CHAPTER 1

Introduction to Statistics

Statisticians have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle.

Maurice Kendall (British statistician)

CONTENTS

This chapter introduces the subject matter of statistics, the art of learning from data. It describes the two branches of statistics, descriptive and inferential. The idea of learning about a population by sampling and studying certain of its members is discussed. Some history is presented.

1.1 INTRODUCTION

Is it better for children to start school at a younger or older age? This is certainly a question of interest to many parents as well as to people who set public policy. How can we answer it?

It is reasonable to start by thinking about this question, relating it to your own experiences, and talking it over with friends. However, if you want to convince

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00001-6

others and obtain a consensus, it is then necessary to gather some objective information. For instance, in many states, achievement tests are given to children at the end of their first year in school. The children's results on these tests can be obtained and then analyzed to see whether there appears to be a connection between children's ages at school entrance and their scores on the test. In fact, such studies have been done, and they have generally concluded that older student entrants have, as a group, fared better than younger entrants. However, it has also been noted that the reason for this may just be that those students who entered at an older age would be older at the time of the examination, and this by itself may be what is responsible for their higher scores. For instance, suppose parents did not send their 6-year-olds to school but rather waited an additional year. Then, since these children will probably learn a great deal at home in that year, they will probably score higher when they take the test at the end of their first year of school than they would have if they had started school at age 6.

A recent study (Table 1.1) has attempted to improve upon earlier work by examining the effect of children's age upon entering school on the eventual number of years of school completed. These authors argue that the total number of years spent in school is a better measure of school success than is a score on an achievement test taken in an early grade. Using 1960 and 1980 census data, they concluded that the age at which a child enters school has very little effect on the total number of years that a child spends in school. Table 1.1 is an abridgment of one presented in their work. The table indicates that for children beginning school in 1949, the younger half (whose average entrance age was 6.29 years) spent an average of 13.77 years, and the older half an average of 13.78 years, in school.

Note that we have not presented the preceding in order to make the case that the ages at which children enter school do not affect their performance in school.

Table 1.1 Total Years in School Related to Starting Age

	Younger half of children		Older half of children	
Year	Average age on starting school	Average number of years completed	Average age on starting school	Average number of years completed
1946	6.38	13.84	6.62	13.67
1947	6.34	13.80	6.59	13.86
1948	6.31	13.78	6.56	13.79
1949	6.29	13.77	6.54	13.78
1950	6.24	13.68	6.53	13.68
1951	6.18	13.63	6.45	13.65
1952	6.08	13.49	6.37	13.53

Source: J. Angrist and A. Krueger, "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, vol. 87, no. 18, 1992, pp. 328–336.

Rather we are using it to indicate the modern approach to learning about a complicated question. Namely, one must collect relevant information, or *data*, and these data must then be described and analyzed. Such is the subject matter of statistics.

1.2 THE NATURE OF STATISTICS

It has become a truism in today's world that in order to learn about something, you must first collect data. For instance, the first step in learning about such things as

1. The present state of the economy
2. The percentage of the voting public who favors a certain proposition
3. The average miles per gallon of a newly developed automobile
4. The efficacy of a new drug
5. The usefulness of a new way of teaching reading to children in elementary school

is to collect relevant data.

Definition Statistics *is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.*

1.2.1 Data Collection

Sometimes a statistical analysis begins with a given set of data; for instance, the government regularly collects and publicizes data about such quantities as the unemployment rate and the gross domestic product. Statistics would then be used to describe, summarize, and analyze these data.

In other situations, data are not yet available, and statistics can be utilized to design an appropriate experiment to generate data. The experiment chosen should depend on the use that one wants to make of the data. For instance, if a cholesterol-lowering drug has just been developed and its efficacy needs to be determined, volunteers will be recruited and their cholesterol levels noted. They will then be given the drug for some period, and their levels will be measured again. However, it would be an ineffective experiment if *all* the volunteers were given the drug. For if this were so, then even if the cholesterol levels of all the volunteers were significantly reduced, we would not be justified in concluding that the improvements were due to the drug used and not to some other possibility. For instance, it is a well-documented fact that any medication received by a patient, whether or not it is directly related to that patient's suffering, will often lead to an improvement in the patient's condition. This is the *placebo effect*, which is not as surprising as it might seem at first, since a patient's belief that she or he is being effectively treated often leads to a reduction in stress, which can result in an improved state of health. In addition, there might have been other—usually unknown—factors that played a role in the reduction of cholesterol levels. Perhaps the weather was unusually warm (or cold), causing the volunteers to spend more or less time outdoors than usual, and this was a factor. Thus, we see that the experiment that calls for giving the drug to all the volunteers is not well designed for generating data from which we can learn about the efficacy of that drug.

A better experiment is one that tries to neutralize all other possible causes of the change of cholesterol level except the drug. The accepted way of accomplishing this is to divide the volunteers into two groups; then one group receives the drug, and the other group receives a tablet (known as a *placebo*) that looks and tastes like the drug but has no physiological effect. The volunteers should not know whether they are receiving the true drug or the placebo, and indeed it is best if the medical people overseeing the experiment also do not know, so their own biases will not play a role. In addition, we want the division of the volunteers into the two groups to be done such that neither of the groups is favored in that it tends to have the "better" patients. The accepted best approach for arranging this is to break up the volunteers "at random," where by this term we mean that the breakup is done in such a manner that all possible choices of people in the group receiving the drug are equally likely. The group that does not receive any treatment (that is, the volunteers that receive a placebo) is called the *control* group.

At the end of the experiment, the data should be described. For instance, the before and after cholesterol levels of each volunteer should be presented, and the experimenter should note whether the volunteer received the drug or the placebo. In addition, summary measures such as the average reduction in cholesterol of members of the control group and members of the drug group should be determined.

Definition *The part of statistics concerned with the description and summarization of data is called* descriptive statistics.

1.2.2 Inferential Statistics and Probability Models

When the experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about the efficacy of the drug. For instance, can we conclude that it is effective in reducing blood cholesterol levels?

Definition *The part of statistics concerned with the drawing of conclusions from data is called* inferential statistics.

To be able to draw a conclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average reduction in cholesterol is lower for the group receiving the drug than for the control group. Can we conclude that this result is due to the drug? Or is it possible that the drug is really ineffective and that the improvement was just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips. Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.)

To be able to draw logical conclusions from data, it is usually necessary to make some assumptions about the chances (or *probabilities*) of obtaining the different data values. The totality of these assumptions is referred to as a *probability model* for the data.

Sometimes the nature of the data suggests the form of the probability model that is assumed. For instance, suppose the data consist of the responses of a selected group of individuals to a question about whether they are in favor of a senator's welfare reform proposal. Provided that this group was *randomly* selected, it is reasonable to suppose that each individual queried was in favor of the proposal with probability p, where p represents the unknown proportion of all citizens in favor of the proposal. The resultant data can then be used to make inferences about p.

In other situations, the appropriate probability model for a given data set will not be readily apparent. However, a careful description and presentation of the data sometimes enable us to infer a reasonable model, which we can then try to verify with the use of additional data.

Since the basis of statistical inference is the formulation of a probability model to describe the data, an understanding of statistical inference requires some knowledge of the theory of probability. In other words, statistical inference starts with the assumption that important aspects of the phenomenon under study can be described in terms of probabilities, and then it draws conclusions by using data to make inferences about these probabilities.

1.3 POPULATIONS AND SAMPLES

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the *population*. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a *sample*.

Definition *The total collection of all the elements that we are interested in is called a* population.

A subgroup of the population that will be studied in detail is called a sample.

In order for the sample to be informative about the total population, it must be, in some sense, representative of that population. For instance, suppose that we are interested in learning about the age distribution of people residing in a given city, and we obtain the ages of the first 100 people to enter the town library. If the average age of these 100 people is 46.2 years, are we justified in concluding that this is approximately the average age of the entire population? Probably not, for we could certainly argue that the sample chosen in this case is not representative of the total population because usually more young students and senior citizens use the library than do working-age citizens. Note that *representative* does not mean that the age distribution of people in the sample is exactly that of the total population, but rather that the sample was chosen in such a way that all parts of the population had an equal chance to be included in the sample.

In certain situations, such as the library illustration, we are presented with a sample and must then decide whether this sample is reasonably representative of the entire population. In practice, a given sample generally cannot be considered to be representative of a population unless that sample has been chosen in a random manner. This is because any specific nonrandom rule for selecting a sample often results in one that is inherently biased toward some data values as opposed to others.

Definition *A sample of k members of a population is said to be a* random sample, *sometimes called a* simple random sample, *if the members are chosen in such a way that all possible choices of the k members are equally likely.*

Thus, although it may seem paradoxical, we are most likely to obtain a representative sample by choosing its members in a totally random fashion without any prior considerations of the elements that will be chosen. In other words, we need not attempt to deliberately choose the sample so that it contains, for instance, the same gender percentage and the same percentage of people in each profession as found in the general population. Rather, we should just leave it up to "chance" to obtain roughly the correct percentages. The actual mechanics of choosing a random sample involve the use of random numbers and will be presented in App. C.

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population by studying the elements of the sample.

*1.3.1 Stratified Random Sampling

A more sophisticated approach to sampling than simple random sampling is the *stratified random sampling* approach. This approach, which requires more initial information about the population than does simple random sampling, can be explained as follows. Consider a high school that contains 300 students in

*The asterisk signifies optional material not used in the sequel.

the first-year class, 500 in the second-year class, and 600 each in the third- and fourth-year classes. Suppose that in order to learn about the students' feelings concerning a military draft for 18-year-olds, an in-depth interview of 100 students will be done. Rather than randomly choosing 100 people from the 2000 students, in a stratified sample one calculates how many to choose from each class. Since the proportion of students who are first-year is $300/2000 = 0.15$, in a stratified sample the percentage is the same and thus there are $100 \times 0.15 = 15$ first-year students in the sample. Similarly, one selects $100 \times 0.25 = 25$ second-year students and $100 \times 0.30 = 30$ third-year and 30 fourth-year students. Then one selects students from each class at random.

In other words, in this type of sample, first the population is *stratified* into subpopulations, and then the correct number of elements is randomly chosen from each of the subpopulations. As a result, the proportions of the sample members that belong to each of the subpopulations are exactly the same as the proportions for the total population. Stratification is particularly effective for learning about the "average" member of the entire population when there are inherent differences between the subpopulations with respect to the question of interest. For instance, in the foregoing survey, the upper-grade students, being older, would be more immediately affected by a military draft than the lower-grade students. Thus, each class might have inherently different feelings about the draft, and stratification would be effective in learning about the feelings of the average student.

1.4 A BRIEF HISTORY OF STATISTICS

A systematic collection of data on the population and the economy was begun in the Italian city-states of Venice and Florence during the Renaissance. The term *statistics*, derived from the word *state*, was used to refer to a collection of facts of interest to the state. The idea of collecting data spread from Italy to the other countries of western Europe. Indeed, by the first half of the 16th century, it was common for European governments to require parishes to register births, marriages, and deaths. Because of poor public health conditions this last statistic was of particular interest.

The high mortality rate in Europe before the 19th century was due mainly to epidemic diseases, wars, and famines. Among epidemics the worst were the plagues. Starting with the Black Plague in 1348, plagues recurred frequently for nearly 400 years. In 1562, as a way to alert the King's court to consider moving to the countryside, the city of London began to publish weekly bills of mortality. Initially these mortality bills listed the places of death and whether a death had resulted from plague. Beginning in 1625, the bills were expanded to include all causes of death.

In 1662 the English tradesman John Graunt published a book entitled *Natural and Political Observations Made upon the Bills of Mortality*. Table 1.2, which notes

Table 1.2 Total Deaths in England

Year	Burials	Plague deaths
1592	25,886	11,503
1593	17,844	10,662
1603	37,294	30,561
1625	51,758	35,417
1636	23,359	10,400

the total number of deaths in England and the number due to the plague for five different plague years, is taken from this book.

Graunt used the London bills of mortality to estimate the city's population. For instance, to estimate the population of London in 1660, Graunt surveyed households in certain London parishes (or neighborhoods) and discovered that, on average, there were approximately 3 deaths for every 88 people. Dividing by 3 shows that, on average, there was roughly 1 death for every 88/3 people. Since the London bills cited 13,200 deaths in London for that year, Graunt estimated the London population to be about

$$13{,}200 \cdot \frac{88}{3} = 387{,}200$$

Graunt used this estimate to project a figure for all England. In his book he noted that these figures would be of interest to the rulers of the country, as indicators of both the number of men who could be drafted into an army and the number who could be taxed.

Graunt also used the London bills of mortality—and some intelligent guesswork as to what diseases killed whom and at what age—to infer ages at death. (Recall that the bills of mortality listed only causes and places of death, not the ages of those dying.) Graunt then used this information to compute tables giving the proportion of the population that dies at various ages. Table 1.3 is one of Graunt's mortality tables. It states, for instance, that of 100 births, 36 people will die before reaching age 6, 24 will die between the ages of 6 and 15, and so on.

Graunt's estimates of the ages at which people were dying were of great interest to those in the business of selling annuities. Annuities are the opposite of life insurance, in that one pays in a lump sum as an investment and then receives regular payments for as long as one lives.

Graunt's work on mortality tables inspired further work by Edmund Halley in 1693. Halley, the discoverer of the comet bearing his name (and also the man who was most responsible, by both his encouragement and his financial support, for the publication of Isaac Newton's famous *Principia Mathematica*), used tables

Table 1.3 Graunt's Mortality Table

Age at death	Deaths per 100 births
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
≥76	1

Note: The categories go up to, but do not include, the right-hand value. For instance, 0–6 means ages 0 through 5 years.

of mortality to compute the odds that a person of any age would live to any other particular age. Halley was influential in convincing the insurers of the time that an annual life insurance premium should depend on the age of the person being insured.

Following Graunt and Halley, the collection of data steadily increased throughout the remainder of the 17th century and on into the 18th century. For instance, the city of Paris began collecting bills of mortality in 1667; and by 1730 it had become common practice throughout Europe to record ages at death.

The term *statistics*, which was used until the 18th century as a shorthand for the descriptive science of states, in the 19th century became increasingly identified with numbers. By the 1830s the term was almost universally regarded in Britain and France as being synonymous with the *numerical science* of society. This change in meaning was caused by the large availability of census records and other tabulations that began to be systematically collected and published by the governments of western Europe and the United States beginning around 1800.

Throughout the 19th century, although probability theory had been developed by such mathematicians as Jacob Bernoulli, Karl Friedrich Gauss, and Pierre Simon Laplace, its use in studying statistical findings was almost nonexistent, as most social statisticians at the time were content to let the data speak for themselves. In particular, at that time statisticians were not interested in drawing inferences about individuals, but rather were concerned with the society as a whole. Thus, they were not concerned with sampling but rather tried to obtain censuses of the entire population. As a result, probabilistic inference from samples to a population was almost unknown in 19th-century social statistics.

It was not until the late 1800s that statistics became concerned with inferring conclusions from numerical data. The movement began with Francis Galton's work on analyzing hereditary genius through the uses of what we would now call regression and correlation analysis (see Chap. 12) and obtained much of its impetus from the work of Karl Pearson. Pearson, who developed the chi-squared goodness-of-fit test (see Chap. 13), was the first director of the Galton laboratory, endowed by Francis Galton in 1904. There Pearson originated a research program aimed at developing new methods of using statistics in inference. His laboratory invited advanced students from science and industry to learn statistical methods that could then be applied in their fields. One of his earliest visiting researchers was W. S. Gosset, a chemist by training, who showed his devotion to Pearson by publishing his own works under the name *Student*. (A famous story has it that Gosset was afraid to publish under his own name for fear that his employers, the Guinness brewery, would be unhappy to discover that one of its chemists was doing research in statistics.) Gosset is famous for his development of the t test (see Chap. 9).

Two of the most important areas of applied statistics in the early 20th century were population biology and agriculture. This was due to the interest of Pearson and others at his laboratory and to the remarkable accomplishments of the English scientist Ronald A. Fisher. The theory of inference developed by these pioneers, including, among others, Karl Pearson's son Egon and the Polish-born mathematical statistician Jerzy Neyman, was general enough to deal with a wide range of quantitative and practical problems. As a result, after the early years of this century, a rapidly increasing number of people in science, business, and government began to regard statistics as a tool able to provide quantitative solutions to scientific and practical problems.

Nowadays the ideas of statistics are everywhere. Descriptive statistics are featured in every newspaper and magazine. Statistical inference has become indispensable to public health and medical research, to marketing and quality control, to education, to accounting, to economics, to meteorological forecasting, to polling and surveys, to sports, to insurance, to gambling, and to all research that makes any claim to being scientific. Statistics has indeed become ingrained in our intellectual heritage.

KEY TERMS

Statistics: The art of learning from data.

Descriptive statistics: The part of statistics that deals with the description and summarization of data.

Inferential statistics: The part of statistics that is concerned with drawing conclusions from data.

Probability model: The mathematical assumptions relating to the likelihood of different data values.

Population: A collection of elements of interest.

Sample: A subgroup of the population that is to be studied.

Random sample of size *k*: A sample chosen in such a manner that all subgroups of size *k* are equally likely to be selected.

Stratified random sample: A sample obtained by dividing the population into distinct subpopulations and then choosing random samples from each subpopulation.

THE CHANGING DEFINITION OF STATISTICS

Statistics has then for its object that of presenting a faithful representation of a state at a determined epoch. (Quetelet, 1849)

Statistics are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (Galton, 1889)

Statistics may be regarded (i) as the study of populations, (ii) as the study of variation, and (iii) as the study of methods of the reduction of data. (Fisher, 1925)

Statistics is a scientific discipline concerned with collection, analysis, and interpretation of data obtained from observation or experiment. The subject has a coherent structure based on the theory of Probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology. (E. Pearson, 1936)

Statistics is the name for that science and art which deals with uncertain inferences—which uses numbers to find out something about nature and experience. (Weaver, 1952)

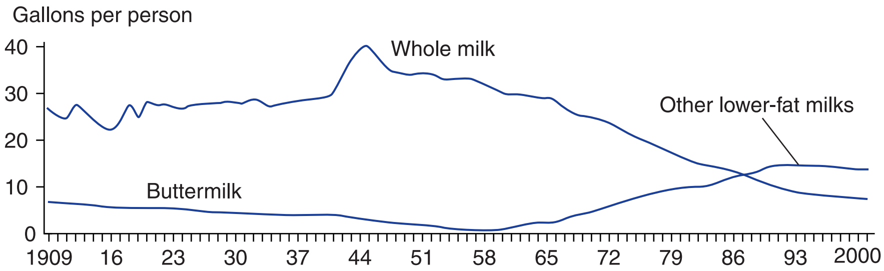
Statistics has become known in the 20th century as the mathematical tool for analyzing experimental and observational data. (Porter, 1986)

Statistics is the art of learning from data. (Ross, 2010)

REVIEW PROBLEMS

1. This problem refers to Table 1.1.
   (a) In which year was there the largest difference between the average number of years of school completed by the younger and older starters?
   (b) Were there more years in which the average number of years completed by the younger starting group exceeded that of the older group, or the opposite?

2. The following is a graph of milk product consumption in the United States from 1909 to 2000. What general conclusion would you draw?



3. The following data yield the percentages of U.S. adults, characterized by educational level, that smoked in the years from 1999 to 2002.
   (a) For which group has there been a steady decline?
   (b) Would you say there is an overall trend?

Cigarette Use in the U.S. (% of all adults)

	1999	2000	2001	2002
Total	**25.8**	**24.9**	**24.9**	**26.0**
Sex				
Male	28.3	26.9	27.1	28.7
Female	23.4	23.1	23.0	23.4
Education				
Non-high school graduate	39.9	32.4	33.8	35.2
High school graduate	36.4	31.1	32.1	32.3
Some college	32.5	27.7	26.7	29.0
College graduate	18.2	13.9	13.8	14.5

4. A medical researcher, trying to establish the efficacy of a new drug, has begun testing the drug along with a placebo. To make sure that the two groups of volunteer patients—those receiving the drug and those receiving a placebo—are as nearly alike as possible, the researcher has decided not to rely on chance but rather to carefully scrutinize the volunteers and then choose the groupings himself. Is this approach advisable? Why or why not?
5. Explain why it is important that a researcher who is trying to learn about the usefulness of a new drug not know which patients are receiving the new drug and which are receiving a placebo.
6. An election will be held next week, and by polling a sample of the voting population we are trying to predict whether the Republican or Democratic candidate will prevail. Which of the following methods of selection will yield a representative sample?

(a) Poll all people of voting age attending a college basketball game.
(b) Poll all people of voting age leaving a fancy midtown restaurant.
(c) Obtain a copy of the voter registration list, randomly choose 100 names, and question them.
(d) Use the results of a television call-in poll, in which the station asked its viewers to call and tell their choice.
(e) Choose names from the telephone directory and call these people.

7. The approach used in Prob. 6e led to a disastrous prediction in the 1936 Presidential election, in which Franklin Roosevelt defeated Alfred Landon by a landslide. A Landon victory had been predicted by the *Literary Digest*. The magazine based its prediction on the preferences of a sample of voters chosen from lists of automobile and telephone owners.
(a) Why do you think the *Literary Digest*'s prediction was so far off?
(b) Has anything changed between 1936 and now that would make you believe that the approach used by the *Literary Digest* would work better today?

8. A researcher is trying to discover the average age at death for people in the United States today. To obtain data, the obituary columns of *The New York Times* are read for 30 days, and the ages at death of people in the United States are noted. Do you think this approach will lead to a representative sample?

9. If, in Prob. 8, the average age at death of those recorded is 82.4 years, what conclusion could you draw?

10. To determine the proportion of people in your town who are smokers, it has been decided to poll people at one of the following local spots:
(a) The pool hall
(b) The bowling alley
(c) The shopping mall
(d) The library
Which of these potential polling places would most likely result in a reasonable approximation to the desired proportion? Why?

11. A university plans on conducting a survey of its recent graduates to determine information on their yearly salaries. It randomly selected 200 recent graduates and sent them questionnaires dealing with their present jobs. Of these 200, however, only 86 questionnaires were returned. Suppose that the average of the yearly salaries reported was $75,000.
(a) Would the university be correct in thinking that $75,000 was a good approximation to the average salary level of all its graduates? Explain the reasoning behind your answer.

(b) If your answer to (a) is no, can you think of any set of conditions relating to the group that returns questionnaires for which $75,000 would be a good approximation?

12. An article reported that a survey of clothing worn by pedestrians killed at night in traffic accidents revealed that about 80 percent of the victims were wearing dark-colored clothing and 20 percent were wearing light-colored clothing. The conclusion drawn in the article was that it is safer to wear light-colored clothing at night.

(a) Is this conclusion justified? Explain.

(b) If your answer to (a) is no, what other information would be needed before a final conclusion could be drawn?

13. Critique Graunt's method for estimating the population of London. What implicit assumption is he making?

14. The London bills of mortality listed 12,246 deaths in 1658. Supposing that a survey of London parishes showed that roughly 2 percent of the population died that year, use Graunt's method to estimate London's population in 1658.

15. Suppose you were a seller of annuities in 1662, when Graunt's book was published. Explain how you would make use of his data on the ages at which people were dying.

16. Based on Table 1.2, which of the five plague years appears to have been the most severe? Explain your reasoning.

17. Based on Graunt's mortality table:

(a) What proportion of babies survived to age 6?

(b) What proportion survived to age 46?

(c) What proportion died between the ages of 6 and 36?

18. Why do you think that the study of statistics is important in your field? How do you expect to utilize it in your future work?

19. The chart on the following page gives the demographic and socioeconomic characteristics of adult smokers in upstate New York in 2006. Use it to determine if the following statements appear to be true. Answer yes or no.

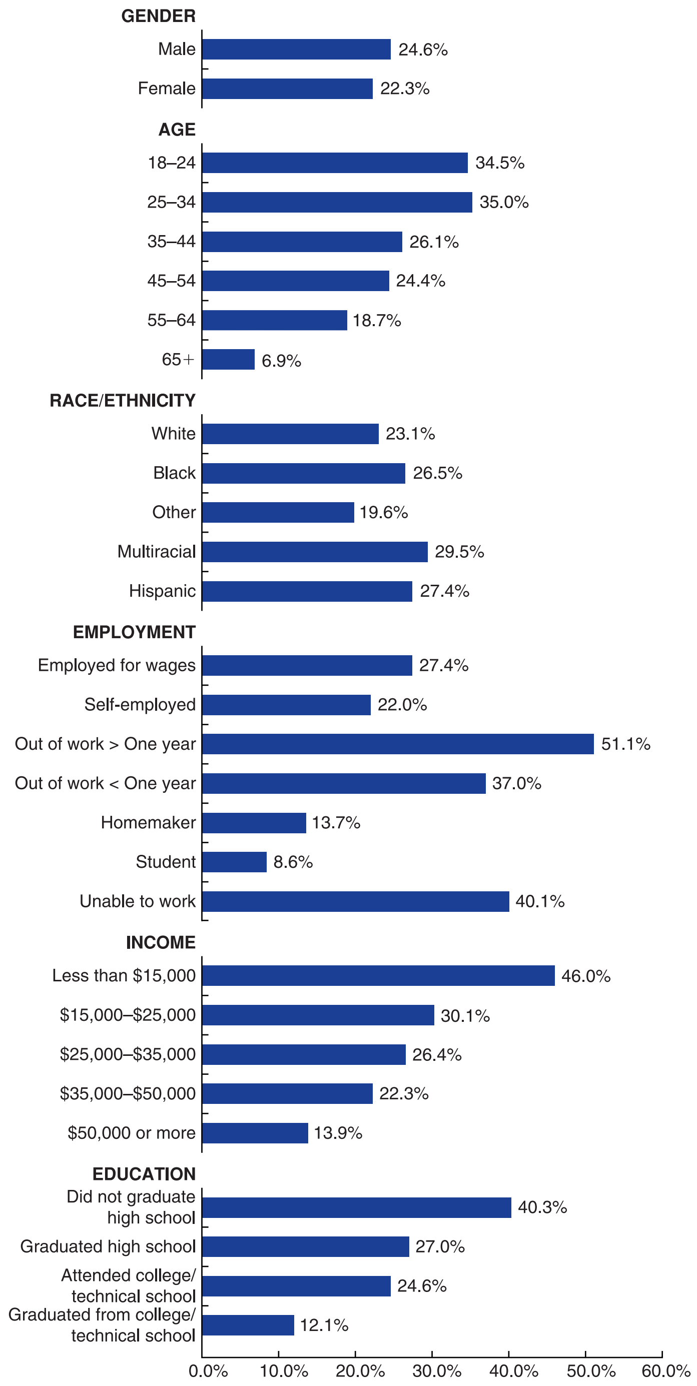
(a) A higher proportion of men than of women are current smokers.

(b) The longer a person has been out of work, the more likely that person is a smoker.

(c) The more education a person has, the more likely that person is to smoke.

(d) Ethnicity does not appear to be related to smoking prevalence.

It should be noted that even when the answer to a preceding question is yes that does not necessarily mean that the characteristic is a cause of smoking, but only that there is a positive association between it and smoking. The concept of association, or *correlation,* will be considered in Chap. 3.



Demographic and Socio-Economic Characteristics of Adult Smokers in Upstate New York, 2006