2 Populations and Samples

CHAPTER OUTLINE

2.1 **Selecting Appropriate Samples**
Explains why the selection of an appropriate sample has an important bearing on the reliability of inferences about a population

2.2 **Why Sample?**
Gives a number of reasons sampling is often preferable to census taking

2.3 **How Samples Are Selected**
Explains how samples are selected

2.4 **How to Select a Random Sample**
Illustrates with a specific example the method of selecting a random sample using a computer statistical package

2.5 **Effectiveness of a Random Sample**
Demonstrates the credibility of the random sampling process

2.6 **Missing and Incomplete Data**
Explains the problem of missing or incomplete data and offers suggestions on how to minimize this problem

✔ LEARNING OBJECTIVES

LEARNING OBJECTIVES

After studying this chapter, you should be able to

1. Distinguish between
   a. populations and samples
   b. parameters and statistics
   c. various methods of sampling
2. Explain why the method of sampling is important

(Continued)

3. State why samples are used
4. Define *random sample*
5. Explain why it is important to use random sampling
6. Select a random sample using a computer statistical program
7. Suggest methods for dealing with missing data

2.1 SELECTING APPROPRIATE SAMPLES

A **population** is a set of persons (or objects) having a common observable characteristic. A **sample** is a subset of a population.

The real challenge of statistics is how to come up with a reliable statement about a population on the basis of sample information. For example, if we want to know how many persons in a community have quit smoking, or have health insurance, or plan to vote for a certain candidate, we usually obtain information on an appropriate sample of the community and generalize from it to the entire population. How a subgroup is selected is of critical importance. Take the classic example of the *Literary Digest* Poll. This poll attained considerable prestige by successfully predicting the outcomes of four presidential elections before 1936. Using the same methods, the *Literary Digest* in 1936 mailed out some 10 million ballots asking persons to indicate their preference in the upcoming presidential election. About 2.3 million ballots were returned, and based on these, the *Literary Digest* confidently predicted that Alfred M. Landon would win by a landslide. In fact, Franklin D. Roosevelt won with a 62% majority. Soon after this fiasco, the *Literary Digest* ceased publication. A postmortem examination of its methods revealed that the sample of 10 million was selected primarily from telephone directories and motor vehicle registration lists, which meant that persons with higher incomes were overrepresented. In 1936, there was a strong relation between income and party preference; thus, the poll's failure was virtually inevitable.

The moral of this incident is clear: The *way* the sample is selected, not its *size,* determines whether we may draw appropriate inferences about the population. Modern sampling techniques can quite reliably predict the winner of a presidential election from a nationwide sample of less than 2000 persons. This is remarkable considering that the nation's population today is more than twice what it was in 1936.

The key to selecting an appropriate sample is that the sample be representative of the population. This means that, whatever variable you are studying, a relatively small sample should very closely approximate the population. The previous two paragraphs present an example in which an extraordinarily large sample did *not* represent the population. The next time a major election

approaches, notice the various polls that attempt to predict the probable winner. Many of these polls will come very close to the actual vote count—and do so with a much smaller sample than was used in 1936.

As another example to illustrate populations and samples, we might identify the health issues of college students and then develop programs or provide health care resources depending upon their needs. Let us consider all the students at your school. Is this a sample or a population? It depends. If you are looking at college students in general, your school would be a subset (i.e., sample) of the population. If only students at your school were sampled, could the information gathered be used to make inferences about all college students in the United States? Probably not. The simple reason is that there is likely to be something unique about your school that makes it nonrepresentative of college students in general.

But let's suppose we are only interested in the health issues for students at your school. Under these circumstances, you would view these students as a population. If you want to know what chronic conditions are present, how much students drink or smoke or use marijuana, what prescription and over-the-counter (OTC) drugs are taken, or what serum cholesterol levels are, how do you go about gathering these data? One way would be to survey or examine every student. Realistically, however, this is simply not possible. So what alternatives do you have? What if you were to take a small sample of the population and then, based on this sample, make accurate inferences about the population. Note that the population may consist of persons, objects, or the observations of a characteristic. The set of observations may be summarized by a descriptive statistic called a **parameter.** When the same characteristic pertains to a sample, it is called a **statistic.**

Suppose your school is large enough to have a pharmacy. What drugs should the pharmacy carry? You would need to determine which prescription drugs are most commonly used by students at your school. To do this, you would select a sample. Remember that you want this sample to represent the population. The preferred, and the most likely, method of obtaining a representative sample is to select a **random sample.** The basic principle behind random sampling is that every subject (in this case, students at your school) has an equal chance of being selected. Although this does not guarantee a representative sample, it is the technique most likely to yield a representative sample.

Random sampling might seem to be the best route, but there are often considerable obstacles. Let us assume that you we have a health questionnaire that includes questions about prescription drugs. Your goal is to select a random sample of students to complete this questionnaire. Selecting a random sample is fairly easy. Using a statistical program, you can assign each student a number and then generate a random sample. You next send the questionnaire to the selected students. Here is the problem: What will your **response rate** be? How many students will return the questionnaire? Without resorting to various techniques for improving the response rate, you may be doing well to get a 50% return. Remember that the primary purpose of random sampling is to obtain a

representative sample and then, based on the sample statistics, to make inferences about the population.

Is a 50% response rate good enough to ensure a representative sample? Often, you simply do not know. Think about the situation at your school, and ask yourself whether students who complete and return a survey are different from those who do not. Researchers often assume that the responses are representative, but that assumption may not be valid. This leads to possible **sampling bias.** Bias occurs because some segment of the population might be either over or underrepresented. For example, females, or freshman, or health science majors may be more likely to respond than males, seniors, or English majors. Thus, the statistics gathered from such a survey may have some **bias,** which means the sample may not be representative of the population. Without getting into specific techniques for increasing the response rate, whatever technique improves the rate is likely to ensure that the sample is representative of the population.

Often, however, a random sample is not feasible. It may be too costly to randomly select a group and then mail out a survey. In the case of a clinical study of a specific treatment for a particular disease, it may be of no value to have three patients from California, one from Idaho, two from Tennessee, and so on. You need to identify an adequate number of subjects quickly and efficiently. At a college or university, you might include all members of one or more classes until you have the sample size you want. This is often referred to as a **convenience sample.** Let us assume that the stats class you are currently in is selected. Most, if not all of you, will probably give your **informed consent** to participate. If the researcher is trying to find out something about your health status, you might complete a questionnaire, or perhaps some type of measurement will be taken, such as height, weight, or blood pressure. Although the researcher is able to gather large amounts of data quickly and efficiently, the key question remains: Is this sample representative of the population? If your class is used, think about how your class differs from and is similar to your school's population. Presumably, you have a higher percentage of students pursuing some type of health major. Are health majors different from other students? You probably also have a disproportionate number of students at a certain grade level. Perhaps there are gender, ethnic, or religious differences that may impact the results. Although a convenience sample may be quick and efficient, and yield large numbers, it still may have limited utility because it is not representative of the population.

Clinical trials frequently employ a nonrandom sample. If a new drug, medical device, or medical procedure is being tested, there may be only one or two test sites. The sample would be those patients who met the criteria for testing the new drug, device, or procedure. The frequent assumption is that, whatever results are found at the test site, similar results will be obtained at other sites and with different subjects. This assumption is tested when the new product is approved for general use and researchers find that either the product is indeed useful or it does not meet the original expectations. A large nonrandom study is the STAR Trial, which was briefly discussed in Chapter 1. In this study, 22,000

postmenopausal women who are at an increased risk of breast cancer are being recruited from more than 500 centers. By the time you read this, there may be some reported results. Because the sample is so large and the scope is so broad (more than 500 centers), there will probably be little or no discussion about the sample. The assumption will be that the results are generalizable to the population of postmenopausal women who are at an increased risk of breast cancer.

2.2 WHY SAMPLE?

You may be wondering, "Why not study the entire population?" There are many reasons. It is *impossible* to obtain the weight of every tuna in the Pacific Ocean. It is too costly to inspect every manifold housing that comes off an assembly line. The Internal Revenue Service does not have the workforce to review every income tax return. Some testing is inherently destructive: tensile strength of structural steel, flight of a solid propellant rocket, measurement of white blood cells. We certainly cannot launch all the rockets to learn the number of defective ones; we cannot drain all the blood from a person and count every white blood cell. Often we cannot justify enumerating the entire population—that is, conducting a census—because for most purposes we can obtain suitable accuracy quickly and inexpensively on the basis of the information gained from a sample alone. One of the tasks of a statistician is to design efficient studies utilizing adequate sample sizes that are not unnecessarily large. How to determine a sample size that is likely to give meaningful results is discussed in Chapter 9.

2.3 HOW SAMPLES ARE SELECTED

At this point, you should begin to recognize the importance of selecting a sample and then using that sample to draw inferences about the population. But how reliable are our inferences regarding a population? The answer depends largely on how precisely the population is specified and on how the sample is selected. Having a poorly specified or enumerated population or an inappropriately selected sample will almost certainly introduce bias. Although you cannot guarantee that bias will be eliminated, you can take steps to control it. The best way to limit bias is to use random sampling, a technique that is simple to apply (it is sometimes called "simple random sampling"). The basic principle behind random sampling is quite straightforward: Each subject in the population has an equal chance of being selected. Cappelleri and Trochim (2000:149) succinctly describe three important reasons for random sampling: "(1) It avoids known and unknown biases on average; (2) it helps convince others that the trial was conducted properly; and (3) it is the basis for statistical theory that underlies hypothesis tests and confidence intervals." As you progress through this chapter, reasons 1 and 2 should become increasingly clear.

As for reason 3, the key point to remember is that a random sample allows us to draw the most representative sample and then, based on the sample statistics, to make inferences about the population. You will study many of these statistical methods as you progress through your statistics class.

Samples can be selected in several other ways. In convenience sampling, as previously noted, a group is selected at will or in a particular program or clinic. These cases are often self-selected. Because the data obtained are seldom representative of the underlying population, problems arise in analysis and in drawing inferences.

Convenience samples are often used when it is virtually impossible to select a random sample. For instance, if a researcher wants to study alcohol use among college students, each member of the population—that is, each college student—ideally would have an equal chance of being sampled. A random sample of 100, 1000, or 10,000 college students is simply not realistic. How will the researcher collect data about alcohol use among college students? Often he or she will survey college students enrolled in a general education course, such as English 101. The underlying assumption on the part of the researcher is that a general education class, which most or all students must take, is a representative sample of college students and therefore accurately represents alcohol use at that college or university. The logical next step is to assume that the colleges or universities surveyed are representative in terms of college and university students' alcohol use. You can see how the use of a convenience sample may eventually lead researchers to inferences about alcohol use among college students that are inaccurate or misleading.

Systematic sampling is frequently used when a **sampling frame**—a complete, nonoverlapping list of the persons or objects constituting the population—is available. We randomly select a first case and then proceed by selecting every nth (say $n = 30$) case, where n depends on the desired sample size. The symbol N is used to denote the size of the entire population.

Stratified sampling is used when we wish the sample to represent the various **strata** (subgroups) of the population proportionately or to increase the precision of the estimate. A simple random sample is taken from each stratum.

In **cluster sampling,** we select a simple random sample of groups, such as a certain number of city blocks, and then interview a person in each household of the selected blocks. This technique is more economical than the random selection of persons throughout the city.

For a complete discussion of the various kinds of sampling methods, you should consult a textbook on the subject.

2.4 HOW TO SELECT A RANDOM SAMPLE

Selecting a random sample is one of the easiest procedures you will learn. The process can be as basic as putting all the names of a population on slips of paper, putting the slips in a container, mixing them up, and selecting however many

names you want. Many calculators will generate random numbers. Also, a random numbers table is included in Appendix E. However, the most widely used technique, and therefore the focus of this section, is the use of a computer program to select random numbers. Without attempting to explain how, we will simply state that computer statistical packages such as SPSS are programmed to allow for the quick selection of random numbers.

In Chapter 3, Table 3.1 lists characteristics of 100 of the 7683 participants in the Honolulu Heart Study, which investigated heart disease among men ages 45–67. This data set is also included in a spreadsheet that is available on this book's Web site at www.mhhe.com/Kuzma5e. Suppose we want to randomly select 25 cases from the 100 cases contained in Table 3.1. If you are using SPSS (we used version 11.5), here is how to select a random sample of 25 from the 100 cases listed in Table 3.1. First, go to the menu and find the Data box, which is the fourth box after File, View, and Edit. From the Data Box, go down to the next-to-last choice, Select cases. Once you have found Select cases, choose Random Sample of Cases. In the Select Cases box, next choose Random Sample of Cases. Then hit the button just below Random Sample of Cases, labeled Sample. Another box, Select Cases: Random Sample, will appear. Under Sample size, you will have two choices: (1) Approximately X% of all cases, or (2) Exactly X cases from the first XX cases. If you choose option 1, all you need to do is pick a number between 1 and 99, and that percentage of cases will be selected. If you choose option 2, you can select the exact number you want where you see the X (blank space in the Dialog box). Where you see XX (another blank in the Dialog box), you will typically enter the total number of the subjects in the data set. Using the data from Table 3.1, we would enter the number 100.

Let us use option 2 and choose 25 from the 100 cases in the Honolulu Heart Study. Enter 25 in the first space and 100 in the second. Click OK. Next, look at your screen with 100 subjects. If you scroll down, you will notice that some of them have a slash across the number. There are 25 numbers selected that do not have a slash. This is your random sample of 25. The 75 numbers that have slashes represent the 75 cases the computer did not randomly select. You may wish to repeat this process. Note that each time you repeat the process a different set of 25 is selected. In the next section, we will use this process to illustrate how a random sample can provide you with a sample average that comes very close to the average of the 100 subjects in this data set. At least, this will happen most, but not all, of the time.

2.5 EFFECTIVENESS OF A RANDOM SAMPLE

Students who encounter random sampling for the first time are somewhat skeptical about its effectiveness. The reliability of sampling is usually demonstrated by defining a fairly small population and then selecting from it all conceivable samples of a particular size, say, three observations. Then, for each sample, the

mean (average) is computed and the variation from the population mean is observed. A comparison of these sample means (statistics) with the population mean (parameter) neatly demonstrates the credibility of the sampling scheme.

In this chapter, we try to establish credibility by a different approach. Let us return to Table 3.1, which lists characteristics of a representative sample of 100 individuals from the 7683 participants in the Honolulu Heart Study, which investigated heart disease among men ages 45–67. Five separate samples of 100 observations each were selected from this population, and the mean ages were compared with the population mean. The results of this comparison are shown in Table 2.1. We can see that the population parameter is 54.36 and that the five statistics representing this mean are all very close to it. The difference between the sample estimate and the population mean never exceeds 0.5 year, even though each sample represents only 1.3% of the entire population. This comparison underscores how much similarity you can expect among sample means.

Let us take another random sample, using SPSS, from the 100 subjects listed in Table 3.1. First, display the data from Table 3.1 on your computer screen. Next, have the computer display the mean (average) age for the 100 subjects. To find the mean, choose Analyze from the menu, and scroll down to the second item, which is Descriptive Statistics. Within the Descriptive Statistics category, you have five choices. Choose the second one, Descriptives. Select variable number 4, which is Age, and then click OK. You will see a display indicating, among other things, that the number (N) of total subjects is 100 and that the mean (average) is 53.67.

Based on the 100 subjects, next select five random samples of 20, 30, and 40 subjects, using the method described in the previous section. If you write down the sample means (averages), notice how most, if not all, of them should be reasonably close to the average age of 53.67. Again, this illustrates how a randomly selected sample can closely approximate the numerical values of the population. Table 2.2 shows the values we obtained from the five samples of 20, 30, and 40.

2.6 MISSING AND INCOMPLETE DATA

In the opening section of this chapter, we discussed the problem of obtaining a random sample from a well-defined population, such as all students at your school. Regardless of whether you are conducting a survey or using some other data collection technique, if you are only able to collect data from some subjects in your sample, the sample may not be representative of the population. If your sample is biased, there is the very real risk that the sample statistics may differ significantly from the population statistics or parameters. For example, what is the average (mean) number of alcoholic drinks consumed by college students, over a specific time period? Suppose you select a random sample from all students at your school and mail each student a questionnaire. Whatever

Table 2.1 Hypertension Study Cases by Diastolic Blood Pressure, Sex, and Dietary Status

ID	Diastolic Blood Pressure (mmHg)	Sex	Vegetarian Status	ID	Diastolic Blood Pressure (mmHg)	Sex	Vegetarian Status
01	88	M	V	42	70	M	NV
02	98	M	V	43	102	M	NV
03	64	M	V	44	84	M	NV
04	80	M	V	45	74	M	NV
05	60	M	V	46	76	M	NV
06	68	M	V	47	84	M	NV
07	58	M	V	48	84	M	NV
08	82	M	V	49	82	M	NV
09	74	M	V	50	82	M	NV
10	64	M	V	51	74	M	NV
11	78	M	V	52	70	M	NV
12	68	M	V	53	92	M	NV
13	60	M	V	54	68	M	NV
14	96	M	V	55	70	M	NV
15	64	M	V	56	70	M	NV
16	78	M	V	57	70	M	NV
17	68	M	V	58	40	M	NV
18	72	M	V	59	83	M	NV
19	76	F	V	60	74	M	NV
20	68	F	V	61	56	F	NV
21	70	F	V	62	89	F	NV
22	62	F	V	63	84	F	NV
23	82	F	V	64	58	F	NV
24	58	F	V	65	58	F	NV
25	72	F	V	66	82	F	NV
26	56	F	V	67	78	F	NV
27	84	F	V	68	82	F	NV
28	80	F	V	69	71	F	NV
29	56	F	V	70	56	F	NV
30	58	F	V	71	68	F	NV
31	82	F	V	72	58	F	NV
32	88	F	V	73	72	F	NV
33	100	F	V	74	80	F	NV
34	88	F	V	75	88	F	NV
35	74	F	V	76	72	F	NV
36	60	F	V	77	68	F	NV
37	74	F	V	78	66	F	NV
38	70	F	V	79	78	F	NV
39	70	F	V	80	74	F	NV
40	66	F	V	81	60	F	NV
41	76	M	NV	82	66	F	NV
				83	72	F	NV

NOTE: ID = identification; mmHg = millimeters of mercury; V = vegetarian; NV = nonvegetarian.

Table 2.2 Effectiveness of a Random Sample ($N = 100$)

Sample Size (1)	Average (Mean) Age				
	1	2	3	4	5
20	52.95	53.50	54.35	53.45	51.80
30	52.80	53.60	53.70	53.57	52.97
40	53.15	53.70	54.10	53.60	53.15

technique you use to distribute, the questionnaire will almost certainly leave you with a significant percentage of nonrespondents. If you use a true random sample, is it possible that the alcohol consumption levels of the respondents may differ from those of nonrespondents? If, instead of a random sample, you use a convenience sample, such as composition or general education classes, is it possible that those students attending class and completing the questionnaire might differ from those absent in terms of alcohol consumption? It is possible.

It is important to recognize that bias may be introduced because of possible differences between respondents and nonrespondents. This may limit the ability to accurately draw inferences about the population. The next question is, What can be done to determine if bias exists? Frequently, the answer is nothing. The researcher simply states that the response rate is adequate or the convenience sample is representative and then proceeds with the analysis. Or the researcher may acknowledge a possible weakness or bias and make some type of qualifying statement about the results. Yet another possibility is to try to identify some objective method of determining if the nonrespondents differ from the respondents.

If this problem were posed to you, at your school, how might you objectively determine if nonrespondents differ from respondents? One possibility would be to determine whether any population data exists for students. For example, most schools could give you the number of males and females, their years in school, and their grade point averages (GPA). If your questionnaire also asked for the respondent's gender, GPA, and year in school, you could easily compute the sample percentage of males and females; the sample percentage of freshman, sophomores, and so on; and the sample GPA. Then you could compare that to the population parameters already identified. What if you find that your sample has a higher GPA than your population. Remember that you are trying to determine how much alcohol is consumed overall, based on a sample. If the GPA is indeed higher for the sample, would you question the validity of taking the sample alcohol consumption data and generalizing to the population? Maybe so. It is quite plausible, and some might argue probable, that students with lower GPAs might consume more alcohol than their peers.

Therefore, a sample with a higher GPA might potentially underestimate the amount of alcohol consumed by students at your school.

The previous discussion focused on the problem of nonrespondents when data are collected only once. What about a situation in which there are multiple observations from the same subject? These are referred to as **longitudinal studies.** This is the case in most clinical trials. Subjects enter the study with some specific criteria such as a disease or condition at a well-defined stage or grade. An example might be patients with moderately elevated serum cholesterol levels entering a study to test the efficacy of a new cholesterol-lowering drug. The study protocol will require serum cholesterol levels to be measured at various times.

Whenever you have a study that requires two or more data collection points, you run the risk that subjects, for a variety of reasons, will drop out, and you now have a problem with missing data. How do statisticians handle the problem of missing data? One potential method, again, is simply to ignore it. Therefore, your analysis includes only those subjects who complete the study. But there are obvious problems with this approach. These center around the reason(s) subjects leave the study. If you are doing a clinical test of a new drug, such as a cholesterol-lowering drug, what are plausible reasons for dropping out? Some reasons might be relatively benign, such as subjects moving away from the study site or simply failing to keep appointments. However, other reasons might directly relate to your ability to determine the efficacy of the new drug. Perhaps a subject's condition worsened rather than improved; perhaps some subjects were experiencing unpleasant or even dangerous side effects; or perhaps one or more subjects died during the study period. All of these reasons would lead you to question the efficacy of the new drug. However, if all subjects with missing data are dropped from the study, what is left is an analysis of only those subjects who either had a beneficial reaction to the new drug or at least avoided some of the more serious adverse events. Obviously, conclusions drawn solely from those who completed the study would be legitimately questioned. Many peer-reviewed journals would reject articles that ignored dropouts and only analyzed data from subjects who completed the study.

This analytical challenge is known as the missing data problem. Fortunately, there are ways of dealing with the problem. One commonly used technique is a type of **carry-forward analysis** called **last observation carry-forward** (LOCF). Ting (2000) describes an example in which subjects in a 4-week clinical study were evaluated at week zero (frequently referred to as the baseline) and at weeks 1, 2, 3, and 4. Drug efficacy measurements were taken at five points during this 4-week study. With LOCF, subjects may drop out anytime after the baseline data is collected. If you are using the LOCF technique for missing data, you "take the last observed value prior to dropout from each of these subjects and treat them as final data (i.e., carry forward the last observation from each subject who dropped out before the week-four evaluation)" (Ting, 2000:104). A hypothetical example of a 4-week trial to lower serum cholesterol is shown in Table 2.3.

Table 2.3 Cholesterol Levels with Missing Data (Simulated)

Subject	Baseline	Week 1	Week 2	Week 3	Week 4	LOCF
1	281	280	272	256	242	NA
2	291	294	298	—	—	298
3	287	275	266	266	—	266
4	266	252	279	249	240	NA
5	295	294	271	268	232	NA
6	273	273	260	288	—	288
7	254	231	240	226	222	NA

Because dropouts are frequently treatment failures, using this technique may give a more accurate picture of the true efficacy of the drug being evaluated. A more complete discussion of carry-forward analysis may be found in an article by Ting (2000).

◆ CONCLUSION

Assessing all individuals in a population may be impossible, impractical, expensive, or inaccurate, so it is usually to our advantage to instead study a sample from the original population. To do this, we must clearly identify the population, be able to list it in a sampling frame, and utilize an appropriate sampling technique. Although several methods of selecting samples are possible, random sampling is usually the most desirable technique. It is easy to apply, limits bias, provides estimates of error, and meets the assumptions necessary for many statistical tests. Missing or incomplete data can also introduce bias. Carry-forward analysis is one technique for accounting for missing or incomplete data. The effectiveness of random sampling can easily be demonstrated by comparing sample statistics with population parameters. The statistics obtained from a sample are used as estimates of the unknown parameters of the population.

◆ VOCABULARY LIST

bias
carry-forward analysis
cluster sample
convenience sample
informed consent
last observation carry-forward
longitudinal study
parameter
population
random sample
response rate
sample
sampling bias
sampling frame
statistic
stratified sample
stratum (*pl* strata)
systematic sample

◆ EXERCISES

2.1 Describe the differences between
a. a parameter and a statistic
b. a sample and a census
c. a simple random sample and a convenience sample

2.2 Describe the steps you would take if you were asked to determine the
a. proportion of joggers in your community
b. number of workers without health insurance at companies with fewer than 100 workers in your community
c. number of pregnant women not obtaining prenatal care in your community
d. number of homeless people in your community

2.3 Why is the way a sample is selected more important than the size of the sample?

2.4 List and briefly explain the importance of the three reasons for random sampling.

2.5 a. Explain why a convenience sample, such as the selection of students in one or more general education classes at your college or university, might not be representative of all students at your institution.
b. Explain why students at your college or university might be representative of college students in general.

2.6 a. In what ways are a random sample, convenience sample, and systematic sample different? In what ways are they similar?
b. Describe a situation in which it would be appropriate and more efficient and cost-effective to use a random sample, convenience sample, systematic sample, or cluster sample.

2.7 Using the data from Table 2.1 and a computer statistical package such as SPSS, choose the variable Diastolic Blood Pressure and find the mean (average) for the 83 subjects. Choose samples of 15, 25, and 35, and find the mean of each of these samples. Note how close the various sample means are to the overall mean of the 83 subjects.

2.8 Suppose in the previous exercise you had selected the sample by taking two simple random samples of 10 from each of the two dietary groups. What label would you apply to such a sample?

2.9 Using the data from Table 3.1 and a computer statistical package such as SPSS, choose the variable Serum Cholesterol Level and find the mean (average) for the 100 subjects. Choose samples of 20, 30, 40, 50, and 60, and find the mean of each of these samples. Note how close the various sample means are to the overall mean of the 100 subjects.

2.10 Describe the population and sample for
a. Exercise 2.7
b. the data in Table 3.1

2.11 Explain why missing or incomplete data may affect the validity and bias the results of a study.