g. randInt(0, 30, 3) will generate 3 random numbers, which would yield the answer {29, 28, 4}.

Besides simple random sampling, there are other forms of sampling that involve random chance. Other well-known random sampling methods are the **stratified samples**, **cluster samples**, and **systematic samples**.

To choose a **stratified sample**, divide the population into groups called “strata” and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n th piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 to 20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person). Compare the fractions $9/25$ and $9/24$. To four decimal places, $9/25 = 0.3600$ and $9/24 = 0.3750$. So these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and non-sampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **non-sampling errors**. A defective counting device can cause a non-sampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. In general, the larger the sample, the smaller the sampling error.

In Statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example 1.11

A study is done to determine the average tuition that College of Lake County transfer, career, and adult education students pay per semester. Each student in the following samples is asked how much tuition they paid for the fall semester. What is the type of sampling in each case?

- a) A sample of 100 College of Lake County students is taken by organizing the students' names by classification (transfer, career or adult education), and then selecting 25 students from each.
- b) A random number generator is used to select a student from the alphabetical listing of all college students in the fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
- c) A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
- d) The transfer, career, and adult education are numbered one, two and three respectively. A random number generator is used to pick two of those categories. All students in those two categories are in the sample.
- e) An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 students they encounter what they paid for tuition in the fall semester. Those 100 students are the sample.

Solution 1.11: a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Try It

1.11 You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores for an elementary Statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick simple random samples.

1. Create a stratified sample by column. Pick three quiz scores randomly from each column.
 - Number each row one through ten.
 - On your calculator, press Math and arrow over to PRB.
 - For column 1, select randInt(and enter 1,10. Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
 - Repeat for columns two through six.
 - These 18 quiz scores are a stratified sample.
2. Create a cluster sample by first randomly picking two of the columns. Use the column numbers: one through six.
 - Press MATH and arrow over to PRB.
 - Select randInt and enter 1,6. Press ENTER. Record the number.
 - Press ENTER and record that number.
 - The two numbers are for two of the columns.
 - The quiz scores (20 of them) in these 2 columns are the cluster sample.
3. Create a simple random sample of 15 quiz scores.
 - Use the numbering 1 through 60.
 - Press MATH. Arrow over to PRB. select randInt and enter 1, 60).
 - Press ENTER 15 times and record the numbers.
 - Record the quiz scores that correspond to these numbers.
 - These 15 quiz scores are the systematic sample.
4. Create a systematic sample of 12 quiz scores.
 - Use the numbering one through 60.
 - Press MATH. Arrow over to PRB. Press 5: randInt(and enter 1, 60).
 - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may need to wrap around (go back to the beginning).

Example 1.12

Determine type of sampling used (simple random, stratified, systematic, cluster, or convenience).

- a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three from a group of boys aged 13 to 14 to form a recreational soccer team.
- b. A pollster interviews all human resource personnel in five different high tech companies.
- c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.

- d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
- e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
- f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution 1.12

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f. convenience

Try It

1.12 Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

If we were to examine two samples representing the same population, they would not be exactly the same, even if we used random sampling methods for the samples. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 1.13

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples. First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

\$128; \$87; \$173; \$116; \$130; \$204; \$147; \$189; \$93; \$153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

\$50; \$40; \$36; \$15; \$50; \$100; \$40; \$53; \$22; \$22

It is unlikely that any student is in both samples.

- a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?
- b. If these samples are not representative of the entire population, would it be wise to use the results to describe the entire population?

c. Now suppose we take a third sample. We choose ten different part-time students from the disciplines of Chemistry, Math, English, Psychology, Sociology, History, Nursing, Physical Education, Art, and Early Childhood Development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

\$180; \$50; \$150; \$85; \$260; \$75; \$180; \$200; \$200; \$150

Is this sample biased? Explain.

Solution 1.13

- a. No. The first sample probably consists of science-oriented students. For example, in addition to the chemistry course, some of them are also Calculus or Biology courses. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are likely taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.
- b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.
- c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Try It

1.13 A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Since asking all 20,000 listeners would be impossible, the station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. Of those sampled, 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music. Is this sample representative of the entire 20,000 listener population?

Variation in Data and in Samples

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data selection methods and your accuracy.

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Critical Evaluation

We need to critically evaluate statistical studies we read about and analyze them before accepting the results of the studies. Common problems to be aware of include:

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples include crash testing cars or medical testing for rare conditions.
- Undue influence: collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context.
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

1.3 | Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When changes in one variable produce changes in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables** (or confounding variables). In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given

a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

1. McClung, M. Collins, D. “Because I know it will!” Placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

Example 1.19

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Solution 1.19

The *population* is men aged 50 to 84.

The *sample* is the 400 men who participated.

The *experimental units* are the individual men in the study.

The *explanatory variable* is oral medication.

The *treatments* are aspirin and a placebo.

The *response variable* is whether a subject had a heart attack.

Example 1.20

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

- Describe the explanatory and response variables in this study.
- What are the treatments?
- Identify any lurking variables that could interfere with this study.
- Is it possible to use blinding in this study?

Solution 1.20

- a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
- b. There are two treatments: a floral-scented mask and an unscented mask.
- c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
- d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Example 1.21

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution 1.21

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Try It

1.21 You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

- a. Describe the explanatory and response variables in the study?
- b. What are the treatments?
- c. What should you consider when selecting participants?
- d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
- e. Identify any lurking variables that could interfere with this study.
- f. How can blinding be used in this study?

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[2]

2. Yudhijit Bhattacharjee, "The Mind of a Con Man," *Magazine*, *New York Times*, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

The committee investigating Stapel concluded that he was guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not so easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics." [3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to detect. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other

research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.

- Data collected from individuals must be guarded carefully to protect their privacy. These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

3. “*Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel*,” Tilburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) (<http://www.retractionwatch.com>) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example 1.22

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected. A researcher is collecting data in a community.

- a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
- b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.

c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution 1.22

- a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
- b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
- c. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Try It

1.22 Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California. The survey is commissioned by the seller of a popular brand of apple juice. There are only two types of juice included in the study: apple juice and cranberry juice. Researchers allow participants to see the brand of juice as samples are poured for a taste test. Among the participants, 25% preferred Brand X, 33% preferred Brand Y and 42% had no preference between the two brands. Brand X then references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

KEY TERMS

Average a number that describes the central tendency of the data

Blinding not telling participants which treatment a subject is receiving

Categorical Variable variables that take on values that are names or labels

Cluster Sampling a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable a random variable (RV) whose outcomes are measured; e.g. the height of trees in the forest is a continuous RV.

Control Group a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Data a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable a random variable (RV) whose outcomes are counted

Double-blinding the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit any individual or object to be measured

Explanatory Variable the independent variable in an experiment; the value controlled by researchers

Informed Consent Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board a committee tasked with oversight of research programs that involve human subjects

Lurking Variable a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Non-sampling Error an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable variables that take on values that are indicated by numbers

Parameter a number that is used to represent a population characteristic

Placebo an inactive treatment that has no real effect on the explanatory variable

Population all individuals, objects, or measurements whose properties are being studied

Proportion the number of successes divided by the total number in the sample

Qualitative Data See **Data**.

Quantitative Data See **Data**.

Random Assignment the act of organizing experimental units into treatment groups using random methods

Random Sampling a method of selecting a sample that gives every member of the population an equal chance of being selected.

Representative Sample a subset of the population that has the same characteristics as the population

Response Variable the dependent variable in an experiment; the value that is measured for change at the end of an experiment

Sample a subset of the population studied

Sampling Bias not all members of the population are equally likely to be selected

Sampling Error the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Statistic a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Stratified Sampling a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let $k = (\text{number of individuals in the population}) / (\text{number of individuals needed in the sample})$. Choose every k^{th} individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments different values or components of the explanatory variable applied in an experiment

Variable a characteristic of interest for each person or object in a population

CHAPTER REVIEW

1.1 Basic Definitions

The mathematical theory of statistics is easier to learn when you know the language. This section presented important terms that will be used throughout the text.

1.2 Data and Sampling

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.2 Experimental Design and Ethics

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some because you support, hurts or reduces benefits to others, and violates some rule.”[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

4. Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, <http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf> (accessed May 1, 2013).

EXERCISES FOR CHAPTER 1

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

3; 4; 11; 15; 16; 17; 22; 44; 37; 16;
14; 24; 25; 15; 26; 27; 33; 29; 35; 44;
13; 21; 22; 10; 12; 8; 40; 32; 26; 27;
31; 34; 29; 17; 8; 24; 18; 47; 33; 34.

Researcher B:

3; 14; 11; 5; 16; 17; 28; 41; 31; 18;
14; 14; 26; 25; 21; 22; 31; 2; 35; 44;
23; 21; 21; 16; 12; 18; 41; 22; 16; 25;
33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

1. population
2. sample
3. parameter
4. statistic
5. variable

Use the following information to answer the next five exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

6. The sampling method was:
a. simple random b. systematic c. stratified d. cluster
7. "Number of times per week" is what type of data?
a. qualitative b. quantitative discrete c. quantitative continuous
8. "Duration (amount of time)" is what type of data?
a. qualitative b. quantitative discrete c. quantitative continuous

9. The colors of the houses around the park are what kind of data?
a. qualitative b. quantitative discrete c. quantitative continuous

10. What is the population for this study? _____

11. The table below contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Use this data to answer the following questions:

- What is the proportion of deaths between 2007 and 2012?
- What percent of deaths occurred before 2001?
- What is the percent of deaths that occurred in 2003 or after 2010?
- What is the fraction of deaths that happened before 2012?
- What kind of data is the number of deaths?
- Earthquakes are quantified according to the Richter scale, which measures the amount of energy they produce (ex. are 2.1, 5.0, 6.7). What type of data is that?
- What contributed to the large number of deaths in 2010? In 2004? Explain.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

12. A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

13. A market researcher polls every tenth person who walks into a store.

14. The first 50 people who walk into a sporting event are polled on their television preferences.
15. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in **Table 1.31**. The second study collected the data in **Table 1.32**.

Group	Showed Improvement	No Improvement	Deterioration
Used Program	142	43	15
Did Not use Program	72	110	18

Table 1.31

Group	Showed Improvement	No Improvement	Deterioration
Used Program	105	74	19
Did Not use Program	88	99	12

Table 1.32

16. Given what you know, which study is correct?
17. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?
18. Both groups that performed the study concluded that the software works. Is this accurate?
19. The company which makes the software uses the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?
20. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from question #18?

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

21. A fitness center is interested in the mean amount of time a client exercises in the center each week.

22. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

23. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

24. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

25. A politician is interested in the proportion of voters in his district who think he is doing a good job.

26. A marriage counselor is interested in the proportion of clients she counsels who stay married.

27. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

28. A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

29. What is the population she is interested in?

- a. all Lake Tahoe Community College students
- b. all Lake Tahoe Community College English students
- c. all Lake Tahoe Community College students in her classes
- d. all Lake Tahoe Community College math students

30. Let X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of a:

- a. variable.
- b. population.
- c. statistic.
- d. data.

31. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

- a. parameter
- b. data
- c. statistic
- d. variable

For the following exercises (32 – 40), identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

32. number of tickets sold to a concert

33. percent of body fat

34. time spent in line to buy groceries

35. number of students enrolled at Evergreen Valley College
36. most-watched television show
37. brand of toothpaste
38. distance to the closest movie theatre
39. age of executives in Fortune 500 companies
40. number of competing computer spreadsheet software packages
41. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
- Using complete sentences, list three things wrong with the way the survey was conducted.
 - Using complete sentences, list three ways that you would improve the survey if it were to be repeated.
42. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
43. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
44. List some practical difficulties involved in getting accurate results from a telephone survey.
45. List some practical difficulties involved in getting accurate results from a mailed survey.
46. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.
47. Name the sampling method used in each of the following situations:
- A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
 - A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.

c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.

d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

48. A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had \$2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?

b. Based on your "gut feeling," do you believe the percentages accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percentage of the population is actually higher or lower than the sample statistics? Why?

Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."

c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?

d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

49. The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey as qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?

- b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
- c. In the last seven days, on how many days did you exercise for 30 minutes or more?
- d. Do you have health insurance coverage?

50. In advance of the 1936 presidential election, a magazine titled *Literary Digest* released the results of an opinion poll predicting that the Republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

- a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
- b. What effect does the low response rate have on the reliability of the sample?
- c. Are these problems examples of sampling error or non-sampling error?
- d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

51. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates. Which of the potential problems with samples discussed in **Section 1.2** could explain this connection?

52. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks: "Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[5] As of April 25, 11 people responded to this question. Each participant answered "NO!" Which of the potential problems with samples discussed in this module could explain this connection?

53. A scholarly article about response rates begins with the following quote: "Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[6] The Pew Research Center for People and the Press admits: "The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more." [7]

- a. What are some possible reasons for the decline in response rate over the past decade?
- b. Explain why researchers are concerned with the impact of the declining response rate of public opinion polls.

REFERENCES

1.1 Basic Definitions

Data and Story

Library, <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).

1.2 Data and Sampling

Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/default.asp> (accessed May 1, 2013). Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/methodology.asp> (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. <http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx> (accessed May 1, 2013).

Data from <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

Dominic Lusinchi, “President’ Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?” *Social Science History* 36, no. 1: 23-54 (2012), <http://ssh.dukejournals.org/content/36/1/23.abstract> (accessed May 1, 2013).

“The *Literary Digest* Poll,” Virtual Laboratories in Probability and Statistics <http://www.math.uah.edu/stat/data/LiteraryDigest.html> (accessed May 1, 2013).

“Gallup Presidential Election Trial-Heat Trends, 1936–2008,” Gallup Politics <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4> (accessed May 1, 2013).

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html> (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, <http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus> (accessed May 1, 2013). Data from San Jose Mercury News

1.3 Experimental Design and Ethics

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, <http://www.hsph.harvard.edu/nutritionsource/vitamin-e/> (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” *athleteinme.com*, <http://www.athleteinme.com/ArticleView.aspx?id=1053> (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” *International Business Times*, July 21, 2011. Also available online at <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443> (accessed May 1, 2013).

The Data and Story

Library, <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html> (accessed May 1, 2013).

M.L. Jackson et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” *Accident Analysis and Prevention Journal*, Jan no. 50 (2013), <http://www.ncbi.nlm.nih.gov/pubmed/22721550> (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. <http://earthquake.usgs.gov/earthquakes/eqarchives/year/> (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. <http://www-fars.nhtsa.dot.gov/Main/index.aspx> (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” <http://www.forbes.com/best-small-companies/list/> (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111: Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Lori Alden, “Statistics can be Misleading,” econoclass.com, <http://www.econoclass.com/misleadingstats.html> (accessed May 1, 2013).

Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, <http://cnx.org/content/m15555/latest/> (accessed May 1, 2013).

lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: <http://www.youpolls.com/details.aspx?id=12328> (accessed May 1, 2013).

Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), <http://poq.oxfordjournals.org/content/70/5/759> (accessed May 1, 2013).

Frequently Asked Questions, Pew Research Center for the People & the Press, <http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls> (accessed May 1, 2013).

CHAPTER 1 SOLUTIONS:

- 1) AIDS patients; 3) The average length of time (in months) AIDS patients live after treatment.
- 5) X = the length of time (in months) AIDS patients live after treatment; 7) b;
- 9) a;
- 11) a. 0.5242 b. 0.03% c. 6.86% d. 823,088/823,856 e. quantitative discrete
f. quantitative continuous g. In both years, underwater earthquakes produced massive tsunamis.
- 13) systematic; 15) simple random;
- 17) There is not enough information given to judge if either one is correct or incorrect.
- 19) Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn.
- 21) a. all clients for the fitness center b. a smaller, selected group of these clients
c. the population mean number of hours spent each week in the fitness center
d. the sample mean number of hours spent each week in the fitness center
e. X = the number of hours spent in the fitness center for a given client in a given week.
f. values for X , such as 2, 1.7, 3.5, ...
- 23) a. all patients of the doctor b. a smaller, selected group of these patients
c. the mean recovery period for all patients d. the mean recovery time for the sample patients e. X = the recovery time of a single patient
f. values for X , such as, 9 months, 4.3 months,...
- 25) a. all voters in the district b. a smaller, selected group of these voters
c. the proportion of all her voters in the district who approve of the politician's performance. d. the proportion of the sample who approve of the politician's job performance.
e. X = whether or not a voter approves of the politician's job performance f. yes, no
- 27) a. all voters in the region (county, state or nation) b. a smaller, selected group of these voters c. the proportion of all voters in the region who will vote for the cause.
d. the proportion of the sample who will vote for the cause. e. X = whether or not a voter will vote for the cause f. yes, no
- 29) a 31) c 33) quantitative continuous; e.g. 21.2%
- 35) quantitative discrete; e.g. 11,234 students 37) qualitative; e.g. Crest, Colgate

- 39) quantitative continuous; e.g. 51 yrs, 63.5 yrs
- 41) a. The survey was conducted using six similar flights. The survey would not be a true representation of the entire population of air travelers. Conducting the survey on a holiday weekend will not produce representative results. b. Conduct the survey during different times of the year. Conduct the survey using flights to and from various locations. Conduct the survey on different days of the week.
- 43) Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.
- 45) Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.
- 47) a. convenience b. cluster c. stratified d. systematic e. simple random
- 49) a. qualitative b. quantitative discrete c. quantitative discrete d. qualitative
- 51) Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate. Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.
- 53) a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

This page is purposely left blank.