

UNIT 13

Samples and Populations

Objectives:

- To distinguish between descriptive statistics and inferential statistics
- To distinguish between a sample and a population
- To distinguish between the population of interest and the accessible interest
- To distinguish between a random sample and a non-random sample
- To understand methods for selecting a simple random sample or a systematic sample
- To understand the limitations of using a convenience sample

All of the topics that we have covered up to this point have focused exclusively on *descriptive statistics*, which is the use of tables, graphs, numerical quantities, etc. to describe a data set. We now want to begin discussing topics leading to *inferential statistics*, which is the use of information gathered from a data set to draw a conclusion about a set of items larger than those on which the data was taken.

To illustrate the difference between descriptive statistics and inferential statistics, suppose that 2000 voters in a city are polled prior to an upcoming election for mayor, and we find that 800 of those polled intend to vote for Olsen for Mayor. If we merely report that "40% of 2000 voters polled in a city intend to vote for Olsen for Mayor," then we would be simply using a percentage to describe a feature of our data; this is an example of using descriptive statistics. On the other hand, if we were to draw the conclusion that "based on the polling data, Olsen does not have enough votes to guarantee that he will win the election for mayor over the only other opponent," then we would be drawing a conclusion about all voters in the city from observations that were made only on some of the voters; this is an example of using inferential statistics.

As another illustration suppose that 500 households in a city are surveyed, and the yearly income is recorded for each household. If we merely report that "the average of yearly incomes recorded for 500 households in a city is \$27,408," then we would be simply using a mean to describe a feature of our data; this is another example of using descriptive statistics. On the other hand, if we were to draw the conclusion that "based on the mean income for the 500 households, the mean yearly income per household for entire the city is below \$30,000," then we would be drawing a conclusion about all households in the city from observations that were made only on some of the households; this is another example of using inferential statistics.

Essentially, descriptive statistics simply refers to a factual presentation of information about a set of items on which data was observed. Inferential statistics goes beyond merely making a statement about the items on which data was observed; the purpose of inferential statistics is to draw one or more conclusions about a larger set of items than those on which the data was observed, whereas there is no such intent with descriptive statistics.

In statistical terminology, a *population* refers to a set of items about which we wish to draw one or more conclusions, and a *sample* refers to a finite set of items which are selected from the population and on which conclusions about the population are based. In the earlier illustration concerning the election of mayor, all voters in the city comprise the population, and the 2000 voters polled make up a sample. In the illustration concerning yearly income per household, all households in the city comprise the population, and the 500 households surveyed make up a sample.

Actually, there is a bit of harmless ambiguity associated with the terms *population* and *sample*. Depending on the context, we may use the terms *population* and *sample* to represent the items on which a specific variable is measured, or we may use the terms *population* and *sample* to represent the actual values of the variable. For instance, in the illustration concerning yearly income per household, we might use the term *population* to refer to all households in the city or to refer to all values of yearly income per household in the city; similarly, we might use the term *sample* to refer to the 500 households surveyed or to refer to observed values of yearly income per household from the 500 households surveyed. We will switch back and forth between these two subtly different interpretations of the terms *population* and *sample* as is convenient in a particular context, but this should not really cause any confusion.

In general, a population may consist of an infinite number of items or a finite number of items; however, a sample must always consist of a finite number of items, since no human being could ever hope to collect data consisting of an infinite number of observations! For example, in the previous illustration concerning the election of mayor, the population of all voters in the city might be huge, but it will still be finite. On the other hand, suppose we wanted to draw some conclusions about the population consisting of the weights of a particular type of pear grown under certain conditions. There is really no end to the number of such pears

that could theoretically ever be grown, although human beings will only be able to ever observe a finite number of such pears. The population of all such pear weights would indeed be infinite.

As a general rule, many large populations, such as all voters in a city, are treated as infinite, even though this is technically not true. The distinction between finite and infinite populations has no practical impact if the size of the sample is less than 10% of the size of the population, which is often the case. Unless otherwise indicated, we shall always assume that a population can be treated as infinite for all practical purposes.

If our intention is to base conclusions about a population on a sample selected from the population, then we most certainly want to select a sample which will be very likely to supply us with an accurate representation of the population. Some famous historical examples can be cited where a very misleading conclusion was made as a result of selecting a sample that was far from an accurate representation of the population about which the conclusion was drawn.

The 1936 presidential election between Landon and Roosevelt provides one such example. The *Literary Digest*, a magazine which had correctly predicted the previous 1932 election, polled ten million voters and assured its readers that Landon would receive more than twice as many votes as Roosevelt. Shortly after Roosevelt won the election easily, the *Literary Digest* ceased to exist. Even though the sample on which the *Literary Digest* based its conclusions was huge, this sample was far from representative of the entire voting population. The hard lesson that had been demonstrated was that no matter how incredibly large a sample may be, conclusions from the sample are effectively worthless if the sample grossly misrepresents the population about which conclusions are to be made. What is sometimes amazing, however, as we shall see later on, is the amount of accuracy that we can achieve with a seemingly small sample when that sample is selected appropriately.

What was wrong with the sample used by the *Literary Digest*? (The *Literary Digest* had used the same sampling procedure in the previous 1932 election, and had predicted that election correctly.) The magazine had selected its sample from its subscribers and by calling people on the telephone. However, in 1936, only the wealthier citizens had telephones and subscribed to the *Literary Digest*, and this was far from representative of the population of voters in the country. Consequently, the sample used by the *Literary Digest* was extremely biased; when this sampling procedure was used for the 1932 election, it was just coincidence, not the accuracy from sampling, that resulted in the *Literary Digest's* correct prediction. Examples similar to what occurred in 1936 also occurred in the 1948 and 1952 presidential elections.

It is very rare that we are able to select a sample from a population with a guarantee that the sample will adequately represent the population. Sometimes, we are not even to sample from the entire population in which we are interested. Suppose, for instance, that we want to obtain a sample of college students for a study. It is unrealistic to think that we could ever have access to the population of all college students, because this population is much too large and is spread out over a huge geographical area. In practice, we must often settle for sampling from an *accessible population*, which would consist only of those items to which we have access. In a study concerning college students, the researchers would most likely only be able to obtain subjects from a few accessible colleges. It is then important to realize that the conclusions drawn from the sample can really only be applied to the accessible population of college students, since no other college student ever had a chance to be selected. In general, the conclusions made from a sample can be applied only to the accessible population from which the sample came, since we cannot be certain that a sample representing the accessible population adequately represents the larger population.

Sometimes it is possible to obtain a list of all the items in a population from which a sample is to be taken. Such a list is called a *sampling frame*. In the previous illustration about a study concerning college students, it is reasonable to think that a list of all the college students in the accessible population could be made available; such a list would be a sampling frame that might be used in the selection of a sample. As another example, if we wanted a sample of blood pressures from employees at a particular factory, we must first select the sample of employees whose blood pressures will be measured. A list of the employees of the factory would be a sampling frame and could be very helpful in the selection of a sample. It is reasonable to assume that such a list could be made available. On the other hand, if we wished to obtain a sample of the amounts of cereal per box from the boxes of cereal produced at a particular factory, no sampling frame would exist.

With some sampling methods, access to a sampling frame is extremely helpful; with other methods, a sampling frame is not a necessity. When a sampling frame is used to obtain a sample, how well the selected sample represents the population of interest will be influenced by the accuracy of the sampling frame. If we were to poll a sample of voters in a certain voting district, obtaining an accurate list of all voters might not be

anywhere near as easy as obtaining an accurate list of all employees of a factory. We might attempt to poll a sample of voters by randomly selecting names from a phone directory; that is, the phone directory would be used as our sampling frame. The inaccuracy of such a sampling frame would arise from the fact that some voters have unlisted phone numbers, some voters do not have phones, and not everyone who has a phone will be necessarily be a registered voter. The *Literary Digest* has already provided us with an example of how wrong our conclusions can be when our sample is selected from an accessible population which is quite different from the population to which we intend to apply our conclusions.

The procedure we use to select items from our accessible population is called a *sampling design*. Each of the many possible sampling designs can be classified as providing either a *random sample* or a *non-random sample*. A random sample is one where items in the population are selected by chance after assigning some non-zero probability of being selected to each item; a non-random sample is one where items are not selected by chance. A non-random sample is often some type of *convenience sample*. As the name implies, a convenience sample is one consisting of items which are easily available. The *Literary Digest* conducted its poll to predict 1936 presidential election on a convenience sample of voters consisting primarily of the voters that were most easily accessible (i.e., its subscribers and those who could be reached by phone).

For another example of a convenience sample, let us again consider a study where we want to obtain a sample of college students. Suppose we have decided to sample from an accessible population consisting of students from a few nearby colleges. Selecting only those students who volunteer to participate in the study would result in a convenience sample. There would then be two limitations to such a study: one limitation resulting from the fact that the sample was selected from an accessible population instead of from the population of interest, and another limitation from the fact that the sample was conveniently selected by using only volunteers.

Using a convenience sample is generally considered a limitation. In certain situations, the limitation from using a convenience sample can be reduced. Instead of merely choosing items which are easily available, it might be possible to select items which are deemed to represent accurately the population of interest based on some expert knowledge and judgment about the characteristics of the population of interest. However, without any real way to make an assessment of the accuracy of the knowledge and judgement, such a sample is still a limitation.

In order to be able to assess the likelihood that selected items adequately represent the sampled population, a random sample is generally preferred over a non-random sample. There are many different types of random samples which can be selected, but the most basic of these is a *simple random sample* of size n . Formally, we define a simple random sample to be one selected in such a way so that each item in the sampled population has an equal chance of being selected and is selected independently of any other item. Note that our definition of a simple random sample is not based just on each item in the sampled population having an equal chance of being selected; we must also insure that items are selected independently of each other. A sampling procedure where each item has an equal chance of being selected, but the items are not selected independently of one another, is not simple random sampling.

To illustrate, let us suppose that we want a sample of $n=2$ names from the following list of four names: Patrice, Julia, Daniel, James. If we use simple random sampling, there are six possible samples which could be selected, and each name would have an equal chance, a $1/2$ probability, of being the selected sample. These six possible samples are as follows:

{Patrice, Julia}	{Patrice, Daniel}	{Patrice, James}
{Julia, Daniel}	{Julia, James}	{Daniel, James}

Each person's name appears in three of the samples, which is why each person has an equal chance, a $3/6 = 1/2$ probability, of being selected.

On the other hand, if we sample by flipping a fair coin and choosing either the two males or the two females depending on whether heads or tails results, the only two possible samples are as follows:

{Patrice, Julia} {Daniel, James}

Each of these samples has a $1/2$ probability of being the selected sample. Since each person's name appears in exactly one of the two possible samples, each person has a $1/2$ probability of being selected. However, since it is only possible to select samples containing people of the same sex, any sample containing people of different sexes has a zero probability of being the selected sample. Even though each person has an equal chance of

being selected, this sampling procedure is not simple random sampling, because males and females are not selected independently of one another.

With simple random sampling, each sample of size n has an equal chance of being chosen. As we have just seen though, if items are not selected independently of one another, each sample does not have an equal chance of being the selected sample. There are many other types of random samples other than simple random samples, but since so many statistical procedures are based on simple random sampling, it is common to refer to a “simple random sample” merely as a “random sample” or to refer to “simple random sampling” merely as a “random sampling.” Hopefully, this will not cause any confusion, as it should be clear from the context of a discussion what is meant.

Although the principle behind simple random sampling is a simple one (hence, the name!), obtaining a simple random sample in practice is often not easy. For example, let us suppose we are able to label all the items in population with different numbers, we write each number on a slip of paper, and we place all the slips of paper in a hat. Are we truly obtaining a simple random sample when we reach into the hat and select the desired number of slips of paper? We can reasonably assume the answer is yes, if we are careful in mixing up and selecting the slips of paper. However, such physical randomization procedures do not always work as well as we might expect. One famous example where physical randomization went awry was with the 1970 draft lottery.

The goal in the 1970 draft lottery was to randomly assign the integers ranging from 1 to 366 to each of the calendar days in 1952, which was a leap year (since Those born in 1952 would be 18 years old in 1970). Each of the calendar days of 1952 was recorded on a slip of paper and put into a capsule. The 366 capsules were mixed into a bin. The capsules were then selected from the bin one at a time; the date in the first capsule selected was assigned the integer 1; the date in the second capsule selected was assigned the integer 2; etc. It turned out however that the mixing of the capsules was not well done, and those with birthdays at the end of the year tended to get lower draft numbers. The correlation between date and draft number was found to be very significant. (If the assignment of integers to calendar days had been truly random, the probability of obtaining a correlation as strong as the one which was found is less than 0.001!) This illustrates the care that must be taken when using physical randomization procedures.

Even when great care is exercised, physical randomization procedures can be a tedious chore. Coin tosses, dice rolling, selecting slips of paper from bins, etc. might be employed satisfactorily when the amount of randomization needed is on a small scale, but these can be difficult and time consuming when randomization on a large scale is needed, as is the case in many practical situations. Consequently, randomization is often best done by using a computer program designed to simulate the generation of random numbers. Programmable calculators and many computer software packages, such as spreadsheets, are often capable of generating random numbers.

For those not having access to appropriate software or to a programmable calculator to generate random numbers, random number tables which have been produced by computer are available. A *random number table* is a sequence of single digits chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 in such a way so that each of the ten integers has an equal probability (1/10) of being selected, and each digit is selected independently of all other digits selected. Table B.1 is a brief table of random numbers generated by computer. More extensive tables of random numbers are available. A book titled *One Million Random Digits* produced by the RAND Corporation has been widely used in sampling design and experimental design.

After appropriately using integers to label all the items from which a sample is to be selected, we can select a simple random sample by reading entries in a random number table beginning with an arbitrary starting point. The way in which the entries in the table were generated ensures that the resulting sample can be treated a simple random sample. To illustrate, let us imagine that we want to select a simple random sample of size $n=8$ from 456 employees of a factory. We shall demonstrate how Table B.1 could be utilized to select the sample.

First we obtain a list of the names of all 456 employee; then we label the names 001, 002, 003, ..., 456. Next, we choose an arbitrary starting point in Table B.1 and read distinct sets of three-digit entries between 001 and 456 inclusive until we obtain a sample of size 8. (The entries of Table B.1 are organized in groups of five in order to make it easier to use the table.) Suppose we choose to start with the 26th row of the first page of Table B.1, and we read the first three digits of each set of five moving across the rows until the desired sample size is obtained. The entries read are as follows:

238 169 423 973 000 120 090 554 115 675 483 747 544 954 888 154 710 711 159 .

Our simple random sample of size $n=8$ would consist of the names labeled as follows:

238 169 423 120 090 115 154 159 .

All other labels we read from the table are ignored because they are not between 001 and 456. Occasionally, it may occur that the same item is selected more than once (although this was not the case in the previous illustration). When this happens, we simply ignore the repeats and continue to read the random number table until the desired sample size is obtained.

As another illustration, let us suppose that a simple random sample of tax returns are to be selected for auditing from a population of 2,980 tax returns. We first label the tax returns 0001, 0002, ..., 2980. The simple random sample can now be selected by reading distinct sets of four-digit entries from an arbitrary starting point in Table B.1.

It is important not to confuse random selection with arbitrary selection. Suppose you are looking at a pile of 2,980 tax returns and decide to select a sample by just reaching into the pile and grabbing the one, after

which you reach into the pile again to grab another one, and so on, until the desired sample size is achieved. Have you truly selected a simple random sample? You may think you have, but what you have really done is select an *arbitrary sample*, which is one chosen by in a haphazard manner. Although an arbitrary sampling might seem to be the same thing as a simple random sampling, arbitrary sampling can not guarantee that each item has an equal chance of being selected. When you reach into the pile to grab a tax return, you most likely will not take the one on top of the pile or the one on the bottom of the pile. However, once you are aware for this, you may deliberately make an effort to take ones near the top or ones near the bottom of the pile, perhaps making these more likely to be selected than others in the pile. In any event, arbitrary sampling, like convenience sampling provides us with no real way to make an assessment of the likelihood that the resulting sample is a good representation of the population of interest, which is a limitation.

Table 13-1
Employees of Mythical Corporation
Ordered Alphabetically by Last Name

<u>Name</u>	<u>Position</u>	<u>Salary</u>
Appleby	Salesman	\$ 55,000
Bernhardt	Senior Vice President	90,000
Birch	Salesman	50,000
Dickenson	Secretary	20,000
Fry	Junior Vice President	85,000
Gray	President	100,000
Jones	Secretary	10,000
Mendel	Salesman	35,000
Newman	Salesman	65,000
Olsen	Salesman	40,000
Pearson	Salesman	60,000
Quill	Head Secretary	25,000
Smith	Salesman	45,000
Therman	Salesman	75,000
Vern	Salesman	70,000

Self Test Problem 13-1. Decide whether each situation is an example of descriptive statistics or an example of inferential statistics; for each situation which is an example of inferential statistics, identify the population and the sample.

- After examining 129 of the trees in a wooded area, a forest ranger concludes that 21% of all trees in the wooded area have been affected by a recent blight.
- The average distance that all employees in an office travel to get to work is found to be 8.4 miles, after obtaining the distance that each employee travels to get to work.
- A forest ranger reports that 27 of the 129 trees examined in a wooded area have been affected by a recent blight.
- Based on the average distance that employees in a particular office travel to get to work, it is concluded that the average distance which all employees in the corporation travel to get to work is about 8.4 miles.

Self Test Problem 13-2. Decide whether or not each sampling procedure is truly simple random sampling.

- (a) A forest ranger selects a sample of trees from a wooded area by traveling through the area and selecting any tree that catches his eye every 10 or 15 minutes.
- (b) A sample 10 students is selected from a class of 52 students by assigning each student's name to one of the cards in a standard deck of 52 cards, shuffling the deck thoroughly, and choosing the students corresponding to the top 10 cards.

Self Test Problem 13-3. A simple random sample of 10 students is to be selected from a class of 52 students.

- (a) Describe how the students in the class would be labeled in order to use a random number table to select the sample.
- (b) Indicate which labels would be selected for the sample, if we were to start from the beginning of the 12th row on the first page of Table B.1.

Even with the aid of a random number table, the selection of a simple random sample can be tedious. Labeling each of the tax returns in the previous illustration could be quite a chore unless there were some already existing label (perhaps all or part of a social security number) which could be used. One alternative to simple random sampling is systematic random sampling. A *systematic random sample* of size n from a population of size N is one selected by randomly selecting one of the first k items and every k th item thereafter, where $k=N/n$. This may sound complicated, but the idea behind systematic sampling is actually very straightforward.

In order to illustrate how systematic random sampling compares with simple random sampling, imagine that we intend to select a sample of size $n=3$ employees from a mythical corporation consisting of $N=15$ employees. In order to select a systematic random sample of size $n=3$, we let $k = 15/3 = 5$. If a sampling frame listing the 15 employees were available, we would randomly select one of the first $k=5$ names on the list and select every 5th name thereafter. Tables 13-1 and 13-2 are two possible sampling frames for the corporation. Each table is a list containing names, positions, and salaries. Table 13-1 is in alphabetical order by name, and Table 13-2 is ordered by salary.

Suppose first that we use Table 13-1 as a sampling frame. We select the systematic random sample by randomly selecting one of the first five names and selecting every 5th name thereafter. It is easy to check that the five possible systematic random samples of names (with salaries in parenthesis) are as follows:

{Appleby(\$55,000)	Gray(\$100,000)	Pearson(\$60,000)}
{Bernhardt(\$90,000)	Jones(\$10,000)	Quill(\$25,000)}
{Birch(\$50,000)	Mendel(\$35,000)	Smith(\$45,000)}
{Dickenson(\$20,000)	Newman(\$65,000)	Therman(\$75,000)}
{Fry(\$85,000)	Olsen(\$40,000)	Vern(\$70,000)}

Each of these samples would have an equal chance (1 out of 5) of being selected. Alternatively, suppose that we use Table 13-2 as a sampling frame. We select the systematic random sample by randomly selecting one of the first five names and selecting every 5th name thereafter. It's easy to check that the five possible systematic random samples of names (with salaries in parenthesis) are as follows:

{Gray(\$100,000)	Newman(\$65,000)	Olsen(\$40,000)}
{Bernhardt(\$90,000)	Pearson(\$60,000)	Mendel(\$35,000)}
{Fry(\$85,000)	Appleby(\$55,000)	Quill(\$25,000)}
{Therman(\$75,000)	Birch(\$50,000)	Dickenson(\$20,000)}
{Vern(\$70,000)	Smith(\$45,000)	Jones(\$10,000)}

Each of these samples would have an equal chance (1 out of 5) of being selected.

If our purpose in selecting a sample of three employees is to represent the distribution of salaries in the corporation, it would seem desirable for the sample to contain one of the five high salaries, one of the five middle salaries, and one of the five low salaries. Each of the possible systematic samples with Table 13-2 as the sampling frame will contain one of the top five salaries, one of the middle five salaries, and one of the

bottom five salaries. This is not true for each of the possible systematic samples with Table 13-1 as the sampling frame. While none of the systematic samples with Table 13-1 as the sampling frame grossly misrepresent the salaries in the corporation, we are more likely to get a good representation of the salaries in the corporation with systematic random sampling by using Table 13-2 as the sampling frame.

Now, how does systematic random sampling compare with simple random sampling? If we were to use simple random sampling instead of systematic random sampling, it would not matter whether our sampling frame was Table 13-1 or Table 13-2. In either case, there are exactly 70 possible random samples of size 3, each of which would have an equal chance (1 out of 70) of being selected. Since there is no guarantee that a simple random sample of size $n=3$ will contain one of the five high salaries, one of the five middle salaries, and one of the five low salaries, we are just as likely to get a good representation of the salaries with a systematic random sample selected from Table 13-1 as with simple random sampling.

The reason that systematic random sampling with Table 13-2 as the sampling frame is better than systematic random sampling with Table 13-1 as the sampling frame and is also better than simple random sampling, is because Table 13-2 is ordered with respect to salaries. In general, when the sampling frame tends to be in ascending or descending order with respect to the variable of interest, we are more likely to obtain a good representation of the population with systematic random sampling than with simple random sampling. When the sampling frame is in random order with respect to the variable of interest, the likelihood of obtaining a good representation of the population is the same for systematic random sampling and simple random sampling.

In many situations, systematic random sampling might be preferred over simple random sampling, because systematic sampling is easier to implement than simple random sampling, since no complicated labeling scheme is really necessary, nor is the generation of many random numbers necessary. Our comparison of Tables 13-1 and 13-2 as sampling frames also leads us to believe that in general we can expect a systematic random sample to be at least as good a representation of the sampled population as a simple random sample. While this is often true, there are some odd situations where systematic random sampling is more likely to produce a poor representation of a population than simple random sampling.

To illustrate, imagine another mythical corporation consisting of $N=20$ employees with four employees

Table 13-2
Employees of Mythical Corporation
Ordered by Salary

<u>Position</u>	<u>Name</u>	<u>Salary</u>
President	Gray	\$100,000
Senior Vice President	Bernhardt	90,000
Junior Vice President	Fry	85,000
Salesman	Therman	75,000
Salesman	Vern	70,000
Salesman	Newman	65,000
Salesman	Pearson	60,000
Salesman	Appleby	55,000
Salesman	Birch	50,000
Salesman	Smith	45,000
Salesman	Olsen	40,000
Salesman	Mendel	35,000
Head Secretary	Quill	25,000
Secretary	Dickenson	20,000
Secretary	Jones	10,000

Table 13-3
Employees of Another Mythical Corporation
Ordered by Salary within Department

<u>Name</u>	<u>Position</u>	<u>Salary</u>
Taylor	Chm of Dept A	\$50,000
Saunders	Asst Chm of Dept A	45,000
Nelson	Member of Dept A	20,000
King	Member of Dept A	20,000
Templar	Chm of Dept B	45,000
Randall	Asst Chm of Dept B	40,000
Sinclair	Member of Dept B	20,000
Wilde	Member of Dept B	15,000
Hopkirk	Chm of Dept C	47,500
Nichols	Asst Chm of Dept C	45,000
Philips	Member of Dept C	25,000
Fitzhugh	Member of Dept C	27,500
Kent	Chm of Dept D	55,000
Jason	Asst Chm of Dept D	42,500
Simon	Member of Dept D	22,500
Curry	Member of Dept D	22,500
Hayes	Head of Dept E	52,500
Collins	Asst Chm of Dept E	47,500
Post	Member of Dept E	17,500
Martin	Member of Dept E	22,500

in each of five different departments. In order to select a systematic random sample of size $n=5$ using Table 13-3 as the sampling frame, we let $k = 20/5 = 4$. We then select the systematic random sample by randomly selecting one of the first $k=4$ names and selecting every 4th name thereafter. It is easy to check that there are four possible systematic random samples of names that could be selected: one sample consists of the five department chairmen, one sample consists of the five assistant department chairmen, and each of the other two samples consists of one member from each of the five departments. If, as before, our purpose in selecting a sample of five employees is to represent the distribution of salaries in the corporation, it would seem to be desirable for the sample to contain a variety of salaries. However, samples containing only department chairmen or only assistant department chairmen will tend to contain the higher salaries, while samples containing only department members will tend to contain the lower salaries. None of these samples provides a reasonable representation of the distribution of salaries in the corporation.

The reason for the poor performance of systematic random sampling in the previous illustration is because the order in which salaries appear in Table 13-3 is periodic in nature, that is, the salaries increase and decrease in a relatively regular cyclical pattern. However, it need not always be the case that systematic random sampling performs poorly in such situations. Suppose Table 13-3 is to be used as the sampling frame to select a systematic random sample of size $n=4$ from the corporation of $N=20$ employees. We select the systematic random sample by randomly selecting one of the first $k = 20/4 = 5$ names and every 5th name thereafter. It is easy to check that there are five possible systematic random samples of names that could be selected, and each such sample consists of exactly one department chairmen, one assistant department chairmen, and two department members. Each such sample will contain a variety of salaries and be a very good representation of the distribution of salaries in the corporation. In general, when the sampling frame exhibits a periodic or cyclical pattern with respect to the variable of interest, the likelihood of obtaining a good representation of the population with systematic random sampling depends on how the values of n and k are related to the cyclical pattern in the sampling frame.

In all our illustrations involving systematic random sampling, $k=N/n$ conveniently turned out to be an integer. When this is not the case, we can simply round k up which may result in a sample size which is one less than actually desired. An earlier illustration concerned the selection of a simple random sample of tax returns from a population of 2,980 returns. If it is reasonable to assume that the tax returns are in random order, systematic random sampling could be considered the same as random sampling. If a systematic random sample of size $n=100$ were to be selected, then $k=2980/100=29.8$. We could randomly select one of the first 30 returns and every 30th return thereafter, with a small chance that we will obtain a sample size one less than the desired 30 (in which case we could simply use the random number table to select one more tax return).

Strictly speaking, systematic random sampling has been defined here when selecting from a population of known size N . The sizes of the population in each of Tables 13-1, 13-2, and 13-3 were made to be unrealistically small, so that we can easily demonstrate important concepts. We might use a modified version of systematic selection with infinite or very large populations. For example, a pollster may decide to interview every 20th person who comes out of a particular building or may pick the 8th name on every 20th page of a phone directory. Whatever modifications to systematic selection are made, we must always be cognizant of how the method used to obtain a sample affects how well the sample will represent the population.

Self Test Problem 13-4. A sample of students is to be selected from a class of 52 students. Each student's name is assigned to one of the cards in a standard deck of 52 cards.

- (a) Describe how a systematic random sample of 13 students could be selected.
- (b) Describe how a systematic random sample of 10 students could be selected.

Self Test Problem 13-5. Five hundred students are taking an exam in a huge auditorium. The students all start the exam at the same time and are allowed to leave when finished. In order to estimate the average amount of time to complete the exam, the length of time spent on the exam is to be recorded for each in a sample of 50 students.

- (a) The students' test booklets are labeled from 001 to 500, and a random number table is used to select the 50 students whose times are recorded. Is this a simple random sample or a systematic random sample?
- (b) The students' test booklets are labeled from 001 to 500, and a random number table is used to select a random number from 1 to 10. The random number selected turns out to be 8, after which the 50 students whose times are recorded are 008, 018, 028, ..., 498. Is this a simple random sample or a systematic random sample?
- (c) A random number table is used to select a random number from 1 to 10, and the random number selected turns out to be 8. The times are recorded for the 8th student to submit the exam and for every tenth student to submit the exam thereafter. Is this a simple random sample or a systematic random sample?
- (d) Which of the sampling methods described in parts (a), (b), and (c) is the best choice and why?

Answers to Self Test Problems

- 13-1** (a) This is an example of inferential statistics; the population consists of all trees in the wooded area, and the sample consists of the 129 trees which the forest ranger selected to examine. (b) This is an example of descriptive statistics. (c) This is an example of descriptive statistics. (d) This is an example of inferential statistics; the population consists of all employees in the corporation, and the sample consists of all the employees in the selected office used to estimate the average distance.
- 13-2** (a) This is more arbitrary sampling than it is simple random sampling, since the ranger's choice of trees cannot be considered truly random. (b) This can be considered simple random sampling, since a well-shuffled deck of cards can be considered as being in random order.
- 13-3** (a) The students would be labeled with 01, 02, ..., 52. (b) The labels selected starting from the beginning of the 12th row on the first page of Table B.1 are 05, 46, 15, 07, 03, 12, 07, 14, 47, and 42.
- 13-4** (a) Since $k = 52/13 = 4$, a random number table could be used to select a random number from 1 to 4 to determine which of the first four cards in the deck to select. Then every 4th card thereafter would also be selected to determine the sample of size 13. (b) Since $k = 52/10 = 5.2$ is not an integer, a random number table could be used to select a random number from 1 to 6 to determine which of the first six cards in the deck to select. Then every 6th card thereafter would also be selected, with a chance that we will obtain a sample size one less than the desired (in which case we could simply use the random number table to select one more card).
- 13-5** (a) a simple random sample (b) a systematic random sample (c) a systematic random sample (d) The sampling method described in part (c) is the best choice, because the completion times are in ascending order, which makes the likelihood of obtaining a good representation of the population higher with systematic random sampling than with simple random sampling.

Summary

Descriptive statistics is the use of tables, graphs, numerical quantities, etc. to describe a data set; *inferential statistics* is the use of information gathered from a data set to draw a conclusion about a set of items larger than those on which the data was taken. A *population* refers to a set of items about which we wish to draw one or more conclusions; a *sample* refers to a finite set of items which are selected from the population and on which conclusions about the population are based. We may use the terms *population* and *sample* to represent the items on which a specific variable is measured, or we may use the terms *population* and *sample* to

represent the actual values of the variable. Many finite but large populations are treated as infinite; the distinction between finite and infinite populations has no practical impact if the size of the sample is less than 10% of the size of the population, which is often the case.

If our intention is to base conclusions about a population on a sample selected from the population, then our goal should be to select a sample which will be very likely to supply us with an accurate representation of the population; obtaining such a sample often depends more on how the sample is selected than on the size of the sample. In practice, we must often settle for sampling from an *accessible population*, which would consist only of those items to which we have access; the conclusions made from a sample can only be applied to the accessible population from which the sample came. Sometimes it is possible to obtain a list of all the items in a population from which a sample is to be taken; such a list is called a *sampling frame*.

The procedure we use to select items from our accessible population is called a *sampling design*. Each of the many possible sampling designs can be classified as providing either a *random sample* or a *non-random sample*. A random sample is one where items in the population are selected by chance after assigning some non-zero probability of being selected to each item; a non-random sample is one where items are not selected by chance. A non-random sample is often some type of *convenience sample*, which is one consisting of easily available items. Using a convenience sample is generally considered a limitation. Sometimes, this limitation can be reduced by choosing items which are deemed to represent accurately the population of interest based on some expert knowledge and judgment about the characteristics of the population of interest.

In order to be able to assess the likelihood that selected items adequately represent the sampled population, a random sample is generally preferred over a non-random sample. Among the many different types of random samples, the most basic is a *simple random sample* of size n . A simple random sample is one selected so that each item in the sampled population has an equal chance of being selected and is selected independently of any other item. Physical randomization procedures do not always work as well as we might expect and can be difficult and time consuming to devise. Programmable calculators and computer software are often capable of generating random numbers. Tables of random numbers generated by computer are also available.

A *systematic random sample* of size n from a population of size N is one selected by randomly selecting one of the first k items and every k th item thereafter, where $k=N/n$. Systematic random sampling can often be easier to implement than simple random sampling, since no complicated labeling scheme is needed. When the sampling frame tends to be in ascending or descending order with respect to the variable of interest, we are more likely to obtain a good representation of the population with systematic random sampling than with simple random sampling. When the sampling frame is in random order with respect to the variable of interest, the likelihood of obtaining a good representation of the population is the same for systematic random sampling and simple random sampling. When the sampling frame exhibits a periodic or cyclical pattern with respect to the variable of interest, the likelihood of obtaining a good representation of the population with systematic random sampling may be higher than, the same as, or lower than that with simple random sampling. Systematic selection is often modified for use with infinite or very large populations.