- Here are other examples of probability sampling: ***Systematic*** random sampling takes every *k*th subject in the sampling frame list. ***Stratified*** random sampling divides the population into groups (strata) and takes a random sample from each stratum. ***Cluster*** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. ***Multistage*** sampling uses combinations of these methods.

Chapter 3 introduces statistics for describing samples and corresponding parameters for describing populations. Hence, its focus is on *descriptive statistics*.

PROBLEMS

Practicing the Basics

2.1. Explain the difference between
- **(a)** Discrete and continuous variables
- **(b)** Categorical and quantitative variables
- **(c)** Nominal and ordinal variables

Why do these distinctions matter for statistical analysis?

2.2. Identify each variable as categorical or quantitative:
- **(a)** Number of pets in family
- **(b)** County of residence
- **(c)** Choice of auto (domestic or import)
- **(d)** Distance (in miles) commute to work
- **(e)** Choice of diet (vegetarian, nonvegetarian)
- **(f)** Time spent in previous month browsing the World Wide Web
- **(g)** Ownership of personal computer (yes, no)
- **(h)** Number of people you have known with AIDS (0, 1, 2, 3, 4 or more)
- **(i)** Marriage form of a society (monogamy, polygyny, polyandry)

2.3. Which scale of measurement (nominal, ordinal, or interval) is most appropriate for
- **(a)** Attitude toward legalization of marijuana (favor, neutral, oppose)
- **(b)** Gender (male, female)
- **(c)** Number of children in family (0, 1, 2, . . .)
- **(d)** Political party affiliation (Democrat, Republican, Independent)
- **(e)** Religious affiliation (Catholic, Jewish, Protestant, Muslim, other)
- **(f)** Political philosophy (very liberal, somewhat liberal, moderate, somewhat conservative, very conservative)
- **(g)** Years of school completed (0, 1, 2, 3, . . .)
- **(h)** Highest degree attained (none, high school, bachelor's, master's, doctorate)
- **(i)** College major (education, anthropology, physics, sociology, . . .)
- **(j)** Test score (0–100 range for scores)
- **(k)** Employment status (employed full time, employed part time, unemployed)

2.4. Which scale of measurement is most appropriate for
- **(a)** Occupation (plumber, teacher, secretary, . . .)
- **(b)** Occupational status (blue collar, white collar)
- **(c)** Social status (lower, middle, upper class)
- **(d)** Statewide murder rate (number of murders per 1000 population)
- **(e)** County population size (number of people)
- **(f)** Population growth rate (in percentages)
- **(g)** Community size (rural, small town, large town, small city, large city)
- **(h)** Annual income (thousands of dollars per year)
- **(i)** Attitude toward affirmative action (favorable, neutral, unfavorable)
- **(j)** Lifetime number of sexual partners

2.5. Which scale of measurement is most appropriate for "attained education" measured as
- **(a)** Number of years (0, 1, 2, 3, . . .)
- **(b)** Grade level (elementary school, middle school, high school, college, graduate school)
- **(c)** School type (public school, private school)

2.6. Give an example of a variable that is **(a)** categorical, **(b)** quantitative, **(c)** ordinal scale, **(d)** nominal scale, **(e)** discrete, **(f)** continuous, **(g)** quantitative and discrete.

2.7. A poll conducted by YouGov for the British newspaper *The Daily Telegraph* in June 2006 asked a random sample of 1962 British adults several questions about their image of the U.S. One question asked, "How would you rate George W. Bush as a world leader?" The possible choices were (great leader, reasonably satisfactory leader, pretty poor leader, terrible leader).
- **(a)** Is this four-category variable nominal, or ordinal? Why?
- **(b)** Is this variable continuous, or discrete? Why?
- **(c)** Of the 93% of the sample who responded, the percentages in the four categories were 1% (great leader), 16%, 37%, 46% (terrible leader). Are these values statistics, or parameters? Why?

2.8. A survey asks subjects to rate five issues according to their importance in determining voting intention

for U.S. senator, using the scale (very important, somewhat important, unimportant). The issues are foreign policy, unemployment, inflation, the arms race, and civil rights. The evaluations can be treated as five variables: foreign policy evaluation, unemployment evaluation, and so on. These variables represent what scale of measurement?

2.9. Which of the following variables could theoretically be measured on a continuous scale? **(a)** Method of contraception used, **(b)** length of time of residence in a state, **(c)** task completion time, **(d)** intelligence, **(e)** authoritarianism, **(f)** alienation, **(g)** county of residence.

2.10. Which of the following variables are continuous when the measurements are as fine as possible? **(a)** Age of mother, **(b)** number of children in family, **(c)** income of spouse, **(d)** population of cities, **(e)** latitude and longitude of cities, **(f)** distance of home from place of employment, **(g)** number of foreign languages spoken.

2.11. A class has 50 students. Use the column of the first two digits in the random number table (Table 2.1) to select a simple random sample of three students. If the students are numbered 01 to 50, what are the numbers of the three students selected?

2.12. A local telephone directory has 400 pages with 130 names per page, a total of 52,000 names. Explain how you could choose a simple random sample of 5 names. Using the second column of Table 2.1 or software or a calculator, select 5 random numbers to identify subjects for the sample.

2.13. Explain whether an experiment or an observational study would be more appropriate to investigate the following:

(a) Whether or not cities with higher unemployment rates tend to have higher crime rates

(b) Whether a Honda Accord or a Toyota Camry gets better gas mileage

(c) Whether or not higher college GPAs tend to occur for students who had higher scores on college entrance exams

(d) Whether or not a special coupon attached to the outside of a catalog makes recipients more likely to order products from a mail-order company

2.14. A study is planned to study whether passive smoking (being exposed to secondhand cigarette smoke on a regular basis) leads to higher rates of lung cancer.

(a) One possible study is to take a sample of children, randomly select half of them for placement in an environment where they are passive smokers, and place the other half in an environment where they are not exposed to smoke. Then 60 years later the observation is whether each has developed lung cancer. Would this study be an experimental study or an observational study? Why?

(b) For many reasons, including time and ethics, it is not possible to conduct the study in (a). Describe a way that *is* possible, and indicate whether it would be an experimental or observational study.

2.15. Table 2.2 shows the result of the 2000 Presidential election and the predictions of several organizations in the days before the election. The sample sizes were typically about 2000. The percentages for each poll do not sum to 100 because of voters reporting as undecided or favoring another candidate.

(a) What factors cause the results to vary somewhat among organizations?

(b) Identify the sampling error for the Gallup poll.

TABLE 2.2

	Predicted Vote		
Poll	Gore	Bush	Nader
Gallup	46	48	4
Harris	47	47	5
ABC	45	48	3
CBS	45	44	4
NBC	44	47	3
Pew Research	47	49	4
Actual vote	48.4	47.9	2.7

Source: www.ncpp.org/

2.16. The BBC in Britain requested viewers to call the network and indicate their favorite poem. Of more than 7500 callers, more than twice as many voted for Rudyard Kipling's *If* than for any other poem. The BBC reported that this was the clear favorite.

(a) Explain what it means to call this a "volunteer sample."

(b) If the BBC truly wanted to determine Brits' favorite poem, how could it more reliably do so?

2.17. A Roper Poll was designed to determine the percentage of Americans who express some doubt that the Nazi Holocaust occurred. In response to the question, "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" 22% said it was possible the Holocaust never happened. The Roper organization later admitted that the question was worded in a confusing manner. When the poll asked, "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" only 1%

said it was possible it never happened.[5] Use this example to explain the concept of response bias.

2.18. Refer to Exercise 2.12 about selecting 5 of 52,000 names on 400 pages of a directory.
- **(a)** Select five numbers to identify subjects for a systematic random sample of five names from the directory.
- **(b)** Is cluster sampling applicable? How could it be carried out, and what would be the advantages and disadvantages?

2.19. You plan to sample from the 5000 students at a college, to compare the proportions of men and women who believe that the legal age for alcohol should be changed to 18. Explain how you would proceed if you want a systematic random sample of 100 students.

2.20. You plan to sample from the 3500 undergraduate students enrolled at the University of Rochester, to compare the proportions of female and male students who would like to see the U.S. have a female President.
- **(a)** Suppose that you use random numbers to select students, but you stop selecting females as soon as you have 40, and you stop selecting males as soon as you have 40. Is the resulting sample a simple random sample? Why or why not?
- **(b)** What type of sample is the sample in (a)? What advantage might it have over a simple random sample?

2.21. Clusters versus strata:
- **(a)** With a cluster random sample, do you take a sample of (i) the clusters? (ii) the subjects within every cluster?
- **(b)** With a stratified random sample, do you take a sample of (i) the strata? (ii) the subjects within every stratum?
- **(c)** Summarize the main differences between cluster sampling and stratified sampling in terms of whether you sample the groups or sample from within the groups that form the clusters or strata.

Concepts and Applications

2.22. Refer to the *Student survey* data file introduced in Exercise 1.11 (page 8). For each variable in the data set, indicate whether it is:
- **(a)** Categorical or quantitative
- **(b)** Nominal, ordinal, or interval

2.23. Repeat the previous exercise for the data file created in Exercise 1.12 (page 9).

2.24. You are directing a study to determine the factors that relate to good academic performance at your school.
- **(a)** Describe how you might select a sample of 100 students for the study.
- **(b)** List some variables that you would measure. For each, provide the scale you would use to measure it, and indicate whether statistical analysis could treat it as (i) categorical or quantitative, (ii) nominal, ordinal, or interval, (iii) continuous or discrete.
- **(c)** Give an example of a research question that could be addressed using data on the variables you listed in (b).

2.25. With ***quota sampling*** a researcher stands at a street corner and conducts interviews until obtaining a quota representing the relative sizes of various groups in the population. For instance, the quota might be 50 factory workers, 100 housewives, 60 elderly people, 30 blacks, and so forth. Is this a probability or nonprobability sampling method? Explain, and discuss potential advantages or disadvantages of this method. (Professional pollsters such as Gallup used this method until 1948, when they incorrectly predicted that Dewey would defeat Truman in a landslide in the presidential election.)

2.26. When the Yankelovich polling organization asked,[6] "Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?" 80% of the sample answered *yes*. When they posed the question, "Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidate they support?" only 40% said *yes*. Explain what this example illustrates, and use your answer to differentiate between sampling error and response bias in survey results.

2.27. In each of the following situations, evaluate whether the method of sample selection is appropriate for obtaining information about the population of interest. How would you improve the sample design?
- **(a)** A newspaper wants to determine whether its readers believe that government expenditures should be reduced by cutting benefits for the disabled. They provide an Internet address for readers to vote *yes* or *no*. Based on 1434 Internet votes, they report that 93% of the city's residents believe that benefits should be reduced.
- **(b)** A congresswoman reports that letters to her office are running 3 to 1 in opposition to

[5] *Newsweek*, July 25, 1994.

[6] *Source*: *A Mathematician Reads the Newspaper*, by J. A. Paulos, Basic Books, 1995, p. 15.

the passage of stricter gun control laws. She concludes that approximately 75% of her constituents oppose stricter gun control laws.

(c) An anthropology professor wanted to compare attitudes toward premarital sex of physical science majors and social science majors. She administered a questionnaire to her large class of Anthropology 437, Comparative Human Sexuality. She found no appreciable difference between her physical science and social science majors in their attitudes, so she concluded that the two student groups were about the same in their relative acceptance of premarital sex.

(d) A questionnaire was mailed to a simple random sample of 500 household addresses in a city. Ten were returned as bad addresses, 63 were returned completed, and the rest were not returned. The researcher analyzed the 63 cases and reported that they represent a "simple random sample of city households."

(e) A principal in a large high school is interested in student attitudes toward a proposed achievement test to determine whether a student should graduate. She lists all of the first-period classes, assigning a number to each. Then, using a random number table, she chooses a class at random and interviews every student in that class about the proposed test.

2.28. A content analysis of a daily newspaper studies the percentage of newspaper space devoted to news about entertainment. The sampling frame consists of the daily editions of the newspaper for the previous year. What potential problem might there be in using a systematic sample with skip number equal to 7 or a multiple of 7?

2.29. In a systematic random sample, every subject has the same chance of selection, but the sample is not a simple random sample. Explain why.

2.30. With a total sample of size 100, we want to compare Native Americans to other Americans on the percentage favoring legalized gambling. Why might it be useful to take a disproportional stratified random sample?

2.31. In a cluster random sample with equal-sized clusters, every subject has the same chance of selection. However, the sample is not a simple random sample. Explain why not.

2.32. Find an example of results of an Internet poll. Do you trust the results of the poll? If not, explain why not.

2.33. To sample residents of registered nursing homes in Yorkshire, UK, I construct a list of all nursing homes in the county, which I number from 1 to 110. Beginning randomly, I choose every tenth home on the list, ending up with 11 homes. I then obtain lists of residents from those 11 homes, and I select a simple random sample from each list. What kinds of sampling have I used?

For multiple-choice questions 2.34–2.37, select the best response.

2.34. A simple random sample of size n is one in which:

(a) Every nth member is selected from the population.

(b) Each possible sample of size n has the same chance of being selected.

(c) There must be exactly the same proportion of women in the sample as is in the population.

(d) You keep sampling until you have a fixed number of people having various characteristics (e.g., males, females).

(e) A particular minority group member of the population is less likely to be chosen than a particular majority group member.

(f) All of the above

(g) None of the above

2.35. If we use random numbers to take a simple random sample of 50 students from the 20,000 students at a university,

(a) It is impossible to get the random number 11111, because it is not a random sequence.

(b) If we get 20001 for the first random number, for the second random number that number is less likely to occur than the other possible five-digit random numbers.

(c) The draw 12345 is no more or less likely than the draw 11111.

(d) Since the sample is random, it is *impossible* that it will be non-representative, such as having only females in the sample.

2.36. Crosson (1994, p. 168) described an analysis of published medical studies involving treatments for heart attacks. In the studies having randomization and strong controls for bias, the new therapy provided improved treatment 9% of the time. In studies without randomization or other controls for bias, the new therapy provided improved treatment 58% of the time. Select the correct response(s).

(a) This result suggests it is better not to use randomization in medical studies, because it is harder to show that new ideas are beneficial.

(b) Many newspaper articles that suggest that a particular food, drug, or environmental agent is harmful or beneficial should be viewed skeptically, unless we learn more about the statistical design and analysis for the study.

(c) This result suggests that you should be skeptical about published results of medical studies that are not randomized, controlled studies.